Self-Modelling in Inference about Absence

A Thesis
Presented to
Wellcome Centre for Human Neuroimaging; Institute of Neurology
University College London

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Matan Mazor

September 2021

Approved for the Division
(Brain Sciences)

Stephen M. Fleming	Karl J. Friston

I, Matan Mazor, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I would like to thank my supervisors, Steve Fleming, for being a wonderful scientific mentor and role-model, and Karl Friston, for your generous advice and support. Steve, I am ever grateful for these four years of training in the supportive, excellence-seeking environment that you cultivated in the UCL MetaLab. I had the privilege of collaborating with amazing scientists, including Lucie Charles, Rani Moran, Roy Tal, Chudi Gong and Nadine Dijkstra. For their help and support, I thank Peter Zeidman, Dan Bang, Max Rollwage, Marion Rouault, and the FIL imaging support team, as well as a supportive international network of scientists, including Felix Henninger, Elisa Filevich, Kristian Lange, Chester Ismay, Brian Maniscalco, Arial Zylberberg, Jorge Morales, my examiners David Lagnado and Heleen Slagter, and the twitter and Stack Exchange communities. I thank Josh Tenenbaum and Tomer Ullman for your mentorship in my visit to MIT and Harvard, and the Bogue Fellowship committee for their decision to support this career-changing visit. This PhD would not have been possible without UCL's Graduate and Overseas Research Scholarships, and it would have been much harder without the support of a Kenneth Lindsay Scholarship. I am grateful to my mentors from Tel Aviv University: Roy Mukamel, Liad Mudrik, Naama Friedman and Roni Katzir, for believing in me as a junior cognitive scientist and patiently guiding me in my first steps. My friends at the Wellcome Centre for Human Neuroimaging and MetaLab - you made these years not only enriching and educating, but also memorable and enjoyable. Dina Silanteva, Alisa Loosen and the Hackney Wick gang - thanks for being the best support bubble in times of a global pandemic. I thank the Mary Ward Cafe at Queen Square and the Hari Krishna volunteers for the vegan food that kept my brain going. My friends Darya Mosenzon, Halely Balaban, Maya Ankaoua, Maayan Keshev, Roni Maimon, Netta Green, Alon Rubin, Ezer Rasin and Yohai Szulszepper - thanks for being there for me. Amnon David Ar - you taught me how to be a painter. I'll forever be indebted for your lessons about observation, friendship and perseverance, which are still shaping the person and researcher I am striving to be today. My parents Ora and Shai, you care about this arbitrary academic title much less than you want me to be happy and stand up for my principles. Toda. Finally, I would like to thank my brother Noam, and our four-legged, hairy and smelly friend, B7, for being my companions and compass in this long journey.

Table of Contents

List of Tables

List of Figures

Abstract

Representing the absence of an object requires one to know that they would know if it were present. This form of second-order, counterfactual reasoning critically relies on access to a mental self-model, specifying expected perceptual and cognitive states under different world states. This thesis addresses open questions regarding inference about absence in perceptual decision making: its reliance on prior metacognitive knowledge, relative encapsulation from metacognitive monitoring, neural underpinning, and relation with default-reasoning. I start by showing that in visual search, implicit metacognitive knowledge about spatial attention supports inference about the absence in the first trial of an experiment, and that this knowledge is dissociable from explicit metacognitive knowledge. Further underscoring the richness and complexity of this knowledge, I find that people are able to accurately predict their future search times, even for complex, unfamiliar displays. Participants' predictions were better aligned with their own search times than with those of other participants, suggesting that this self-knowledge is person-specific. I then ask what factors contribute to confidence in decisions about presence and absence. Reverse-correlation analysis reveals stimulus features that contribute to detection decisions and confidence. I discuss these findings in the context of sensory noise estimation. Using functional MRI, I find that a network of frontal and parietal regions that are implicated in decision confidence are mostly invariant to whether subjective confidence is rated with respect to decisions about presence or absence. In interpreting these results, I formulate computational models that monitor fluctuations in external stimulus strength and in internal attentional states. Finally, in six behavioural experiments, different levels of the cognitive hierarchy are found to be sensitive to different notions of absence. I conclude with a discussion of ways in which inference about absence can be used by cognitive scientists for probing implicit metacognitive beliefs and studying the mental self-model.

Impact Statement

This thesis is submitted in the strange world of 2021. Twice a week, I start my day with a rapid lateral flow covid-19 test. I wipe a swab inside my nostrils, transfer the sample into a liquid and then place two drops on a test kit. I then wait to read the results: two lines indicate a positive result, and one line a negative one. But why does the test need to have two lines? Why can't a line indicate a positive result, and zero lines a negative one?

For a positive test result, this additional control line doesn't add much. Its importance is for interpreting a negative test. Two lines indicate that covid-19 antigens were detected, one line indicates that covid-19 antigens were not detected and that the test is working, and zero lines indicate that antigens were not detected, but that this is not very informative, because other things that should have been detected were not detected either. Without this additional control line, we have no way of telling between these last two options.

Detecting the presence of covid-19 antigens in a sample is conceptually similar to other detection and search tasks, such as detecting the presence or absence of a red sock in a drawer. But unlike the covid test, upon not finding a red sock I don't have a control line to indicate that the sock would have been found if it were present, or that my vision is intact. Instead, I need to rely on some knowledge about my perception and attention - for example that I would not have missed the sock if it were there. This is a unique feature of decisions about the *absence* of things: without a positive control, they must rely on some form of *self-modelling*.

In a series of studies I examine behavioural and neural activation patterns in visual detection and visual search tasks, and ask whether they provide evidence for reliance on self-modelling, specifically when participants report the absence of stimuli. In carrying out this research I adopted high standards of transparency and openness: all experiments were pre-registered and time-locked with respect to data acquisition, all data (including raw neuroimaging data) is openly shared, and my analysis scripts are openly available. The thesis itself is written using the R package thesisdown, making all statistical analysis 100% reproducible. Some findings from these studies are published in eLife, and as a Registered Report in Neuroscience of Consciousness, and other parts are currently being reviewed for publication.

This thesis opens a novel research programme, using inference about absence to ask questions about self-modelling. In these first studies I focused on healthy adults, but in the future these ideas hold promise for understanding psychological conditions such as obsessive compulsive disorder, where the bar for inferring absence (for example, of germs on one's hand) is set exceptionally high. I hope my PhD output inspires further research on inference about absence, and its reliance on self-modelling.

To my fellow non-human primates of Queen Square, whose experience of London was very different to mine.

Introduction

You are in the grocery shop. On your grocery list are one carton of oat milk and one guava. You search through the shelves and find your favourite oat milk. You place the carton in your basket and move on to the fruit aisle. You visually scan the fruit boxes, but you already have a strong feeling that you will not find guavas in this store. You would have already smelled the guavas if they were anywhere around you. But then again, maybe something is wrong with your sense of smell? You grab a mandarin and sniff it. Your sense of smell is intact. You can be confident that there are no guavas around.

0.1 Inference about absence

Finding the oat milk carton was straightforward. As soon as you identified it you were convinced in its presence, no reflection or deliberation required. In contrast, concluding that no guavas were present took you longer and involved more complex cognitive processes. You had to rely on the absence of smell or sight of the fruit to reach a conclusion. In philosophical writings, this is known as Argument from Ignorance (*Argumentum ad ignorantiam*): the fallacy of accepting a statement as true only because it hasn't been disproved (Locke, 1836). Although logically unsound, *Argumentum ad ignorantiam* is widely applied by humans in different situations and contexts (Oaksford & Hahn, 2004). One particular context which invites such reasoning is that of inference about absence. Positive evidence is rarely available to support inference about absence, and so it is almost exclusively made on the basis of a failure to find evidence for presence.

Basing inference on the absence of evidence can sometimes be rational from a Bayesian standpoint (Oaksford & Hahn, 2004). For this to be the case, the individual must know the sensitivity and specificity of the perceptual or cognitive system at hand. For example, in order for the inference "I don't smell a guava, therefore there are no guavas in this store" to be logically sound, I need to know that the probability of me not smelling a guava is very low if it is nearby, and so is the probability of me imagining the smell of a guava when it is not there. In other words, in order to make valid inferences about absences I need to know things about myself and my cognitive processes (see next section 0.2.2 for a formal unpacking of this logical derivation). In the above example, this is evident in that my certainty in the absence of a guava increased after smelling the mandarin. Critically, smelling the mandarin did not provide me with any additional information about the layout of the shop or

Figure 1: Guavas.

the seasonal availability of tropical fruit, but about my own perceptual system.

The following section introduces a computational formulation of this self-knowledge account, based in formal semantics and Bayesian theories of cognition, and exemplifies how different patterns of results can be interpreted in light of this formulation. This formulation is then followed by descriptions of several independent lines of experimental findings that all demonstrate a role for self-knowledge in inference about absence.

0.2 Probabilistic reasoning, criterion setting, and self knowledge

The intimate link between inference about absence and self-knowledge has been recognized in the fields of linguistics, formal logic, and artificial intelligence. In *default-reasoning logic* (Reiter, 1980), a failure to provide a proof for a statement is transformed into a proof for the negation of the statement using the *closed world assumption*: the assumption that a proof would have been found if it was available. Similarly, Linguist Benoît de Cornulier's refers to *epistemic closure*: the notion that all there is to be known is in fact known. This is reflected in his two definitions of *knowing whether* (De Cornulier, 1988):

Symmetrical definition:

'John knows whether P' means that:

1. If P, John knows that P.
2. If not-P, John knows that not-P.

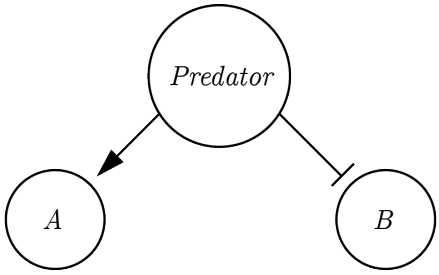
Dissymmetrical definition:

'John knows whether P' means that:

1. If P, John knows that P.
2. John knows that 1 holds.

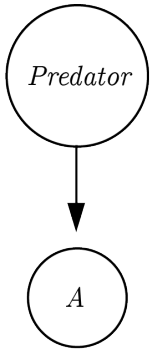
0.2.1 Second-order cognition

The symmetric definition entails a *first-order process*, as no knowledge about the system itself is used in the process of inferring the world state. This definition applies to scenarios in which it is possible to have direct evidence against the veracity of a proposition. For example, a hypothetical organism can be equipped with sensors A and B that are tuned to the presence or absence of a predator, respectively. This organism can be said to know whether there is a predator around or not. It will know that a predator is nearby if A is on and B is off, and it will know there is no predator around if B is on and A is off (similar to the *Neuron-Antineuron* architecture in Gold & Shadlen (2001)). Such an organism can be said to implement the symmetrical definition of to know whether presented above. The symmetric architecture is redundant: assuming perfect information flow there is a perfect negative correlation between the activations of sensors A and B. Conversely, the asymmetric definition only necessitates one sensor that is sensitive to the presence of a predator. The organism will know that the predator is around if the sensor is activated, and will conclude that it is not around if the sensor is not activated. This inference is dependent on the confidence of the organism that the sensor will always be activated by the presence of a predator (the negative test validity of its sensor, see section 0.2.2). In that sense, the asymmetric



	Predator present	Predator absent
Sensor A	1	0
Sensor B	0	1

Figure 2: A symmetric implementation of a predator-detector.



	Predator present	Predator absent
Sensor A	1	0

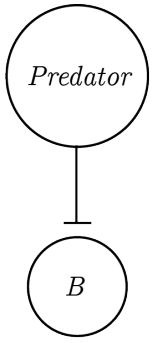
Figure 3: An asymmetric implementation of a predator-detector.

definition entails a *second-order process*. This implementation assumes that the absence of a predator is a *default state*. Making this assumption makes the system leaner: instead of having two sensors, only one sensor is needed to mark deviations from a default state (Reiter, 1980). This default-reasoning has an interesting property: it is *non-monotonous*. Accepting the default state (the absence of a predator in the above example) can only be done tentatively and can potentially be overridden by future evidence. This is not true for the deviant state (here, the presence of a predator), which once accepted cannot be retracted based on the absence of new evidence. In other words, while beliefs about the absence of a predator can be overturn by evidence for presence, beliefs about the presence of a predator cannot be overturn by the absence of evidence for presence.

The asymmetric architecture requires that the organism knows that the presence of a predator would activate sensor A. Only then can the organism take the absence of input from A as evidence for the absence of a predator. Without this knowledge, the organism will be able to represent the presence of a predator (when A is activated), but not its absence. Indeed, it has been pointed out that Reiter's Default Logic is an

autoepistemic logic, which is based on an agent's ability to introspect over their own belief states (Denecker, Marek, & Truszczynski, 2011).

The mirror architecture is also possible: taking the presence of a predator to be a default state and using a sensor to mark deviations from this state, i.e., the absence of a predator. This architecture is perfectly equivalent to the previous one for systems



	Predator present	Predator absent
Sensor B	0	1

Figure 4: An asymmetric implementation of a predator-detector with a pessimistic prior.

that are composed of sensors only. All activated sensors in the first architecture are silenced in the second architecture and vice versa. However, for multi-layered systems that generate higher-level representations from sensory input, the second architecture becomes unreasonably huge. In such systems, if the default state is taken to be "everything is happening," then for every sensory input the system should generate the abstract representation of all possible *combinations* of sensory inputs that were not experienced — $2^n - 1$ in total, n being the number of sensors. This number becomes unrealistic even with a modest number of 100 sensors (2^{99}, or more than a million million million million millions), and is even less realistic for complex systems that are equipped with eyes, thalami and cortices.

This has dramatic consequences for systems that need to flexibly represent a rich space of entities or events, using a set of finite building blocks such as sensors and atomic concepts. Such hierarchical, complex systems are compelled to implement an architecture analog to the one in figure 4, namely to represent presences only, and infer absence by relying on their own self-representation. In other words, the maintenance of a reliable self-representation can be costly, but not nearly as costly as the alternative of representing absences and presences in a symmetrical way.

0.2.2 Computational models of detection

In psychological experiments of near-threshold detection, participants are required to decide whether a stimulus (for example a faint dot) was present or absent from a display. Using De Cornulier's formulation, we can ask which of the two definitions better describes the inferential machinery that is engaged in detection tasks. Is

it the case that participants perceive positive evidence for the absence of a target (symmetrical definition), or alternatively, do they rely on the metacognitive belief that they would have seen the target if it was present (dissymetrical definition)?

The High-Threshold model

The *high-threshold model* of visual detection (Blackwell, 1952) formalizes this process in a way that shares conceptual similarity with De Cornulier's dissymemetrical definition. According to this model, the probability of detecting the signal d scales with stimulus intensity. If participants detect the signal, they respond with 'yes.' The parameter d is a perceptual parameter: it captures variables such as objective stimulus intensity (for example, in units of luminance) and sensory sensitivity (for example, of photoreceptors in the retina, or neurons in the visual cortex). The value of this parameter corresponds to the degree to which statement 1 in the dissymetrical definition is true: "If P [a stimulus is presented] John knows that P," or to the reliability of the excitatory edge feeding into sensor B in figure 3. Critically, in the high-threshold model no similar parameter exists to control the probability of detecting the absence of a signal. In other words, the presence/absence asymmetry is expressed in the absence of a direct edge from 'stimulus absent' to a 'no' response (leftmost dashed line in Fig. 5). In this model, 'no' responses are controlled by the 'guessing' parameter g. Unlike d, the g parameter is under participants' cognitive control, and can be optimally set to maximize accuracy based on beliefs about the probability of a stimulus, the incentive structure, and critically, metacognitive beliefs about the perceptual sensitivity parameter d.

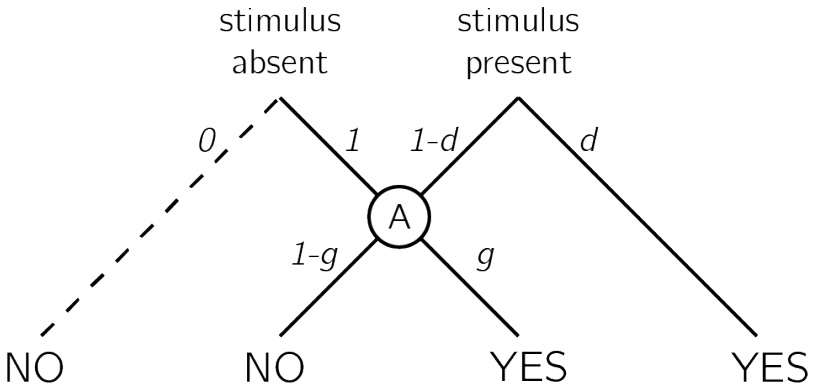


Figure 5: In discrete high-threshold models the presence of a signal can sometimes lead directly to a 'yes' response, but the absence of a signal is never sufficient to lead to a 'no' response. Agents enter node A when a stimulus is not detected. 'No' responses are then controlled by the parameter g - a 'guessing parameter' that determines the probability of responding 'yes' in case no stimulus was detected.

Given accurate knowledge about the parameter d and the prior probability of signal presence, observers can use *Bayes' rule* to extract the *negative test validity* (Oaksford & Hahn, 2004): the probability that a signal is absent, given that the they

did not perceive a signal. Formally, this equals $p(\neg T|\neg e)$, where T stands for my theory (here, a signal is present) and e for the availability of evidence (here, I can see the signal). Using Bayes' rule, this quantity is determined by the system's *correct rejection rate* ($p(\neg e|\neg T)$), *hit rate* ($p(e|T)$), and the prior probability of T. In the high threshold model, the correct rejection rate is always 1 (the threshold is never exceeded by noise alone), so the negative test validity equals:

$$p(\neg T|\neg e) = \frac{\overbrace{p(\neg e|\neg T))}^{CR}(1-p(T))}{1-p(e)} = \frac{1-p(T)}{1-p(e)} \tag{1}$$

where

$$p(e) = \overbrace{p(e|\neg T)}^{FA}(1-p(T)) + \overbrace{p(e|T)}^{Hit}p(T) = \overbrace{p(e|T)}^{Hit}p(T) \tag{2}$$

Subjects can then use the negative test validity to inform their setting of the g parameter. For example, consider a setting where you know that a target will appear on exactly half of the trials ($p(T) = 0.5$), and that half of the targets will be detected ($p(e|T) = 0.5$). Using the above formula, and given that in the high-threshold model $p(e|\neg T) = 0$, you can conclude that $p(\neg T|\neg e) = \frac{1-0.5}{1-0.5*0.5} = \frac{2}{3}$. In other words, given that a target was not detected, it is twice as likely that no target was present than that a target was present. This information can now be used to inform your setting of the g parameter before the next experimental trial.

Signal Detection Theory

Given its simplicity, the high-threshold model is useful for demonstrating the utility of self-knowledge for inference about absence. Without veridical knowledge about the sensitivity parameter d, subjects cannot tell whether they can rely on the absence of evidence when making inference about the absence of a stimulus. Continuous and graded models of perception based on Signal Detection Theory (SDT) express the same asymmetrical nature of presence/absence judgments, where clear evidence can be available for presence but less so for absence (see appendix A for an overview of Signal Detection Theory). In signal detection terms, this is expressed as high between-trial variance in sensory strength when a signal is present, but low variance when a signal is absent (see Fig. 6). Here, instead of controlling the parameter g, participants control the placement of a decision criterion. Only trials in which the sensory signal (also termed perceptual evidence, or decision variable) exceeds this criterion will be classified as 'stimulus present' trials. Optimal positioning of the criterion is dependent on beliefs about the likelihood of a stimulus to be present, as well as the spread of the signal and noise distributions and the distance between them [the stimulus-conditional *Probability Density Functions*; Gold & Shadlen (2001)]. Due to the unequal-variance structure, sensory strength in trials where a stimulus is present will be on average farther from the decision criterion compared to when no stimulus is present. As a result, similar to the setting of the g parameter in the high-threshold model, the exact placement of the SDT decision criterion will affect accuracy more when a stimulus is absent, compared to when a stimulus is present.

Common to both frameworks is the reliance on knowledge about one's own perception (the d parameter in the first case, the shape and position of the sensory distributions in the second) for optimally setting a heuristic for response on trials in which no clear evidence is available for the presence of a signal. As a result, these models draw a strong link between participants' beliefs about their own perception and their behaviour on target-absent trials. In what follows I provide empirical examples for how humans make inference about the absence of objects and memories, and link those examples to the core idea, that inference about absence critically relies on access to a self-model.

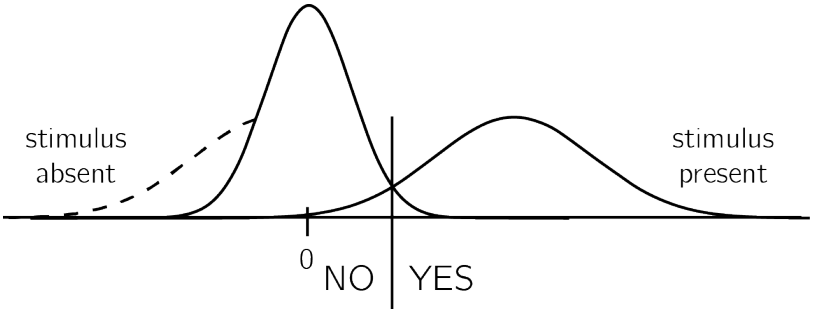


Figure 6: In unequal-variance SDT models, decisions are made based on the relative position of the sensory sample to a decision criterion. The presense/absence asymmetry manifests in the fact that only in some 'target-present' trials, but not in 'target-absent' trials, the sensory sample falls far away from the decision criterion. The dashed line represents the missing tail of the noise distribution: without it, definite evidence is sometimes available for presence, but never for absence.

0.3 Detection: "I would have noticed it"

We start our exploration of inference about absence in cognition with perhaps the most basic of psychophysical tasks - visual detection. In visual detection, participants report the presence or absence of a target stimulus, commonly presented near perceptual threshold. In such tasks, accuracy alone cannot reveal a difference in processing between decisions about presence and decisions about absence, because task accuracy is a function of both 'yes' and 'no' responses.

However, when asked to report how confident they are in their decision, subjective confidence reports reveal a metacognitive asymmetry between judgments about presence and absence. Decisions about target absence are accompanied by lower confidence, even for correctly rejected 'stimulus absence' trials (Kanai, Walsh, & Tseng, 2010; Mazor, Friston, & Fleming, 2020; Meuwese, Loon, Lamme, & Fahrenfort, 2014). Put differently, often participants cannot tell if they missed an existing target, or correctly perceived the absence of a target.

For example, in a study by Meuwese, Loon, Lamme, & Fahrenfort (2014), participants were asked to rate their confidence after performing either a perceptual detection

task ("Was there an animal present?") or a categorization task ("Was the animal a bird?"). Stimuli were identical for the two conditions, apart from phase-scrambled 'noise' images that were only shown on detection blocks. Metacognitive sensitivity was quantified as the area under the response-conditional type-II receiver-operating characteristic curve (AUROC2; see Appendix A.3). This measure reflects the agreement between confidence ratings and objective accuracy. AUROC2 was higher for the categorization than for the detection task even when performance on the primary tasks was equated. This difference originated from degraded metacognitive ability for trials in which the subjects reported not detecting an animal. More specifically, it was driven by lower confidence ratings for correct rejection trials rather than high confidence ratings for misses.

These and similar observations of a metacognitive disadvantage for inference about absence (Kanai, Walsh, & Tseng, 2010; Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018; Mazor, Friston, & Fleming, 2020; Meuwese, Loon, Lamme, & Fahrenfort, 2014), as well as a similar pattern in response times [decisions about absence tend to be slower than decisions about presence; Mazor, Friston, & Fleming (2020)] fit well with the high-threshold and unequal-variance SDT models described above. Only in a subset of target-present trials, but in no target-absent trial, can participants make a decision without deliberation (without passing in the A node in the high-threshold model, or based on a sample very far from the decision criterion in unequal-variance SDT). On these trials, participants can be highly confident in that a target was present – more confident than when deciding that a target was present after deliberation. These high-confidence trials will only be available when a target is indeed present, giving rise to a metacognitive disadvantage for inference about absence.

In line with a central role for self-monitoring in inference about absence, the lower metacognitive sensitivity for 'stimulus absence' judgments diminishes or reverses when targets are masked from awareness by means of an attentional manipulation (Kanai, Walsh, & Tseng, 2010; Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018). For example, when an attentional-blink paradigm is used to control stimulus visibility, participants are significantly more confident in their correct rejection trials than in their misses. What is it in attentional manipulations that improves participants' metacognitive insight into their judgments about stimulus absence? One compelling possibility is that a blockage of sensory information at the perceptual stage is not accessible to awareness [and is thus phenomenally transparent; Metzinger (2003)], whereas fluctuations in attention are accessible to introspection [and are thus phenomenally opaque; Limanowski & Friston (2018)]. This monitoring of one's attention state makes it possible to use premises such as "I would not have missed the target" in rating confidence in absence under attentional, but not under perceptual manipulations of visibility. Put in more formal terms, attentional manipulations increase metacognitive access to the likelihood function going from world-states to perceptual states, thereby allowing trial-to-trial tuning of the decision criterion or the g parameter.

Studies contrasting detection responses and confidence ratings under different levels of attention provide more support for this metacognitive account of detection 'no' responses. For example, participants are more likely to report the absence of a target in a specific location if their attention was directed to this location before

stimulus onset, compared to when their attention was directed to a different location (Rahnev et al., 2011). Similarly, participants are more likely to correctly report the absence of a target embedded in a stimulus (for example, a grating embedded in noise) when the stimulus is presented at the center of their visual field, compared to the periphery (Odegaard, Chang, Lau, & Cheung, 2018; Solovey, Graney, & Lau, 2015). Note that both effects are the exact opposite of what is expected based on that attention boosts sensory gain (Parr & Friston, 2019), because an increase in sensory gain without a change to the decision criterion would make false alarms, not correct rejections, more prevalent. They are however consistent with the idea that participants deploy a metacognitive strategy, shifting their decision criterion to accord with the expected strength of evidence given their current attentional state. If participants overestimate the effect of attention on their visual sensitivity, decision criterion, as measured in Signal Detection Theory, will be higher for attended versus unattended stimuli (see Fig. 7). Indeed, detection criterion is typically found to be lower for unattended stimuli (Odegaard, Chang, Lau, & Cheung, 2018; Rahnev et al., 2011; Solovey, Graney, & Lau, 2015).

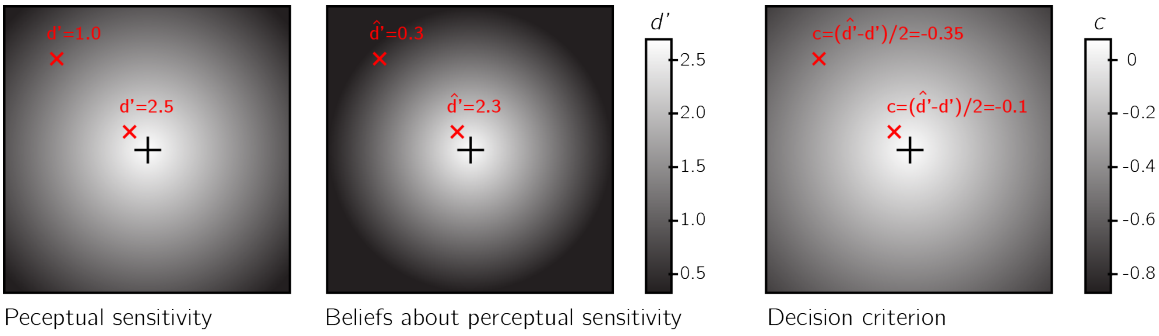


Figure 7: Left panel: Sensitivity to near-threshold stimuli is lower in the visual periphery. For example, d' equals 1.0 in top left of the screen, but is much higher near the center. Right panel: the perceptual decision criterion is lower (more 'yes' responses) in the visual periphery. Middle panel: if the effect of eccentricity on visual sensitivity is overestimated in participants' mental self-model (here d' in the top left corner is estimated to be 0.3), a lowering of the decision criterion in the visual periphery as observed in Odegaard et al. (2018) is expected.

0.4 Visual search: "I would have found it"

In visual search tasks, participants are presented with an array of stimuli and are asked to report, as quickly and accurately as possible, whether a target stimulus was present or absent in the array. Moving one step up the complexity ladder, the accumulation of information in visual search is not only a function of stimulus strength and sensory precision, but is also affected by the endogenous allocation of attention to items in an observed scene. As a result, search time varies as a function of the number of

distractors, their perceptual similarity to the target and their spatial arrangement, among other factors (for a review, see J. Wolfe & Horowitz, 2008). These factors affect not only the time taken to report the presence of a target, but also the time taken to report its absence. For example, when searching for an orange target among red and green distractors, the number of distractors has virtually no effect on search time (e.g., D'Zmura, 1991) - a phenomenon known as 'pop-out.' The bottom-up pop-out of a target can explain the immediate recognition of the presence of a target, irrespective of distractor set size. But this perceptual pop-out cannot, by itself, explain the immediate recognition of target absence, because in target absence trials there is nothing in the display to pop out.

Computational models of visual search provide different accounts for search termination in target-absent trials. In *Feature Integration Theory*, visual search comprises a pre-attentive, automatic process, and a later stage that is under participants' cognitive control. According to this model, difficult searches for a conjunction of features (*conjunction searches*, for example, searching for a purple 7 among orange and purple digits) terminate with a 'no' response once participants finished scanning all the items in the display [a *self-terminating exhaustive search*; A. M. Treisman & Gelade (1980)]. However, this model predicts that search-time variability in such conjunction target-absent trials should be lower than in conjunction target-present trials - a pattern that is not observed in empirical data (Moran, Zehetleitner, Liesefeld, Müller, & Usher, 2016; J. M. Wolfe, Palmer, & Horowitz, 2010). Furthermore, Feature Integration Theory does not provide an explicit account of target-absent responses in highly efficient parallel searches. In early versions of the *Guided Search* model, 'target

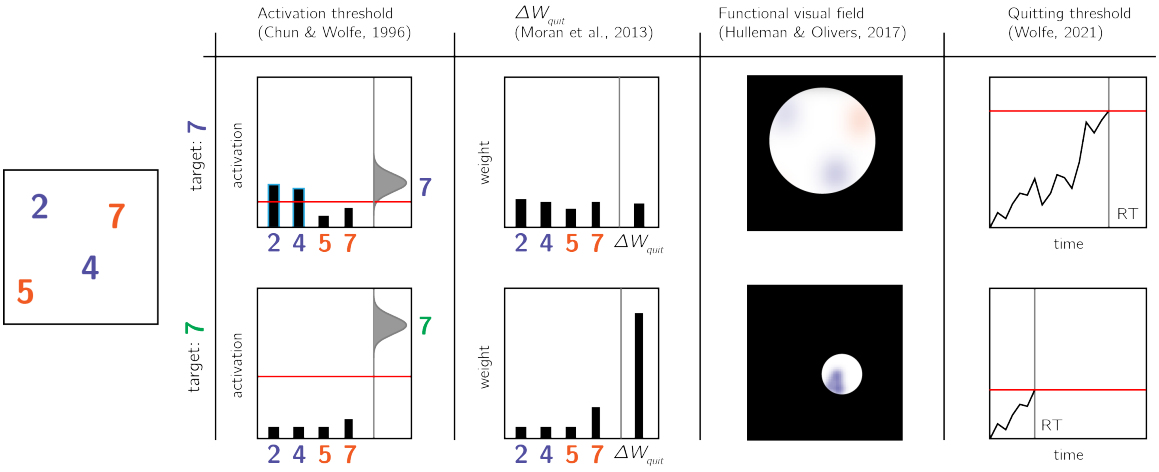


Figure 8: Models of search termination. For the same visual array (left panel) search terminated immediately for one target (a green 7, lower row), but takes longer for another target (a purple 7, upper row). Different models of visual search explain this difference by postulating search termination mechanisms that are sensitive to the counterfactual difficulty of finding a hypothetical target.

absent' judgments are the result of exhausting the search only on items that surpass a

learned 'activation threshold' (Chun & Wolfe, 1996; J. M. Wolfe, 1994). In difficult searches, the activation threshold is set to a low value, thereby requiring the scanning of multiple items before a 'no' response can be delivered. In contrast, in easy searches the activation threshold is set to a high value, reflecting a belief that a target would be highly salient (see Fig. 8). Furthermore, some very long searches are terminated once subjects concluded that "it rarely takes this long to find a target" (J. M. Wolfe, 1994).

A more recent version of the Guided Search model (*Competitive Guided Search*) described visual search as a stochastic process where items are selected for inspection based on their dynamic weight in a salience map. Critically, this model also included a *quitting unit* that is selected with a certain probability (Moran, Zehetleitner, Müller, & Usher, 2013). The search terminates once an item is recognized as the target, or once the quitting unit is selected. In this model, the salience of the quitting unit changes following the rejection of distractors. This incremental change is controlled by a parameter (Δw_{quit}) that is "under strategic control of the observer." For difficult searches, this parameter can be set to a low value, so that more items can be scanned before search termination. In very easy 'pop-out' searches this parameter can be set to a high value, making it possible to terminate a search after rejecting only one item.

In the latest formulation of the Guided Search model (J. M. Wolfe, 2021), searches are terminated once a noisy accumulator reaches a *quitting threshold*. Setting the quitting threshold high allows participants to scan more items before concluding that a target is absent. The mechanism by which participants calibrate the quitting threshold is not specified in the model.

Finally, in a fixation-based model of visual search, the number of items that are concurrently scanned within a single fixation (the *functional visual field*) is dependent on search difficulty: with more items for easy searches and less items for more difficult ones (Hulleman & Olivers, 2017).

Importantly for our point here, the activation threshold, Δw_{quit}, the quitting threshold and the functional visual field all share high similarity with the SDT criterion or the high-threshold g parameter, and are influenced by explicit or implicit beliefs about the subjective salience of a hypothetical target in the array – a form of self-knowledge.

Usually, search times in target-present and target-absent trials are highly correlated, such that if participants take longer to find the target in a given display, they will also take longer to conclude that it is absent from it (J. M. Wolfe, 1998). This alignment speaks to the accuracy of the mental self-model: participants take longer to conclude that a target is missing when they believe they would take longer to find the target, and these beliefs about hypothetical search times are generally accurate. In the two upper panels of Fig. 9 I provide two examples of cases where beliefs about search behaviour perfectly align with actual serach behaviour, leading to optimal search termination.

However, self-knowledge about attention in visual search is not always accurate. For example, when searching for an unfamiliar letter (for example, an inverted N) among familiar letters (for example, Ns), the unfamiliar letter draws immediate attention without a need for serially attending to each item in the display. However,

participants are slow in concluding that no unfamiliar letter is present, exhibiting a search time pattern consistent with a serial search for 'target absent' responses only (Wang, Cavanagh, & Green, 1994; Zhang & Onyper, 2020). In the context of my proposal here, this can be an indication for a blind-spot of the mental self-model, failing to represent the fact that an unfamiliar letter would stand out (see Fig. 9, lower panel).

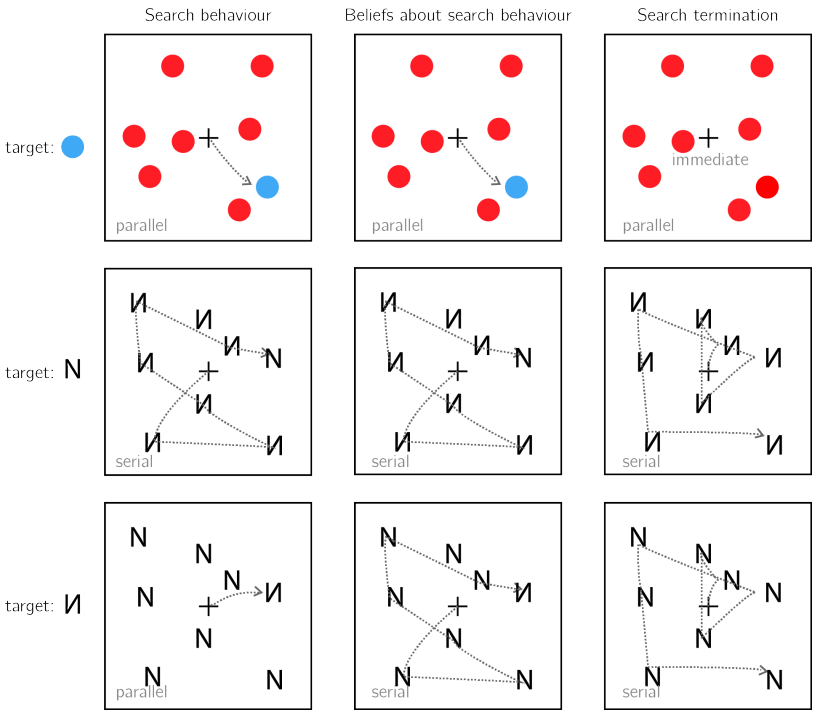


Figure 9: Upper panel: A target that is marked by a unique colour imemdiately captures attention (left). This fact is available to participants' self-model (middle). As a result, participants can immediately terminate a search when no distractor shares the color of the target (right). Middle panel: When searching for the letter N among inverted Ns, the target does not immediately capture attention, and the serial deployment of attention is necessary (left). Participants are aware of this (middle). As a result, participants perform an exhaustive serial search before concluding that a target is absent (right). Lower panel: When searching for an inverted N among canonically presented Ns, the inverted letter immediately captures attention (left). This fact is not specified in the self-model (middle). As a result, participants perform an unncessary exhaustive serial search before concluding that a target is absent (right).

0.5 Memory: "I would have remembered it"

Inference about absence not only applies to external objects (such as guavas, or visual items on the screen), but also to mental variables such as memories and thoughts. For example, upon being introduced to a new colleague, one can be certain that they have not met this person before. In the memory literature, this is known as *Negative recognition*: remembering that something did not happen (Brown, Lewis, & Monk, 1977). In the lab, a typical recognition memory experiment comprises a learning phase and a test phase. In the learning phase participants are presented with a list of items, and in the test phase they are asked to classify different items as 'old' (presented in the learning phase) or 'new' (not presented in the learning phase).

Recognition memory is often modeled using threshold or signal detection models, or a combination of the two [*Dual Process models*; Wixted (2007); Yonelinas, Dobbins, Szymanski, Dhaliwal, & King (1996)]. For example, in SDT models (Banks, 1970), participants compare a 'memory trace' against an internal criterion to determine whether the item should be classified as old or new. Like perceptual detection, the placement of the decision threshold reflects beliefs about the expected signal for old and new items. If participants believe that learned items would produce very salient memory traces, they can safely increase the decision criterion without risking mistaking old items for being new.

The role of self-knowledge in negative recognition is exemplified in the *mirror effect*: items that are more likely to be correctly endorsed as 'old' are also more likely to be correctly rejected as 'new.' In SDT terms, this effect can be described as the adjustment of the decision criterion to the expected memory trace of an item, had it been present [its *memorability*; Brown, Lewis, & Monk (1977)]. For example, Brown, Lewis, & Monk (1977) found that when asked to memorize a list of names, subjects are more confident in remembering that their own name was on the list, but also in correctly remembering when it was *not* on the list. For this effect to manifest, it is not sufficient that subjects' memory was better for their own name. They also had to know this fact, and to use it in their counterfactual thinking ("I would remember if my name were on the list").

The mirror effect has also been demonstrated for the name of one's hometown (Brown, Lewis, & Monk, 1977), for word frequency [rare words are more likely to be correctly endorsed or rejected with confidence; Brown, Lewis, & Monk (1977); Glanzer & Bowles (1976)], word imaginability (Cortese, Khanna, & Hacker, 2010; Cortese, McCarty, & Schock, 2015) and for study time [subjects are more likely to correctly classify new items as new if learned items are presented for longer; Stretch & Wixted (1998); Starns, White, & Ratcliff (2012)].

In a clever set of experiments, Strack, Förster, & Werth (2005) established a causal link from metacognitive beliefs about item memorability and decisions about the absence of memories. In two experiments, participants in one group were led to believe that high-frequency words (words that are used relatively often) are more memorable than low-frequency words, while participants in a second group were led to believe that low-frequency words were more memorable than high-frequency words. This manipulation affected participants' tendency to reject high-frequency or low-

frequency items in a later recognition-memory task. Participants who believed that high-frequency words were more memorable were more likely to classify high-frequency words as 'new,' suggesting that their metacognitive belief informed their inference about the absence of a memory ('I would have remembered this word'). Inversely, participants who believed that low frequency words were more memorable showed the opposite pattern.

One formal description of this inferential process is provided by the *likelihood ratio* rule. According to this model, subjects compare the likelihood of incoming evidence under two competing models of the world - the presence or absence of a memory trace, and choose the model under which the incoming evidence is more likely. In order to be able to compare the likelihood of an observation under alternative models, subjects must have a model of their cognition that is sufficiently detailed to yield conditional probability distributions. In experiments where the probabilities of an item to be old or new is equal, the likelihood ratio strategy is optimal (Neyman & Pearson, 1933). As a cartoon example, a participant may expect the perceived memory trace for frequent words to be centered around 0.3, and around 0.6 for infrequent words. Using the likelihood ratio rule, this participant will be more confident in that a word is new if the observed memory is 0 and the word is infrequent, compared to when the word is frequent. The likelihood ratio approach has been successful in explaining several features of recognition memory, including the mirror effect in negative recognition (Glanzer, Adams, Iverson, & Kim, 1993; Glanzer, Hilford, & Maloney, 2009).

Just like in the cases of near-threshold detection and visual search, the intuitive metacognitive knowledge behind the mirror effect may not be available for explicit report, at least not in the absence of direct experience with the task itself. In their explicit memorability reports, subjects often have little to no declarative metacognitive knowledge of which items are more likely to be remembered, even under conditions that give rise to a mirror effect. For example, although more frequent words are more likely to be forgotten (and incorrectly classified as old), participants tended to judge them as more memorable than infrequent words (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Benjamin, 2003; Greene & Thapar, 1994; Wixted, 1992). However, participants showed metacognitive insight into the negative effect of word frequency on memorability when memorability was rated after (and not before) negative recognition judgments (Benjamin, 2003; Guttentag & Carroll, 1998). Thus, the implicit metacognitive knowledge that supports accurate negative recognition may become available for explicit report only when participants introspect about their recognition attempts.

0.6 The development of a self-model

As exemplified above, the inferential processes that result in judgments of absence share important commonalities, regardless of whether it is the absence of an isolated target stimulus, of one target in an array of distractors, or of a non-physical entity such as a memory. First, in all three cases, to infer absence agents must possess some self-knowledge (under what conditions are they likely to miss a target, how long they should expect to search before finding a target in an array of distractors, or which

items are likely or unlikely to be remembered). Second, agents must be able to use this counterfactual knowledge and compare it with their current state (for example, having no recollection of an item, or not seeing a target stimulus).

At what developmental stage do humans master the necessary self knowledge and inferential machinery to make efficient and accurate inference about absence? In the context of memory, evidence suggests that the necessary self-knowledge and the capacity for counterfactual thinking exist in primary form already in early childhood, but continue to develop until adulthood. For example, children as young as 5 were able to give meaningful assessments the memorability of hypothetical life events and to use this metacognitive knowledge to inform their judgments about the nonoccurrence of an event, but this ability did not reach full maturation until the age of 9 (Ghetti & Alexander, 2004). Other studies identified a qualitative transition between the ages 7 and 8 in the ability of children to rely on expected event memorability for inference about the absence of a memory (Ghetti, Castelli, & Lyons, 2010; Ghetti, Lyons, Lazzarin, & Cornoldi, 2008). This developmental discontinuity was attributed to the development of counterfactual thinking and second-order theory of mind. Indeed, the ability to infer that something did not happen based on that it would have been remembered critically relies on one's ability to ascribe mental states to their counterfactual self.

In perception, the ability to represent absences lags behind the ability to represent presences, but reaches maturation much earlier than in the case of memory. In a study by Coldren & Haaf (2000), 4 month-old infants were familiarized with a pair of identical letters (e.g., the letter 'O'), presented side by side. In the test phase, one of the letters was replaced with a novel letter, which differed from the familiar letter either in the presence or the absence of a distinctive feature. For example, when infants that were familiarized with the letter O were tested on a display of one O and one Q, the novel letter (Q) was marked by the presence of a distinctive feature. Conversely, for infants that were familiarized with the letter Q, the novel letter O was marked by the absence of a distinctive feature. Infants showed preferential looking at the novel letter only when this letter was marked by the presence, not the absence, of a distinctive feature. A similar feature-positive effect was still evident in the learning behaviour of preschool children. When presented with two similar displays, 4 and 5 year old children were able to learn to approach the display with a distinctive feature but were at chance when trained to approach a display that is marked by the absence of a distinctive feature (Sainsbury, 1971).

Together, these results suggest that the capacity to infer the absence of physical and mental entities and the ability to use beliefs about absence to guide learning develop through infancy and early childhood. In the context of this thesis, the development of this capacity may reflect, at least in part, the gradual expansion of different aspects a mental self-model, and the development of the capacity to use this model for counterfactual reasoning. For example, a baby that is not drawn to the new letter 'O' after being habituated to the letter 'Q' may not yet represent the absence of the distinguishing feature, because they lack the implicit self knowledge to know that they would notice the lower diagonal line if it was present. More abstractly, a 7 year-old may not be able to confidently tell that they did not spread a lotion on a

chair [a highly memorable action, due to its bizarreness; Ghetti, Lyons, Lazzarin, & Cornoldi (2008)], because they lack the self-knowledge to know that if they had, they would remember doing so.

0.7 This thesis

This thesis revolves around inference about absence in perception, and its reliance on self-modeling. First, in Chapter 1 I look at inference about absence in visual search. Like detection and memory, in visual search too inference about the absence of a target item must rely on some form of self-knowledge (see section 0.4). This study sought to pinpoint the origin of this knowledge. For example, is the knowledge that some visual searches are easier than others available to subjects in everyday life, or is it learned from experience in the artificial context of performing many trials of the same visual search task again and again? Due to the typical many-trials/few-subjects structure of lab-based experiments, classical visual search studies could not tell between these alternative options. By collecting data from a large number of online participants, in this first study we were able to reliably characterize participants' search termination in the first few trials of an experiment.

In Chapter 1, Exp. 2, we found that participants gave accurate estimates of search difficulty, showing good metacognitive knowledge of key findings in the visual search literature. However, metacognitive estimates were given at the end of the experiment, allowing participants to base their estimates on their recent task experience. In Chapter 2 I asked participants to estimate search times for new search arrays, before performing these searches. Here I found that an internal model of visual search is rich and accurate, and also that it is person-specific in that it is better aligned with subjects' own search behaviour than with the search behaviour of other participants.

In Chapters 3 and 4 I looked at a different perceptual task in which participants make inference about presence and absence: near-threshold detection. Detection was compared against discrimination as a control task in which stimulus category is inferred in a symmetrical setting (for example, right versus left motion). In Chapter 3, I used reverse correlation to ask what perceptual features contribute to decision and confidence in detection and discrimination. In three experiments (one lab-based and two conducted online), I replicated the 'positive evidence bias' in discrimination confidence, and found that a similar bias exists in detection response, establishing a link between confidence in disrimination and decisions about presence and absence.

In Chapter 4 I used functional magnetic resonance imaging (fMRI) to compare brain activation in decisions about stimulus type, stimulus presence, and stimulus absence, as well as in confidence ratings in these decisions. I found a quadratic modulation of brain activity by confidence in prefrontal and parietal cortices. This modulation was stronger for detection judgments (decisions about stimulus presence or absence) than for discrimination judgments (decisions about stimulus type). Computational models of internal and external precision monitoring captured some, but not all aspects of the observed data.

Finally, in Chapter 5 I focused on three behavioural asymmetries in detection (in

confidence, response time, and metacognitive sensitivity), and asked whether similar asymmetries exist for the detection not of entire objects or stimuli, but also of stimulus parts, stimulus features, and expectation violations. The idea to look at presence and absence of sub-stimulus entities drew inspiration from reports of *visual search asymmetries*, where finding a stimulus that is marked by the presence of a feature relative to distractors is easier than finding a stimulus that is marked by the absence of a feature. Results from six pre-registered experiments indicated at least two sources of asymmetry between presence and absence, that independently contribute to differences in response time and confidence on the one hand, and in metacognitive sensitivity on the other hand.

Chapter 1

Efficient search termination without task experience: the role of second-order knowledge about visual search

Matan Mazor & Stephen M. Fleming

As a general rule, if it is easy to detect a target in a visual scene, it is also easy to detect its absence. To account for this, models of visual search explain search termination as resulting either from counterfactual reasoning over second-order representations of search efficiency, automatic extraction of ensemble statistics of a display, or heuristic adjustment of a search termination strategy based on previous trials. Traditional few-subjects/many-trials lab-based experiments render it impossible to disentangle the unique contribution of these different processes to absence pop-out - the immediate recognition that a feature is missing from a display. In two pre-registered large-scale online experiments (N1=1187, N2=887) we show that search termination times are already aligned with target identification times in the very first trials of the experiment, before any experience with target presence. Exploratory analysis reveals that second-order knowledge about search efficiency can be used to guide decisions about search termination even if it is not available for explicit report. We conclude that for basic stimulus properties, efficient inference about absence is independent of task experience, and relies instead on implicit second-order knowledge.

1.1 Introduction

Searching for the only blue letter in an array of yellow letters is easy, but searching for the only blue X in an array of yellow Xs and blue Ts is much harder (A. M. Treisman & Gelade, 1980). This difference manifests in the time taken to find the target letter, but also in the time taken to conclude that the target letter is missing. In other words, easier searches not only make it easier to detect the presence of a target, but also to infer its absence. Differences in the speed of detecting the presence

of a target have been attributed to pre-attentional mechanisms (A. M. Treisman & Gelade, 1980) and guiding signals (J. M. Wolfe, 2021; J. M. Wolfe & Gray, 2007) that can sometimes make the target item 'pop out' immediately without any attentional effort. In target-absent trials, however, there is nothing in the display to pop out. This raises a fundamental question: what makes some decisions about target absence easier than others?

Models of search termination offer three classes of answers to this question, based on counterfactual reasoning, ensemble perception, and task heuristics. According to counterfactual models, decisions about target absence are guided by prior beliefs about search efficiency ("If it were present, I would have found the red book by now"). These comprise beliefs about regularities in the environment ("it it were present, the book would have been on this shelf"), and second-order beliefs about one's own perception and attention ("the red cover would have immediately drawn my attention"). In recent versions of the Guided Search model (J. M. Wolfe, 2012, 2021), for example, search termination is triggered by a noisy quitting signal accumulator reaching a *quitting threshold*, which can be adapted to maximize long-time search efficiency, and be affected by prior second-order beliefs about the effects of set size and crowding on search difficulty (J. M. Wolfe, 2012). Similarly, in Competitive Guided Search, the probability of terminating a search is a function of several factors, including a free parameter that indexes counterfactual beliefs about finding a target, had it been present (Moran, Zehetleitner, Müller, & Usher, 2013). Finally, in a fixation-based model of visual search, the number of items that are concurrently scanned within a single fixation (the *functional visual field*) depends on the expected difficulty of finding a hypothetical target: with more items for easy searches and fewer items for more difficult ones (Hulleman & Olivers, 2017).

Ensemble perception accounts of visual search postulate that some global properties of a display can be extracted automatically and immediately, and that in some cases these global properties are sufficient to conclude that a target is absent. For example, according to Feature Integration Theory, pre-attentive activation in *feature maps* can provide participants with information about the presence or absence of a feature in the display (A. M. Treisman & Gelade, 1980). The absence of a relevant feature is then sufficient to make an immediate 'target absent' decision, without processing any individual stimulus.

Finally, heuristic-based models suggest that quitting parameters are acquired by participants as they perform a task, sometimes by following very simple rules. For example, in one model, an internal *activation threshold* decreases following incorrect and increases following correct 'no' responses (Chun & Wolfe, 1996). A higher activation threshold results in the scanning of less distractors, giving rise to shorter search times for easier searches. This simple heuristic provides an excellent fit to data from a visual search task with hundreds of trials, and does so without requiring that subjects hold any prior knowledge or expectations about search efficiency.

In traditional visual search experiments, where participants perform hundreds of trials of similar searches, it is impossible to disentangle the contributions of these three putative mechanisms to search termination. Yet, the three accounts make different predictions for the earliest trials of a visual search experiment, where participants

encounter the stimuli for the first time. In these trials, quitting time cannot reflect the adaptive adjustment of a threshold based on previous trials, or the statistical learning of regularities in the experiment. Instead, efficient search termination without task experience must rely on an immediate perception of ensemble properties of the display, prior second-order knowledge about one's own search efficiency, or a combination of both.

In two pre-registered experiments we focus on feature search for colour and shape. Focusing on the first four trials of the task, we ask whether prior experience with the task and stimuli is necessary for efficient search termination in feature searches. Unlike typical visual search experiments that comprise hundreds or thousands of trials, here we collect only a handful of trials from a large pool of online participants. This unusual design allows us to reliably identify search time patterns in the first trials of the experiment. By making sure that the first displays do not include the target stimulus, we are able to ask what knowledge is available to participants about their expected search efficiency prior to engaging with the task.

To anticipate our results, we find that efficient search termination for single features does not depend on task experience. In an exploratory analysis on a subset of participants, we further show that efficient search termination is also independent of explicit metacognitive knowledge about the task. We argue that without second-order knowledge about one's own perception and attention, ensemble perception alone is not sufficient for efficient search termination, and interpret our results as revealing a role for implicit second-order knowledge of search efficiency in search termination.

1.2 Experiment 1

In Experiment 1, we examined search termination in the case of colour search. When searching for a deviant colour, the number of distractors has virtually no effect on search time (*colour pop-out*; e.g., D'Zmura, 1991), for both 'target present' and 'target absent' responses. Here we asked whether efficient quitting in colour search (*color absence pop-out*) is dependent on task experience. A detailed pre-registration document for Experiment 1 can be accessed at osf.io/yh82v/.

1.2.1 Participants

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 1187 Participants (median reported age: 33; range: [18-81]) were recruited via Prolific, and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. Following our pre-registration, we collected data until we reached 320 included participants for each of our pre-registered hypotheses (after applying our pre-registered exclusion criteria). The entire experiment took around 3 minutes to complete (median completion time: 3.19 minutes). Participants were paid £0.38 for their participation, equivalent to an hourly wage of £ 7.14.

1.2.2 Procedure

A static version of Experiment 1 can be accessed on matanmazor.github.io/termination. Participants were first instructed about the visual search task. Specifically, that their task is to report, as accurately and quickly as possible, whether a target stimulus was present (press 'J') or absent (press 'F'). Then, practice trials were delivered, in which the target stimulus was a rotated *T*, and distractors rotated *L*s. The purpose of the practice trials was to familiarize participants with the structure of the task. For these practice trials the number of items was always 3. Practice trials were delivered in short blocks of 6 trials each, and the main part of the experiment started only once participants responded correctly on at least five trials in a block (see Figure 1.1). In the main part of the experiment, participants searched for a red dot among blue dots or a mixed array of blue dots and red squares. Set size was set to 4 or 8, resulting in a 2-by-2 design (search type: color or color×shape, by set size: 4 or 8). Critically, and unbeknown to subjects, the first four trials were always target-absent trials (one of each set-size × search-type combination), presented in randomized order. These trials were followed by the four corresponding target-present trials, presented in randomized order. The final four trials were again target-absent trials, presented in randomized order.

1.2.3 Randomization

The order and timing of experimental events was determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized in a way that ensures registration time-locking (Mazor, Mazor, & Mukamel, 2019).

1.2.4 Data analysis

Rejection criteria

Participants were excluded for making more than one error in the main part of the experiment, or for having extremely fast or slow reaction times in one or more of the tasks (below 250 milliseconds or above 5 seconds in more than 25% of the trials).

Error trials, and trials with response times below 250 milliseconds or above 1 second were excluded from the response-time analysis. All pre-registered analyses without RT-based exclusion are reported in appendix B.

Data preprocessing

To control for within-block trial order effects, a linear regression model was fitted separately for each block and participant, predicting search time as a function of trial serial order within the block ($RT \sim \beta_0 + \beta_1 i$, with i denoting the mean-centered serial position within a block). Search times were corrected by subtracting the product of the slope and the mean-centered serial position, in a block-wise manner.

Subject-wise search slopes were then extracted for each combination of search type (color or conjunction) and block number by fitting a linear regression model to the

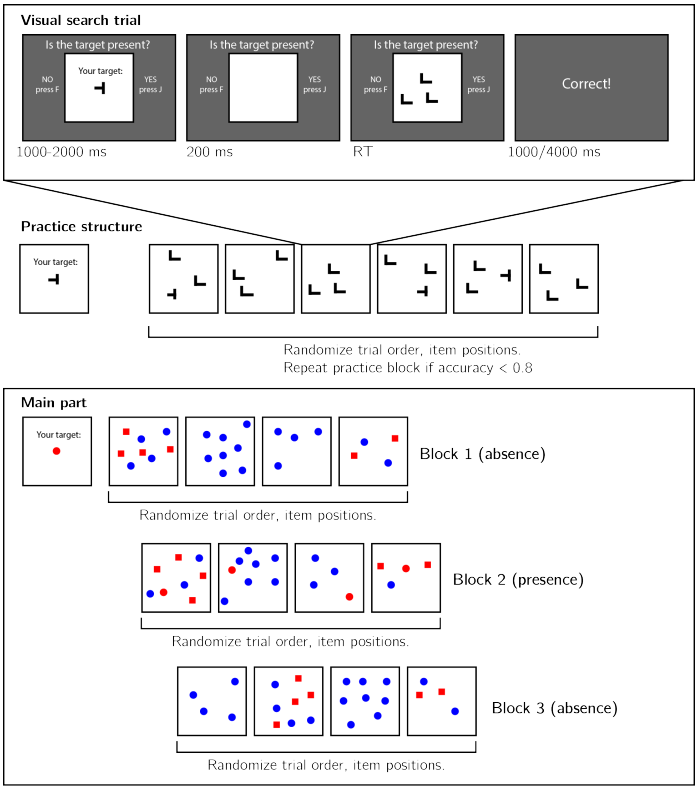


Figure 1.1: Experimental design. Top panel: each visual search trial started with a screen indicating the target stimulus. The search display remained visible until a response was recorded. To motivate accurate responses, the feedback screen remained visible for one second following correct responses and for four seconds following errors. Middle panel: after reading the instructions, participants practiced the visual search task in blocks of 6 trials, until they had reached an accuracy level of 0.83 correct or higher (at most one error in a block of 6 trials). Bottom panel: the main part of the experiment comprised 12 trials only, in which the target was a red dot. Unbeknown to subjects, only trials 5-8 (Block 2) were target-present trials, and the remaining trials were target-absent trials. Each 4-trial block followed a 2 by 2 design, with factors being set size (4 or 8) and distractor type (color or conjunction; blue dots only or blue dots and red squares, respectively).

reaction time data with one intercept and one set-size term.

Hypotheses and analysis plan

Experiment 1 was designed to test several hypotheses about the contribution of metacognitive knowledge to search termination, the state of this knowledge prior to engaging with the task, and the effect of experience on this metacognitive knowledge. The specifics of our pre-registered analysis can be accessed in the following link:

https://osf.io/ea385. We outline some possible search time patterns and their pre-registered interpretation in Fig. 1.2. Analysis comprised a positive control based

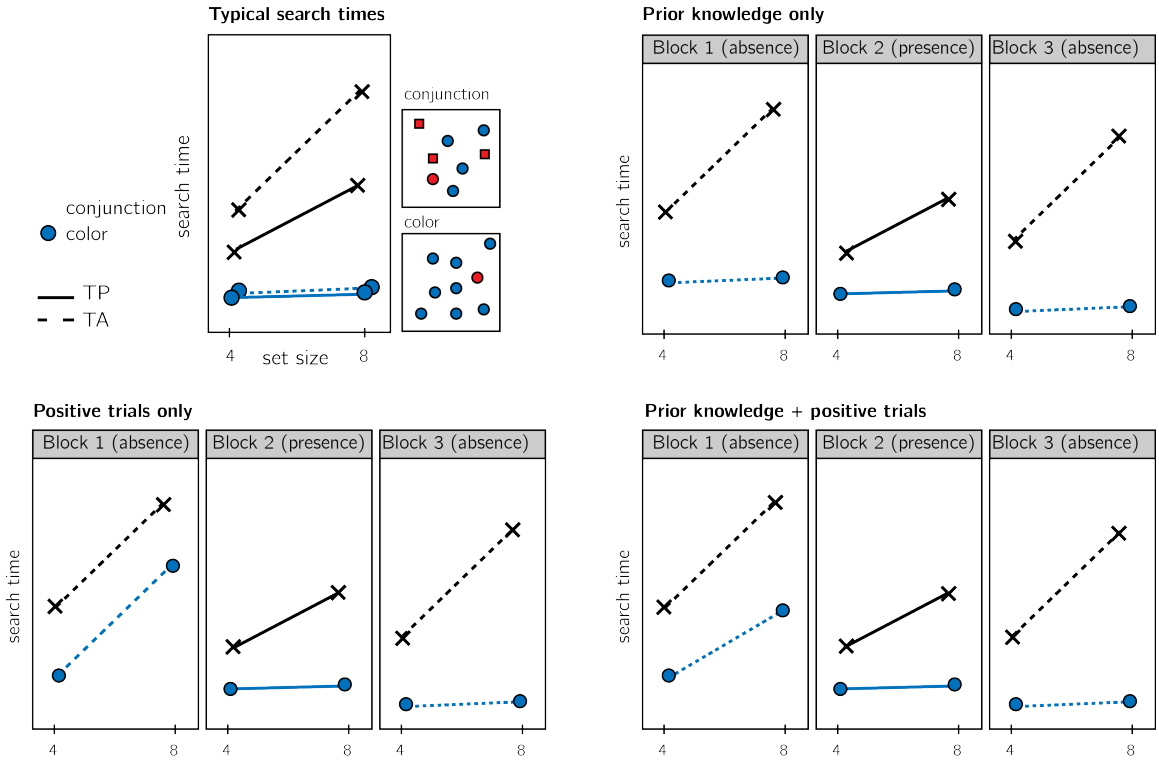


Figure 1.2: Visualization of Hypotheses. Top left: typical search times in visual search experiments with many trials (where TP = Target Present responses; TA = Target Absent responses). Set size (x axis) affects search time in conjunction search, but much less so in color search. However, it is unclear whether this pattern also holds in the first target-absent trials in an experiment. Different models make different predictions about target-absent search times in the first block of the experiment. Top right: one possibility is that the same qualitative pattern will be observed in our design, with an overall decrease in response time as a function of trial number. This would suggest that the second-order knowledge necessary to support efficient inference about absence was already in place before engaging with the task. Bottom left: an alternative pattern is that the same qualitative pattern will be observed for blocks 2 and 3, but not in block 1. This would suggest that for inference about absence to be efficient, participants had to first experience some target-present trials. Bottom right: alternatively, some degree of second-order knowledge may be available prior to engaging with the task, with some being acquired by subsequent exposure to target-present trials. This would manifest as different slopes for conjunction and color searches in blocks 1 and a learning effect for color search between blocks 1 and 3.

on target-present trials, a test of the presence of a pop-out effect for target-absent color search in block 1, and a test for the change in slope for target-absent color search between blocks 1 and 3. All hypotheses were tested using a within-subject t-test, with a significance level of 0.05. Given the fact that we only have one trial per cell, one excluded trial is sufficient to make some hypotheses impossible to test on a given participant. For this reason, for each hypothesis separately, participants were included only if all necessary trials met our inclusion criteria. This meant that some hypotheses were tested on different subsets of participants.

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. We used R [Version 4.0.5; R Core Team (2019)] and the R-packages *BayesFactor* [Version 0.9.12.4.2; Richard D. Morey & Rouder (2018)], *cowplot* [Version 1.1.1; Wilke (2019)], *dplyr* [Version 1.0.7; Wickham, François, Henry, & Müller (2020)], *ggplot2* [Version 3.3.5; Wickham (2016)], *jsonlite* [Version 1.7.2; Ooms (2014)], *lsr* [Version 0.5; Navarro (2015)], *MESS* [Version 0.5.7; Ekstrøm (2019)], *papaja* [Version 0.1.0.9997; Aust & Barth (2020)], *pwr* [Version 1.3.0; Champely (2020)], *reticulate* [Version 1.20; Ushey, Allaire, & Tang (2020)], and *tidyr* [Version 1.1.3; Wickham & Henry (2020)] for all our analyses. A detailed pre-registration document for Experiment 1 can be accessed at osf.io/yh82v/. All analysis scripts and anonymized data are available at github.com/matanmazor/termination.

1.2.5 Results

Overall mean accuracy was 0.95 (standard deviation =0.06). Median reaction time was 623.98 ms (median absolute deviation = 127.37). In all further analyses, only correct trials with response times between 250 and 1000 ms are included.

Hypothesis 1 (positive control): Search times in block 2 (target-present) followed the expected pattern, with a steep slope for conjunction search ($M = 12.52$, 95% CI [10.08, 14.95]) and a shallow slope for color search ($M = 3.91$, 95% CI [2.13, 5.70]; see middle panel in Fig. 1.3A). Color search slope was significantly lower than 10 ms/item and thus met our criterion for being considered 'pop-out' ($t(961) = -6.69$, $p < .001$). Furthermore, the difference between the slopes was significant ($t(749) = 6.50$, $p < .001$). This positive control served to validate our method of using two trials per participant for obtaining reliable group-level estimates of search slopes.

Hypothesis 2: Our central focus was on results from block 1 (target-absent). Here participants didn't yet have experience with searching for the red dot. Similar to the second block, conjunction search slope was steep ($M = 18.41$, 95% CI [14.95, 21.87]). A clear pop-out effect for color absence was also evident ($M = 0.15$, 95% CI $[-\infty, 2.31]$, $t(886) = -7.51$, $p < .001$). Furthermore, the average search slope for color search in this first block was significantly different from that of the conjunction search ($t(413) = 6.55$, $p < .001$; see leftmost panel in Fig. 1.3A), indicating that a color-absence pop-out is already in place prior to direct task experience. This result is

in line with the *prior-knowledge only* model (see Fig. 1.2), in which participants have valid expectations for efficient color search, prior to engaging with a task.

Pre-registered hypotheses 3-5 were designed to test for a learning effect between blocks 1 and 3, before and after experience with observing a red target among blue distractors. Given the overwhelming pop-out effect for target-absent trials in block 1, not much room for additional learning remained. Indeed, results from these tests support a prior-knowledge only model.

Hypothesis 3: Like in the first block, in the third block color search complied with our criterion for 'pop-out' ($M = 2.27$, 95% CI $[-\infty, 3.86]$, $t(979) = -7.98$, $p < .001$), and was significantly different from the conjunction search slope ($t(745) = 11.16$, $p < .001$; see rightmost panel in Fig. 1.3A). This result is not surprising, given that a pop-out effect was already observed in block 1.

Hypothesis 4: To quantify the learning effect for color search, we directly contrasted the search slope for color search in blocks 1 and 3. We find no evidence for a learning effect ($t(799) = -1.15$, $p = .250$). Furthermore, a Bayesian t-test with a scaled Cauchy prior for effect sizes (r=0.707) provided strong evidence in favour of the absence of a learning effect ($BF_{01} = 12.98$).

Hypothesis 5: In case of a learning effect for pop-out search, Hypothesis 5 was designed to test the specificity of this effect to color pop-out by computing an interaction between block number and search type. Given that no learning effect was observed, this test makes little sense. For completeness, we report that the change in slope between blocks 1 and 3 was similar for color and conjunction search ($M = -3.58$, 95% CI $[-10.52, 3.36]$, $t(320) = -1.01$, $p = .311$).

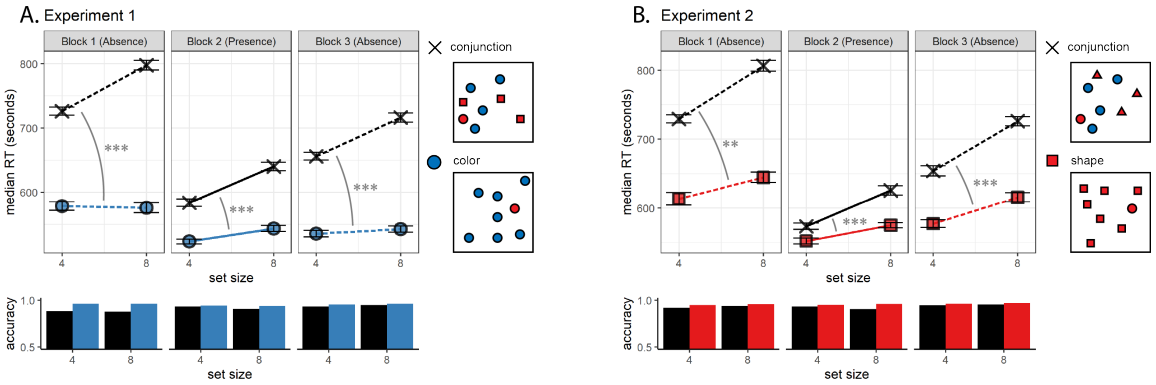


Figure 1.3: Main Results for Expeiments 1 (A) and 2 (B). Upper panel: median search time by distractor set size for the two search tasks across the three blocks (12 trials per participant). Correct responses only. Lower panel: accuracy as a function of block, set size and search type. Error bars represent the standard error of the median (estimated with bootstrapping). Significance stars correspond to the difference in slope between conjunction and feature search within a block. *: p<0.5, * * : p<0.01, * * * : p<0.001

1.2.6 Additional analysis: first trial only

We considered the possibility that our results do not reflect true absence pop-out without task experience, but instead might reflect participants' ability to rapidly adjust their termination times based on feedback from previous trials, even within the four trials of the first block. To rule out such within-block learning effects, we tested whether participants showed a color-absence pop-out effect on the very first trial of the experiment. To this end, we analyzed first trial response times as a function of search type (conjunction or color) and set-size. Since these first trials were slower overall (median RT in the first trial: 881.30 ms compared to 630.34 ms in the last trial), for this exploratory analysis we did not exclude trials based on response times.

Even in this between-subject analysis, with only one trial per participant, we found a significant positive search slope for conjunction search (23.31 ms/item, $p < 0.01$), but not for color search (-5.13 ms/item, $p = .43$; note that this negative slope is not apparent in Fig. 1.4A because the figure presents median reaction times, rather than means). The difference in slopes between conjunction and color, quantified as the interaction between set size and search type in a two-way between-subject analysis of variance, was also significant ($F(1, 1, 041) = 6.74$, $MSE = 466,761.60$, $p = .010$, $\hat{\eta}^2_G = .006$; see Fig. 1.4A). In other words, a color-absence pop-out was already detectable in the very first trial of the experiment.

1.3 Experiment 2

Experiment 1 provided unequivocal evidence that color-absence pop-out occurs prior to experiencing color pop-out in the context of the same task. Experiment 2 was designed to extend these findings to another stimulus feature that is also found to efficiently guide attention: shape. Unlike colour space, which spans three dimensions only, the space of possible shapes is relatively unconstrained such that having prior knowledge of the expected effect of different shapes on attention might require a richer mental model of attentional processes. Furthermore, colour is agreed to be a 'guiding attribute of attention,' while it is unclear which shape features guide attention (J. M. Wolfe & Horowitz, 2017). In this experiment we also included an additional control for prior experience with visual search tasks, and asked if knowledge about search efficiency is available for explicit metacognitive report.

1.3.1 Participants

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 887 participants (median reported age: 33; range: [18-75]) were recruited via Prolific, and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. We collected data until we reached 320 included participants for hypotheses 1-4 (after applying our pre-registered exclusion criteria). The entire experiment took around 4 minutes to

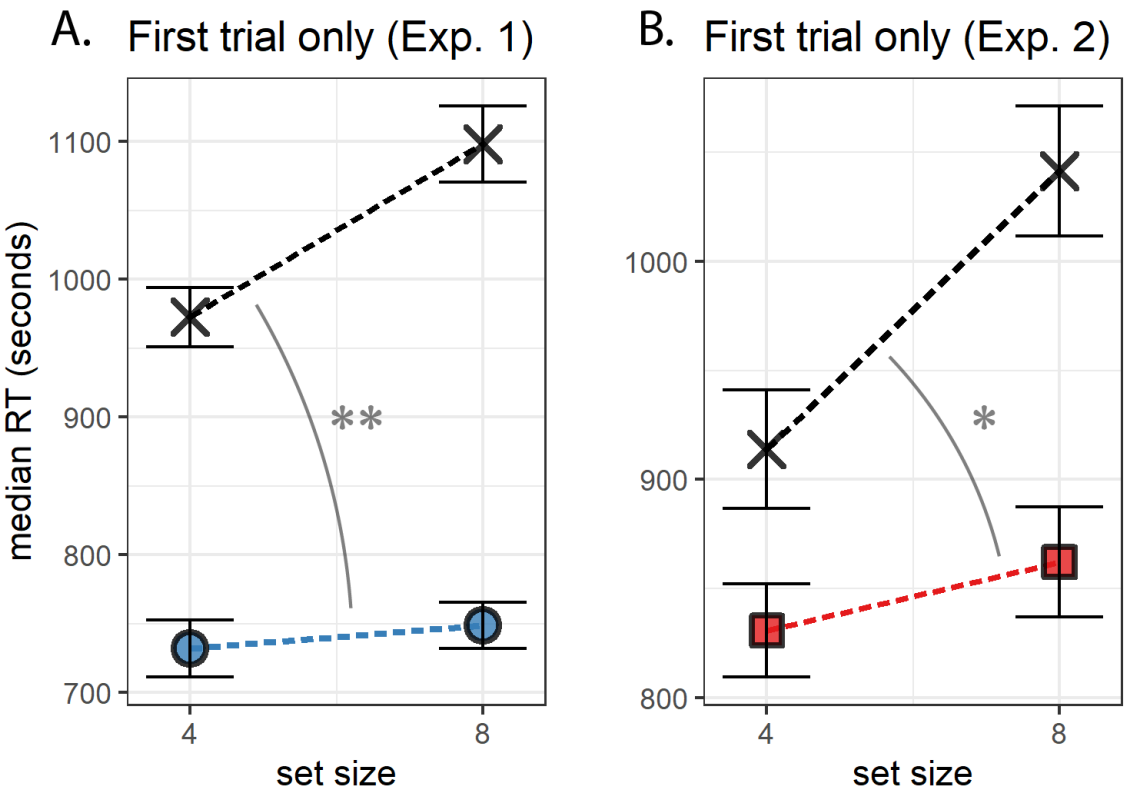


Figure 1.4: Median search time by distractor set size for Experiments 1 and 2, looking at the first trial of each participant only. Same conventions as in Fig. 1.3.

complete (median completion time in our pilot data: 3.93 minutes). Participants were paid £0.51 for their participation, equivalent to an hourly wage of £7.78.

1.3.2 Procedure

A static version of Experiment 2 can be accessed on matanmazor.github.io/termination. Experiment 2 was identical to Experiment 1 with the following exceptions. First, instead of color search trials, we included shape search trials, where the red dot target is present or absent in an array of red squares. Second, to minimize the similarity between conjunction and shape searches, conjunction trials included blue dots and red triangles as distractors. Third, to test participants' explicit metacognition about their visual search behaviour, upon completing the main part of the task participants were presented with the four target-absent displays (shape and conjunction displays with 4 or 8 items), and were asked to sort them from fastest to slowest. Finally, participants reported whether they had participated in a similar experiment before, where they were asked to search for shapes on the screen. Participants who responded 'yes' were asked to tell us more about this previous experiment. This question was included in order to examine whether efficient target-absent search in trial 1 reflects prior experience with similar visual search experiments.

Our pre-registered analysis plan for Experiment 2, including rejection criteria and data preprocessing, was identical to our analysis plan for Experiment 1, and can be accessed in the following link: https://osf.io/v6mnb.

1.3.3 Results

Overall mean accuracy was 0.96 (standard deviation =0.06). Median reaction time was 644.60 ms (median absolute deviation = 123.89). In all further analyses, only correct trials with response times between 250 and 1000 ms are included.

Hypothesis 1 (positive control): Search times in block 2 (target-present) followed the expected pattern, with a steep slope for conjunction search (M = 15.08, 95% CI [12.34, 17.83]) and a shallow slope for shape search (M = 5.84, 95% CI [3.90, 7.78]; see middle panel of Fig. 1.3B). The slope for shape search was significantly lower than 10 ms/item and thus met our criterion for being considered 'pop-out' ($t(754) = -4.21$, $p < .001$). Furthermore, the difference between the slopes was significant ($t(584) = 4.98$, $p < .001$).

Hypothesis 2: Our central focus was on results from block 1 (target-absent). Here participants didn't yet have experience with finding the red dot. Similar to the second block, the slope for conjunction search was steep (M = 19.53, 95% CI [16.03, 23.04]). The slope for shape search was numerically lower than 10 ms/item, but not significantly so (M = 8.03, 95% CI [$-\infty$, 10.50], $t(608) = -1.31$, $p = .095$). Still, the average search slope for shape search in this first block was significantly different from that of the conjunction search ($t(326) = 2.77$, $p = .006$; see leftmost panel of Fig. 1.3B), indicating that a processing advantage for detecting the absence of a shape compared to the absence of shape-color conjunction was already in place before experience with target presence.

Moreover, this processing advantage was not different from what is expected based on shape search slope in block 2 (target presence). A conservative estimate for the ratio between target absence and target presence search slopes is 2 (J. M. Wolfe, 1998). Based on this ratio of 2 and the observed target-presence search slope of 6 ms/item, target absence search slope is expected to be 12 ms/item, or higher. Indeed, search slope for shape absence was not significantly different from, and numerically lower than, twice the search slope for shape presence as measured in block 2 ($t(548) = -1.16$, $p = .246$; $BF_{01} = 10.66$). In other words, our failure to find a pop-out effect for shape absence was not due to participants being suboptimal in their quitting times, but because finding a red dot among red squares is truly more difficult than finding a red dot among blue dots.

Hypothesis 3: As in the first block, in the third block the slope for shape search was numerically lower than 10 ms/item, but not significantly so (M = 8.85, 95% CI [$-\infty$, 10.68], $t(723) = -1.03$, $p = .151$). Importantly, the slope for shape search in block 3 was significantly different from the slope for conjunction search ($t(565) = 6.02$, $p < .001$) and not significantly different from twice the search slope for shape presence ($t(653) = 1.04$, $p = .299$; $BF_{01} = 13.29$; see rightmost panel of Fig. 1.3B).

Hypothesis 4: To quantify a potential learning effect for shape search between blocks 1 and 3, we directly contrasted the search slope for shape search in these two

'target-absent' blocks. We find no evidence for a learning effect ($t(542) = -0.03$, $p = .974$). Furthermore, a Bayesian t-test with a scaled Cauchy prior for effect sizes (r=0.707) provided strong evidence against a learning effect ($BF_{01} = 20.72$). Like in Experiment 1, these results are most consistent with a *prior-knowledge only* model (see Fig. 1.2), in which participants already know to expect that shape search should be easier than conjunction search, prior to having direct experience with target-present trials.

1.3.4 Additional Analyses

First trial only

As in Exp. 1, here we also extended our pre-registered analysis with an exploratory between-subject analysis, focusing on the first trial of the experiment. Here too, we observed a significant positive search slope for conjunction search (43.65 ms/item, $p < 0.001$), but not for shape search (9.80 ms/item, $p = .40$). The difference in slopes between conjunction and shape, quantified as the interaction between set size and search type in a two-way betwee-subject analysis of variance, was significant ($F(1, 781) = 4.25$, $MSE = 209,989.78$, $p = .040$, $\hat{\eta}^2_G = .005$; see Fig. 1.4B). This result reveals that efficient recognition of shape absence is already detectable in the very first trial of the experiment.

Task experience

At the end of the experiment, participants were asked if they have ever participated in a similar experiment before, where they were asked to search for a target item. 796 out of 887 participants answered 'no' to this question. For those participants, a highly efficient search for a distinct shape in the first trials of the experiment, if found, cannot be due to prior experience of performing a visual search task with similar stimuli. Notably, however, participants who reported having no prior experience with a visual search task still showed efficient search termination for shape distractors ($M = 7.32$, 95% CI [4.21, 10.43]), and were significantly more efficient in terminating shape search than conjunction search in the first 4 target-absent trials ($t(296) = 2.68$, $p = .008$). Efficient search termination for shape search is therefore not dependent on prior visual search trials, neither within the same experiment nor in previous ones.

Search time estimates

Upon completing the main part of Experiment 2, participants positioned the four search arrays (shape and conjunction searches with 4 or 8 distractors) on a perceived difficulty axis (see Fig. 1.5A). We used these difficulty ratings to ask whether the advantage for detecting the absence of a distinct shape over the absence of a shape/color conjunction depended on explicit access to metacognitive knowledge about search difficulty. The decision to quit early in shape-absent trials may depend on an internal belief that the target shape would have drawn attention immediately, but this belief may be inaccessible to introspection. If introspective access is not a necessary condition

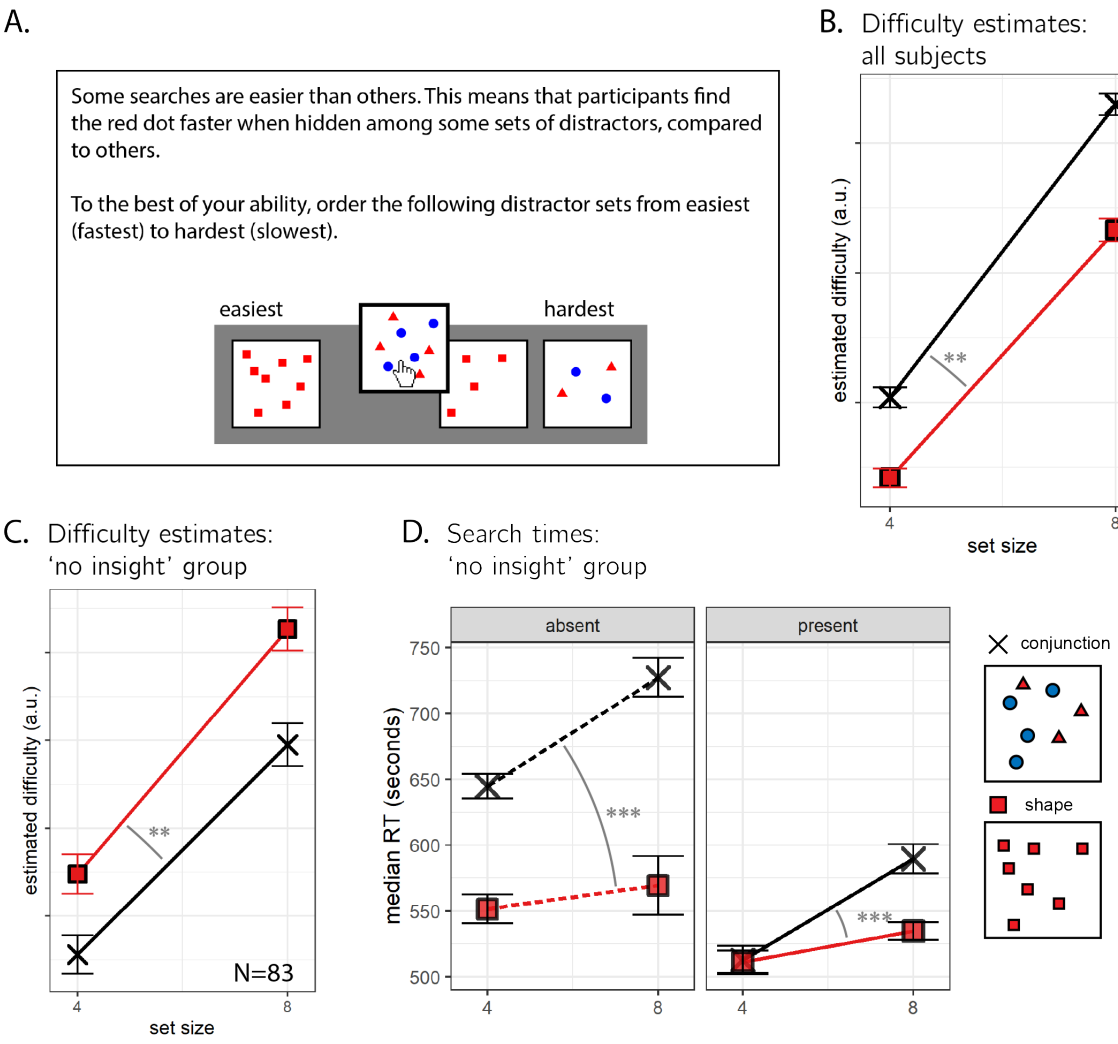


Figure 1.5: A: After completing the visual search component of Experiment 2, participants were asked to position the four searches (shape and conjunction searches with 4 or 8 distractors) on a perceived difficulty axis. B. As a group, participants' estimates revealed metacognitive knowledge of the set size effect and of the fact that shape search is harder. C. A subset of 84 participants erroneously believed that shape search was more difficult than conjunction search. D. Even among these participants, search slopes in target-absence blocks followed the typical pattern, with a steeper slope for conjunction search. Same plotting conventions as Fig. 1.3.

for efficient quitting in visual search, some participants may not be able to reliably introspect about the difficulty of different searches but still be able to quit efficiently in shape search.

For this analysis, we only considered the ratings of participants who engaged with the array-sorting trial, and moved some of the arrays before continuing to the next trial

(N=789). Searches with 8 distractors were rated as more difficult than searches with 4 distractors, in line with the set-size effect ($t(788) = 31.62$, $p < .001$). Furthermore, conjunction searches were rated as more difficult than shape searches ($t(788) = 5.11$, $p < .001$). Finally, we fitted single-subject linear regression models to the two search types, predicting search-time estimates (the position of each condition on a continuous perceived difficulty scale) as a function of set size. Similar to actual search slopes, these slopes derived from subjective estimates were also shallower for shape than for conjunction search, reflecting a belief that the effect of set size in shape search is not as strong as the effect of set size in conjunction search ($M = 6.45$, 95% CI [2.81, 10.08], $t(788) = 3.48$, $p = .001$; see Fig. 1.5B).

Subjective search time estimates revealed that by the end of the experiment, the average participant considered the slope of shape search to be shallower than that of conjunction search. This suggests that at least some participants had introspective access to their visual search behaviour. But were those participants whose estimates reflected a shallow slope for shape search the same ones that were more efficient in detecting the absence of a shape in the display? The slopes of retrospective estimates for shape search were not reliably correlated with actual search slopes for shape absence in block 1 ($r = .08$, 95% CI [−.06, .22]) or 2 ($r = .02$, 95% CI [−.12, .16]). However, this result should be interpreted carefully in light of the low reliability of single subject estimates that are derived from one trial per cell. Indeed, search slopes for shape absence in blocks 1 and 3 were not reliably correlated themselves ($r = .05$, 95% CI [−.10, .19]).

To answer this question using a more severe test (Mayo, 2018), we focused on the subset of participants whose difficulty orderings reflected the erroneous belief that shape search was more difficult than conjunction search ($N = 83$; see Fig. 1.5C). If efficient search termination depends on accurate explicit metacognitive knowledge about search efficiency, search termination in this subset of participants is not expected to be more efficient in shape compared to conjunction search, and is even expected to show the opposite pattern. In contrast with this prediction, search slopes for shape-absence trials were shallower than for conjunction-absence trials ($M_d = 12.45$, 95% CI [5.21, 19.69], $t(82) = 3.42$, $p = .001$; see Fig. 1.5D). This indicates that efficient identification of shape absence is not dependent on explicit metacognitive knowledge about search efficiency.

1.4 Discussion

How do people decide that a target is absent from a visual scene? In this study we considered three candidate answers to this question: counterfactual reasoning ("I would have detected the target if it were present"), ensemble perception ("I immediately see that the target is missing") and task heuristics ("Based on previous trials, responding now would balance accuracy and response time"). The third option is different from the first two: while a heuristic calibration of a termination rule may shape search behaviour in classic lab-based experiments comprised of many repetitive trials, it is not available to subjects in one-shot searches in their everyday lives, nor is it available

to them in the first trial of the experiment.

To isolate the effect of previous trials on search termination, we focused on the first trials of a visual search task, before participants experience finding the target. Across two experiments, we found that no prior experience with color or shape pop-out in previous trials was needed for participants to be able to terminate the search early when a target would have been found immediately. In other words, participants were sensitive to the counterfactual efficiency with which a hypothetical target would have been detected even in the first trials of the experiment. This result rules out a purely heuristic-based account of search termination and suggests that in these first few trials, participants are relying on prior second-order knowledge about visual attention (e.g., 'red pops out,' or 'a dot would catch my attention'), on a pre-attentional identification of target absence via ensemble statistics, or on a combination of the two.

Do participants employ a counterfactual heuristic, drawing on implicit metacognitive knowledge about search efficiency, or instead immediately perceive the absence of a target via ensemble scene statistics? We suggest that without second-order knowledge of their own perception and attention, ensemble perception alone is not sufficient to account for absence pop-out. Ensemble perception allows observers to extract summary statistical information from sets of similar stimuli, without directly perceiving any single stimulus (Whitney & Yamanashi Leib, 2018). According to this account, if participants immediately perceive that the search array comprises only squares, they might not need to rely on any counterfactual thinking or self-knowledge to conclude that no circle was present. Importantly, however, for the global statistical property 'the array comprises only squares' to be extracted from a display without representing individual squares, the visual system must represent, explicitly or implicitly, that a non-square item would have been detected by the visual system if they had been present. This second-order representation can be implemented, for example, as a threshold on curvature-sensitive neurons ('a round object would have induced a higher firing rate in this neuron population'), or more generally as a likelihood function going from polygons to firing patterns ('The perceived input is most likely under a world state where the display includes only polygons').

As an illustration, assume that Sarah, a participant in our experiment, does not know that a red item would immediately catch her attention in an array of blue distractors. Not only can Sarah not report this fact, this knowledge is not represented and cannot influence her cognitive system. Sarah is now searching for a red dot, and sees a uniform array of blue dots. How can she know that she hasn't missed a red dot? In the absence of second-order knowledge about search efficiency, Sarah would have to scan the dots one by one before committing to a 'target absent' response. Therefore, whether or not ensemble perception plays a role in absence pop-out, second-order knowledge about search efficiency is necessary to exlain the effects we observe.

Should this second-order knowledge be considered metacognitive? We argue that it should, and note that it is not a prerequisite for metacognitive knowledge to be accessible to consciousness. Metacognitive knowledge was originally assumed by Flavell (1979) to mostly affect cognition without accessing consciousness at all (i.e. without inducing a 'metacognitive experience'). Different aspects of metacognitive monitoring, including an immediate *Feeling of Knowing* when presented with a problem,

have been attributed to implicit metacognitive mechanisms that share a conceptual similarity with the ones described in the previous paragraph (Reder & Schunn, 1996). Indeed, metacognitive knowledge is sometimes measured as an ability to flexibly adapt information gathering thresholds: similar to a decision to terminate a search, the decision to stop gathering more information is widely accepted to be guided by metacognitive factors in developmental (Leckey et al., 2020; Siegel, Magid, Pelz, Tenenbaum, & Schulz, 2021) and comparative (Watanabe, Grodzinski, & Clayton, 2014) psychology.

Our findings complement and extend previous work in which participants had introspective awareness of attentional capture (Adams & Gaspelin, 2020, 2021): our results suggest that on top of the ability to monitor attention, people also hold valid second-order knowledge about attentional processes, that allows them to make predictions and guide their information gathering decisions. A schematic model of attention has been suggested to be implemented in the brains of many animal species, including all mammals and birds, and to facilitate attention control and monitoring (Graziano, 2013). This kind of implicit second-order knowledge, perhaps together with a capacity to extract ensemble statistics from a display, may be crucial for representing the *absence* of objects. The critical difference between inferring *X is absent* and simply lacking the belief *X is present* is a counterfactual belief that *X* would have been detected, had it been presented (Mazor, 2021; Mazor & Fleming, 2020). In turn, studying the processes underpinning efficient inference about absence can shed light on the role of higher-order representations in perception - because such counterfactual beliefs rest on representing, perhaps implicitly, how one's own perceptual system might respond under various conditions.

1.5 Conclusion

Our findings reveal that some knowledge about search efficiency is available to participants already in the first trials of the experiment, before engaging with the task or knowing what distractors to expect. These results reflect the same qualitative response time patterns as those commonly obtained in typical (few subjects/many trials) visual search experiments. Given that no target was present in these trials, participants must have been sensitive to the counterfactual likelihood of them finding the target, had it been present. In Experiment 2 we showed that this second-order knowledge about search difficulty was often accessible to report, but that this was not a necessary condition for efficient search termination. We conclude that efficient inference about absence is critically dependent on implicit second-order knowledge about visual search. In the next chapter I look more closely at participants' explicit second-order knowledge of their visual search behaviour, by asking for prospective search-time estimates for unfamiliar, complex stimuli.

Chapter 2

Internal models of visual search are rich, person-specific, and mostly accurate

Matan Mazor, Max Siegel & Joshua B. Tenenbaum

Having an internal model of one's attention can be useful for effectively managing limited perceptual and cognitive resources. For example, in the previous Chapter I argued that knowledge of the attentional capture of shape or colour singletons allows participants to immediately recognize the absence of objects in displays. While this and other work has hinted to the existence of an internal model of attention, it is still unknown how rich and flexible this model is, whether it corresponds to one's own attention or alternatively to a generic person-invariant schema, and whether it is specified as a list of facts and rules, or alternatively as a probabilistic simulation model. To this end, we designed a task to test participants' ability to estimate their own behavior in a visual search task with novel displays. In four online experiments (two exploratory and two pre-registered), prospective search time estimates reflected accurate metacognitive knowledge of key findings in the visual search literature, including the set-size effect, higher efficiency of feature- over conjunction- searches, and visual search asymmetry for familiar and unfamiliar stimuli. We further find that participants' estimates fit better with their own search times compared to the search times of other participants. Together, we interpret our findings as suggesting that people hold an internal model of visual search that is rich, person specific, and mostly accurate.

2.1 Introduction

In order to efficiently interact with the world, agents construct *mental models*: simplified representations of the environment and of other agents that are accurate enough to generate useful predictions and handle missing data (Forrester, 1971; Friston, 2010; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). For example, participants' ability to predict the temporal unfolding of physical scenes has been attributed to an 'intuitive

physics engine': a simplified model of the physical world that uses approximate, probabilistic simulations to make rapid inferences (Battaglia, Hamrick, & Tenenbaum, 2013). Similarly, having a simplified model of planning and decision-making allows humans to infer the beliefs and desires of other agents based on their observed behaviour (Baker, Saxe, & Tenenbaum, 2011). Finally, in motor control, an internal model of one's motor system and body allows subjects to monitor and control their body (Wolpert, Ghahramani, & Jordan, 1995). This internal forward model has also been proposed to play a role in differentiating self and other (Blakemore, Wolpert, & Frith, 1998). In recent years, careful experimental and computational work has advanced our understanding of these internal models: their scope, the abstractions that they make, and the consequences of these abstractions for faithfully and efficiently modeling the environment.

Agents may benefit from having a simplified model not only of the environment, other agents, and their motor system, but also of their own perceptual, cognitive and psychological states. Chapter 1 demonstrated the utility in a mental self-model for efficient inference about absence without task experience. But the utility in a mental self-model goes beyond visual search. For example, it has been suggested that knowing which items are more subjectively memorable is useful for making negative recognition judgments ["I would have remembered this object if I saw it"; Brown, Lewis, & Monk (1977)]. Similarly, children guided their decisions and evidence accumulation based on model-based expectations about the perception of hidden items (Siegel, Magid, Pelz, Tenenbaum, & Schulz, 2021). In the context of perception and attention, Graziano & Webb (2015) argued that having a simplified Attention Schema - a simplified model of attention and its dynamics - is crucial for monitoring and controlling one's attention, similar to how a body-schema supports motor control.

Indeed, people are not only capable of predicting the temporal unfolding of physical scenes, or the behaviour of other agents, but also the workings of their own attention under hypothetical scenarios. In one study, participants held accurate beliefs about the serial nature of visual search for a conjunction of features, and the parallel nature of visual search for a distinct color (Levin & Angelone, 2008). Similarly, the majority of third graders knew that the addition of distractors makes finding the target harder, particularly if the distractors and target are of the same color (Miller & Bigi, 1977). These and similar studies established the existence of metacognitive knowledge about visual search, as a result raising new questions about its structure, limits, and origins. We identify three such open questions. First, do internal models of visual search represent search difficulty along a continuum, or alternatively classify search displays as being either 'easy' or 'hard?' Second, to what extent is knowledge about visual search learned or calibrated based on first-person experience? And third, are internal models of visual search structured as a list of facts and laws, or as an approximate probabilistic simulation?

Here we take a first step toward providing answers to these three questions, using visual search as our model test case for internal models of perception and attention more generally. Participants estimated their prospective search times in visual search tasks and then performed the same searches. Similar to using colliding balls (Smith & Vul, 2013) and falling blocks (Battaglia, Hamrick, & Tenenbaum, 2013) to study intuitive

physics, here we chose visual search for being thoroughly studied and for following robust behavioural laws. In Experiments 1 and 2, we used simple colored shapes as our stimuli, and compared participants' internal models to scientific theories of attention that distinguish parallel from serial processing. We found that participants represented the relative efficiency of different search tasks along a continuum, but had a persistent bias to assume serial search. In experiments 3 and 4 we used unfamiliar stimuli from the Omniglot dataset (Lake, Salakhutdinov, Gross, & Tenenbaum, 2011) with the purpose of testing the richness and compositional nature of participants' internal models, and their reliance on person-specific knowledge. We find that participants do remarkably well in predicting their search times for novel stimuli. Furthermore, we show that internal models of visual search are person-specific, in that they are better fitted to one's own search behaviour compared with the search behaviour of other participants. Although estimation time analysis failed to provide direct evidence for online simulation, we suggest that a graded, person-specific representation of visual search is most consistent with a simulation account.

2.2 Experiments 1 and 2: shape, orientation, and color

An internal model of visual search may take a similar form to that of a scientific theory, by specifying an ontology of concepts and a set of causal laws that operate over them (Gerstenberg & Tenenbaum, 2017; Gopnik & Meltzoff, 1997). For example, participants may hold an internal model of visual search that is similar to Anne Treisman's *Feature Integration Theory*. According to this theory, visual search comprises two stages: a pre-attentive parallel stage, and a serial focused attention stage (A. Treisman, 1986; A. Treisman & Sato, 1990). In the first stage, visual features (such as color, orientation, and intensity) are extracted from the display to generate spatial 'feature maps.' Targets that are defined by a single feature with respect to their surroundings can be located based on these feature maps alone (*feature search*; for example searching for a red car in a road full of yellow taxis). Since the extraction of a feature map is pre-attentive, in these cases search can be completed immediately. In contrast, sometimes the target can only be identified by integrating over multiple features (*conjunction search*; for example if the road has not only yellow taxis, but also red buses). In such cases, attention must be serially deployed to items in the display until the target is identified.

A simplifying assumption of Feature Integration Theory is that there is no transfer of information between the pre-attentive and focused attention stages. In other words, observers cannot selectively direct their focused attention to items that produced strong activations in the pre-attentive stage. *Guided Search* models (J. M. Wolfe, 1994, 2021; J. M. Wolfe, Cave, & Franzel, 1989) assume instead that participants use these pre-attentive guiding signals in their serial search. Compared to Feature Integration Theory, Guided Search models provide much better fit to empirical data, at the expense of being more complex and rich in detail. To date, it is unknown where do internal models of visual search fall on this performance-complexity trade-off:

do people differentiate between 'easy' and 'hard' searches like in Feature Integration Theory, or do they represent search difficulty on a continuum, more like Guided Search?

In Experiments 1 and 2 we used stimuli that lend themselves to a categorical distinction between parallel and serial search: simple geometrical shapes of different colors and orientations. We asked whether participants' internal models of visual search predict which search displays demand serial deployment of attention and which don't. Critically, participants gave their search time estimates before they were asked to perform searches involving these or similar stimuli, so their search time estimates reflected prior beliefs about search efficiency. Experiment 2 was designed to replicate and generalize the results of Exp. 1 to a new stimulus dimension (orientation) and distractor set sizes. Our hypotheses and analysis plan for Experiment 2, based on the results of Experiment 1, were pre-registered prior to data collection (pre-registration document: osf.io/2dpq9). Raw data and full analysis scripts are available at github.com/matanmazor/metaVisualSearch.

2.2.1 Participants

For Exp. 1, 100 participants (no demographic data was collected) were recruited from Amazon's crowdsourcing web-service Mechanical Turk. Exp. 1 took about 20 minutes to complete. Each participant was paid $2.50. The highest performing 30% of participants received an additional bonus of $1.50. For Exp. 2, 100 participants (median reported age: 35.00; range: [19-80]) were recruited from the Prolific crowdsourcing web-service. The experiment took about 15 minutes to complete. Each participant was paid £1.5. The highest performing 30% of participants received an additional bonus of £1.

2.2.2 Procedure

The study was built using the Lab.js platform (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2019) and hosted on a JATOS server (Lange, Kühn, & Filevich, 2015). Static versions of all four experiments are available at github.com/matanmazor/metaVisualSearch.

Familiarization

First, participants were acquainted with the visual search task. The instructions for this part were as follows:

> In the first part, you will find a target hidden among distractors. First, a gray cross will appear on the screen. Look at the cross. Then, the target and distractors will appear. When you spot the target, press the spacebar as quickly as possible. Upon pressing the spacebar, the target and distractors will be replaced by up to 5 numbers. To move to the next trial, type in the number that replaced the target.

The instructions were followed by four trials of an example visual search task (searching for a *T* among 7 *L*s). Feedback was delivered on speed and accuracy. The purpose of this part of the experiment was to familiarize participants with the task.

Estimation

After familiarization, participants estimated how long it would take them to perform various visual search tasks involving novel stimuli and various set sizes. On each trial, they were presented with a target stimulus and a display of distractors and were asked to estimate how long it would take to find the target if it was hidden among the distractors (see Fig. 2.1).

To motivate accurate estimates, we explained that these visual search tasks will be performed in the last part of the experiment, and that bonus points will be awarded for trials in which participants detect the target as fast or faster than their search time estimate. The number of points awarded for a successful search changed as a function of the search time estimate according to the rule $points = \frac{1}{\sqrt{secs}}$. This rule was chosen for being exponential with respect to the log response times, incentivizing participants to be consistent in their ratings across short and long search tasks (see Appendix C.1). The report scale ranged from 0.1 to 4 seconds in Exp. 1 and to 2 seconds in Exp. 2. After one practice trial (estimating search time for finding one *T* among 3 randomly positioned *L*s), we turned to our stimuli of interest. In Experiment 1, participants estimated how long it would take them to find a red (#FF5733) square among green (#16A085) squares (color condition), red circles (shape condition) and a mix of green squares, red circles, and green circles (shape-color conjunction condition), for set sizes 1, 5, 15 and 30. Together, participants estimated the expected search time of 12 different search tasks (see Figure 2.1, upper right panel). In Experiment 2, participants rated how long it would take them to find a red tilted bar (20° off vertical) among green titled bars (color condition), red vertical bars (orientation condition) and a mix of green tilted and red vertical bars (orientation-color conjunction condition) for set sizes 2, 4, and 8. Together, participants estimated the expected search time of 9 different search tasks (see Figure 2.1, lower right panel). In both experiments, the order of estimation trials was randomized between participants.

Visual Search

Participants performed three consecutive search tasks for each of the 12 (Exp. 1) or 9 (Exp. 2) search types. The order of presentation was randomized between participants. No feedback was delivered about speed. To motivate accurate responses, error trials were followed by a 5-second pause.

2.2.3 Results

Accuracy in the visual search task was reasonably high in both Experiments (Exp. 1: $M = 0.93$, 95% CI [0.90, 0.96]; Exp. 2: $M = 0.82$, 95% CI [0.77, 0.87]). Error trials and visual search trials that took shorter than 200 milliseconds or longer than 5

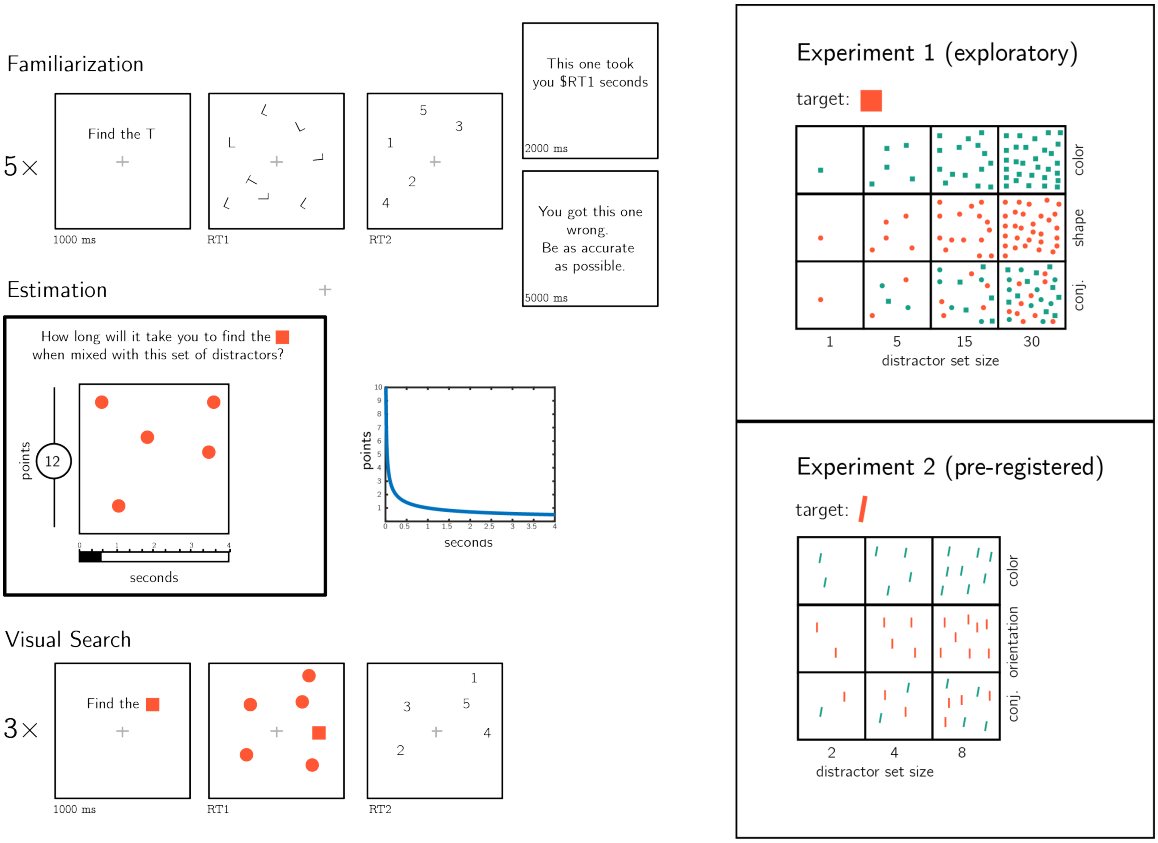


Figure 2.1: Experimental design. Participants first performed five similar visual search trials and received feedback about their speed and accuracy. Then, they were asked to estimate the duration of novel visual search tasks. Bonus points were awarded for accurate estimates, and more points were awarded for risky estimates. Finally, in the visual search part participants performed three consecutive trials of each visual search task for which they gave a search time estimates. Right panels: stimuli used for Experiments 1 and 2.

seconds were excluded from all further analysis. Participants were excluded if more than 30% of their trials were excluded based on the aforementioned criteria, leaving 89 and 74 participants for the main analysis of Experiments 1 and 2, respectively.

Search times

For each participant and distractor type, we extracted the slope of the function relating RT to distractor set size. As expected, search slopes for color search were not significantly different from zero in Exp. 1 (-0.40 ms/item; $t(88) = -0.45$, $p = .652$, $\text{BF}_{01} = 7.74$) and Exp. 2 (0.51 ms/item; $t(73) = 0.07$, $p = .946$, $\text{BF}_{01} = 7.80$). This is consistent with color being a basic feature that is not dependent on serial attention for its extraction by the visual system (A. Treisman, 1986; A. Treisman & Sato, 1990). The slope for shape search was close, but significantly higher than zero (5.66 ms/item;

$t(88) = 4.35$, $p < .001$), and the slope for orientation was numerically higher than zero (11.05 ms/item) but not significantly so ($t(73) = 1.50$, $p = .139$, $BF_{01} = 2.70$). In both Experiments, conjunction search gave rise to search slopes significantly higher than zero (Exp. 1: 14.80 ms/item ($t(88) = 9.16$, $p < .001$; Exp. 2: 72.14 ms/item ($t(73) = 7.50$, $p < .001$; see Figure 2.2).

Estimation accuracy

We next turned to analyze participants' prospective search time estimates, and their alignment with actual search times. In both Experiments, participants generally overestimated their search times. This was the case for all search types across the two Experiments (see Figure 2.2, left panels: all markers are above the dashed $x = y$ diagonal). This is expected, based on our bonus scheme that incentivized conservative estimates (see Appendix C.1). Despite this bias, estimates were correlated with true search times, supporting a metacognitive insight into visual search behaviour (see Fig. 2.2, left panels. Within subject Spearman correlations, Exp. 1: $M = 0.28$, 95% CI [0.21, 0.35], $t(88) = 7.77$, $p < .001$; Exp 2: $M = 0.16$, 95% CI [0.07, 0.26], $t(73) = 3.48$, $p = .001$).

To test participants' internal models of visual search, we analyzed their estimates as if they were search times, and extracted *estimation slopes* relating estimates to the number of distractors in the display (see Fig. 2.2, right panels). Estimation slopes (expected ms/item) were steeper than search slopes for all search types. In particular, although search time for a deviant color was unaffected by the number of distractors, participants estimated that color searches with more distractors should take longer (mean estimated slope in Exp. 1: 17.76 ms/item; $t(88) = 6.35$, $p < .001$; in Exp 2: 29.43 ms/item; $t(73) = 2.63$, $p = .010$). In other words, at the group level, participants showed no metacognitive insight into the parallel nature of color search.

Although they were significantly different from zero, in both Experiments estimation slopes for color search were significantly shallower than for conjunction search (Exp. 1: $t(88) = 4.08$, $p < .001$, Exp. 2: $t(73) = 3.87$, $p < .001$). In contrast, although true search slopes were shallower for shape and orientation than for conjunction (p's<0.001), the difference in estimation slopes was not significant (difference between shape and conjunction slopes: $t(88) = 1.65$, $p = .103$; difference between orientation and conjunction slopes: $t(73) = 1.18$, $p = .244$).

A graded representation of search efficiency

In Feature Integration Theory, searches come in two flavours: parallel and serial. If participants' model of visual search shares this simplifying assumption, the results from the previous section indicate that their models also wrongly specify that shape and orientation searches are serial just like conjunction search. In contrast, an internal model of visual search may represent search efficiency along a continuum, with some searches being highly efficient, some highly inefficient, and others fall somewhere in between the two ends. This is more in line with Guided Search models (Hoffman, 1979; J. M. Wolfe, 2021; J. M. Wolfe, Cave, & Franzel, 1989). To decide between these two

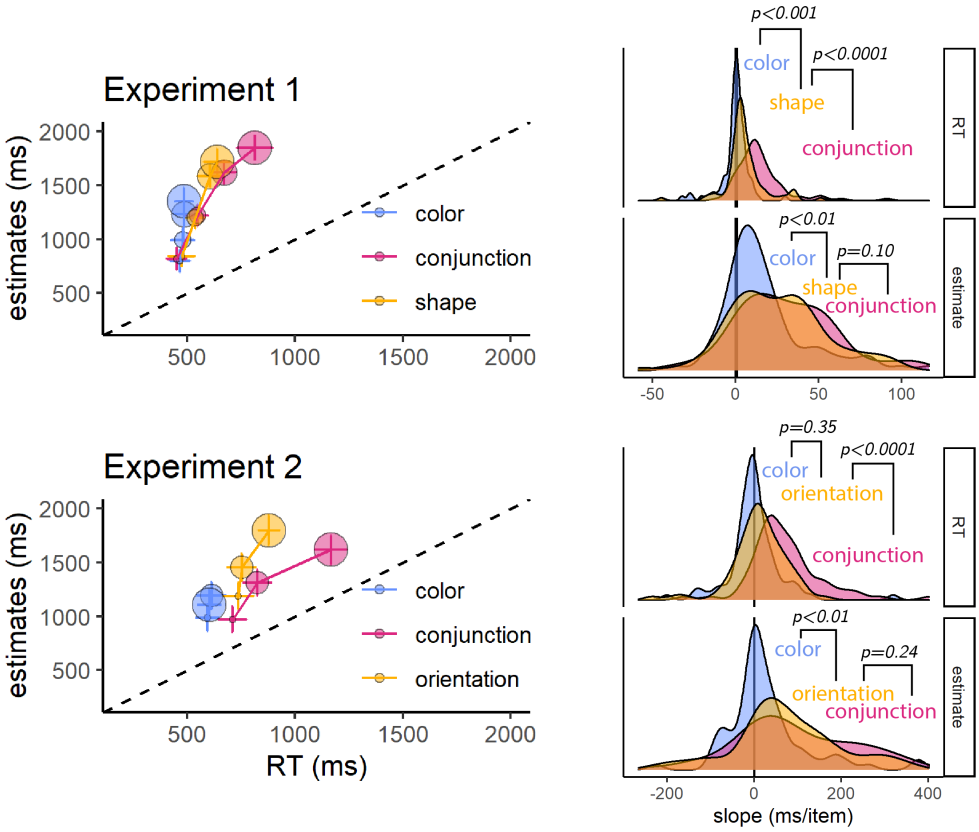


Figure 2.2: Left panels: median estimated search times plotted against true search times for the different search types (coded by color), and set sizes (coded by circle size; from small to large), for Exp. 1 (upper panel) and 2 (lower panel). Error bars represent the standard error of the median. Right panels: distribution of search (top) and estimated (bottom) slopes for the three search types in Exp. 1 (upper panel) and 2 (lower panel). The dashed line indicates $y = x$.

options, we focused on the slopes for shape and orientation. These searches were more efficient than conjunction search, but not as efficient as colour search. We tested if this efficiency gradient was represented in search time estimates of single individuals, or alternatively, emerged at the group level only. To this end, we scaled subject-specific RT and estimate slopes with respect to conjunction slopes $\beta_{scaled} = \frac{\beta}{\beta_{conjunction}}$. If representations of search efficiency are dichotomous, single participants can represent shape search either as being equally difficult as conjunction search, or as equally easy as color search. This predicts that the distribution of scaled estimate slopes should peak either at 1 or at the same value as color search. Instead, scaled estimate slopes for both shape and orientation peaked at values lower than 1 and higher than color search, indicating a graded representation of search efficiency in the internal models of single subjects (See Fig. 2.3. Exp. 1: median: 0.85; mode: 0.92; One sided Wilcoxon test against 1: $V = 914.00$, $p = .040$; One sided Wilcoxon test against color slope:

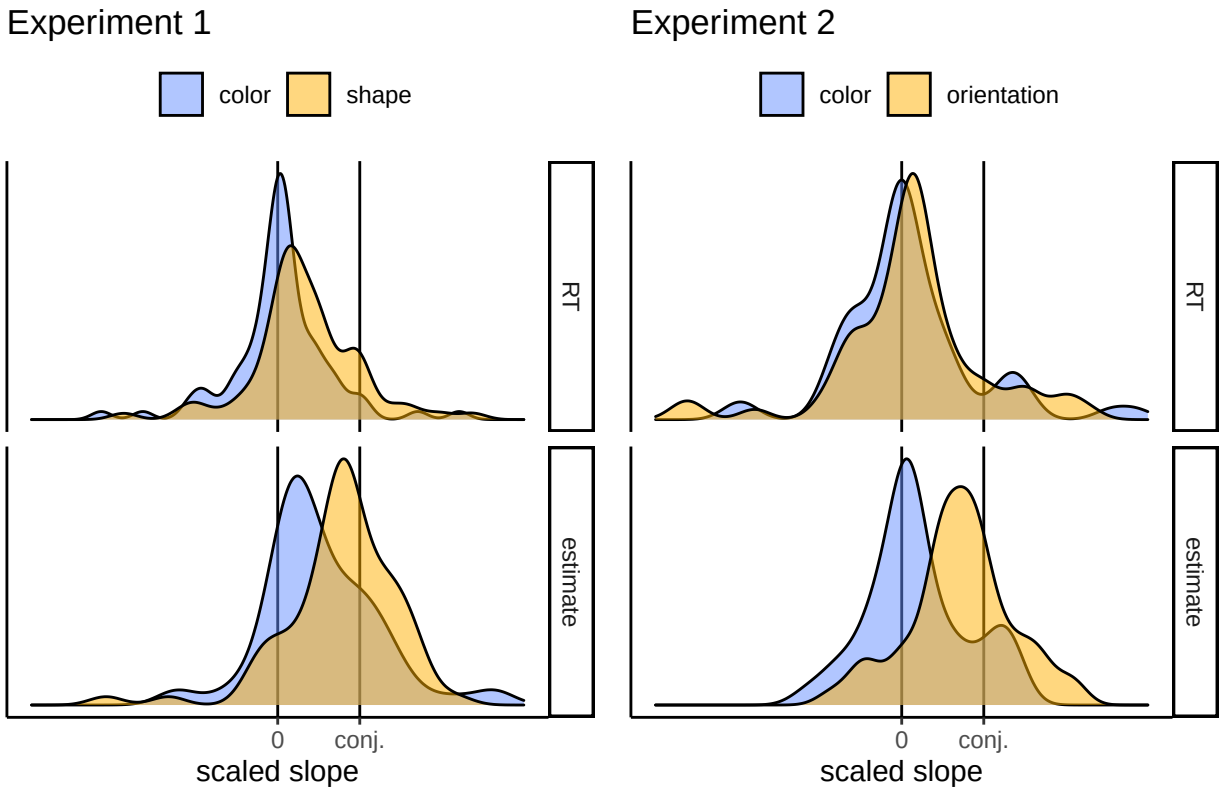


Figure 2.3: Normalized slopes for feature searches in Experiments 1 (left) and 2 (right). Search and estimate slopes were normalized with respect to conjunction slopes, to yield subject specific estimates.

$V = 1,488.00$, $p = .047$. Exp. 2: median: 0.75; mode: 0.70; One sided Wilcoxon test against 1: $V = 405.00$, $p = .013$; One sided Wilcoxon test against color slope: $V = 969.00$, $p = .001$).

2.3 Experiments 3 and 4: complex, unfamiliar stimuli

In Experiments 1 and 2 an internal model of visual search allowed participants to accurately estimate how long it would take them to find a target stimulus in arrays of distractor stimuli. Participants had insight into the set-size effect and into the fact that conjunction searches are more difficult than feature searches. We also found that participants' internal models of visual search represented search efficiency along a gradient, and were systematically biased to overestimate the effect of set-size, even in feature searches in which the number of distractors had no effect on search time.

In Experiments 3 and 4 we asked how rich this model is, by using displays of complex stimuli with which participants are unlikely to have any prior experience

(letters from a medieval Alphabet and from the Futurama TV series). Here, insight into the set size effect and its absence in feature searches would not be useful for generating accurate search time estimates. Instead, participants' internal model of visual search must be capable of extracting relevant features from rich stimuli, and use these features to generate stimulus-specific predictions based on some intricate model of how visual search works. Using these more complex stimuli further allowed us to ask if search-time estimates rely on person-specific knowledge. Exp. 4 followed Exp. 3 and was pre-registered (pre-registration document: osf.io/dprtk). Raw data and full analysis scripts are available at github.com/matanmazor/metaVisualSearch.

2.3.1 Participants

For Exp. 3, 100 participants (median reported age: 31; range: [19-65]) were recruited from the Prolific crowdsourcing web-service. The experiment took about 15 minutes to complete. Participants were paid £1.5. The highest performing 30% of participants received an additional bonus of £1. For Exp. 4, 200 participants (median reported age: 33; range: [19-81]) were recruited from the Prolific crowdsourcing web-service. We recruited more participants for Exp. 4 in order to have sufficient statistical power for our inter-subject correlation analysis. The experiment took about 8 minutes to complete. Participants were paid $1.27. The highest performing 30% of participants received an additional bonus of $0.75.

2.3.2 Procedure

The procedure for Experiments 3 and 4 was similar to that of Exp. 1 with several changes.

Stimuli were letters drawn by Mechanical Turk workers (Lake, Salakhutdinov, Gross, & Tenenbaum, 2011), instead of geometrical shapes (see Fig. 2.4). In Exp. 3, we used letters from the *Alphabet of the Magi*. In Exp. 4, we used letters from the *Futurama* television series as well as Latin letters. We explained to participants that they will search for a specific letter (the target letter) among copies of another letter (the distractor letter). In Exp. 3, both target and distractor were letters from the Alphabet of the Magi, and distractors were drawn by different Mechanical Turk workers. In Exp. 4, on half of the trials the target was a Latin letter and distractors were Futurama letters and on the other half the target was a Futurama letter and distractors were Latin letters. In these experiments, distractors were copies of the same letter drawn by the same Mechanical Turk worker. This was important for our visual search asymmetry analysis (see below). In the familiarization part, we used as target and distractors two letters from the Alphabet of the Magi (Exp. 3) and two letters from the Futurama alphabet (Exp. 4). Importantly, these letters were only used for training, and did not appear in the Estimation or Visual search parts. In the Estimation part participants gave search time estimates for 8 search tasks, all involving 10 distractors, and in the Visual Search part they performed these search tasks. To minimize random variation in spatial configurations (which was important for the search asymmetry analysis), in Exp. 4 letters appeared on an invisible clock

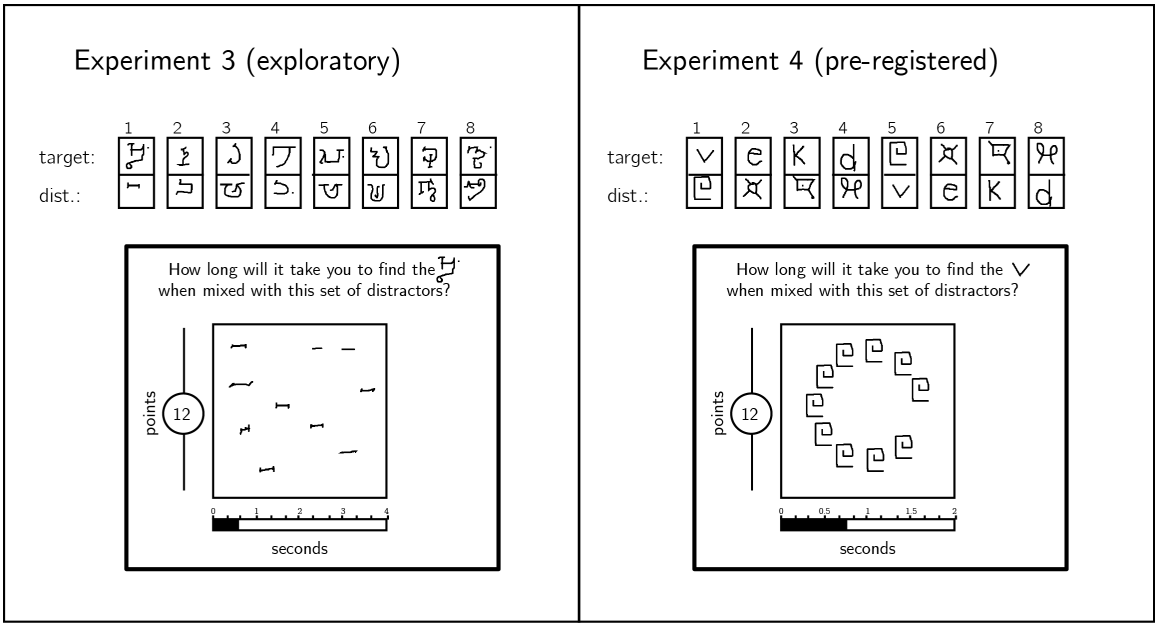


Figure 2.4: In Exp. 3, stimuli were characters from the Alphabet of the Magi, and distractors were drawn by different Mechanical Turk users. In Exp. 4, stimuli were characters from the Latin and Futurama alphabets. Stimulus pairs 1-4 and 5-8 are identical except for the target assignment. In Exp. 4, all distractors in a display were drawn by the same Mechanical Turk user, and were presented on an invisible clock face.

face. Finally, the report scale ranged from 0.1 to 4 seconds in Exp. 3 and to 2 seconds in Exp. 4.

2.3.3 Results

Accuracy in the visual search task was high in both experiments (Exp. 3: $M = 0.89$, 95% CI [0.86, 0.92]; Exp. 4: $M = 0.97$, 95% CI [0.96, 0.98]). Error trials and visual search trials that took shorter than 200 milliseconds or longer than 5 seconds were excluded from all further analysis. Participants were excluded if more than 30% of their trials were excluded based on the aforementioned criteria, leaving 88 and 200 participants for the main analysis of Experiments 3 and 4, respectively.

Estimation accuracy

In both experiments, search time estimates were positively correlated with true search times (within-subject Spearman correlations in Exp. 3: $M = 0.44$, 95% CI [0.37, 0.52], $t(86) = 12.16$, $p < .001$; Exp. 4: $M = 0.10$, 95% CI [0.05, 0.15], $t(191) = 3.67$, $p < .001$; see Figures 2.5 and 2.7A). The correlation between search time and search time estimates was significantly weaker in Experiment 4 ($\Delta M = 0.35$, 95% CI [0.26, 0.43], $t(181.02) = 7.60$, $p < .001$)). This difference in correlation strength is likely the

result of a narrower range of search times in Exp. 4 (with median search times 566 - 684 ms, per display) than in Exp. 3 (649 - 1615 ms).

Importantly, in both experiments all searches involved exactly 10 distractors, so a positive correlation could not be driven by the effect of distractor set size. Furthermore, since participants had no prior experience with our stimuli, their estimates could not have been informed by explicit knowledge about specific letters ('The third letter in the *Alphabet of the Magi* pops out to attention when presented between instances of the fourth letter,' or 'the fifth letter in the *Futurama Alphabet* is difficult to find when presented among *d*s'). These positive correlations reveal a more intricate knowledge of visual search. Our next two analyses were designed to test whether estimates were based on person-specific knowledge, and whether their generation involved a simulation of the search process.

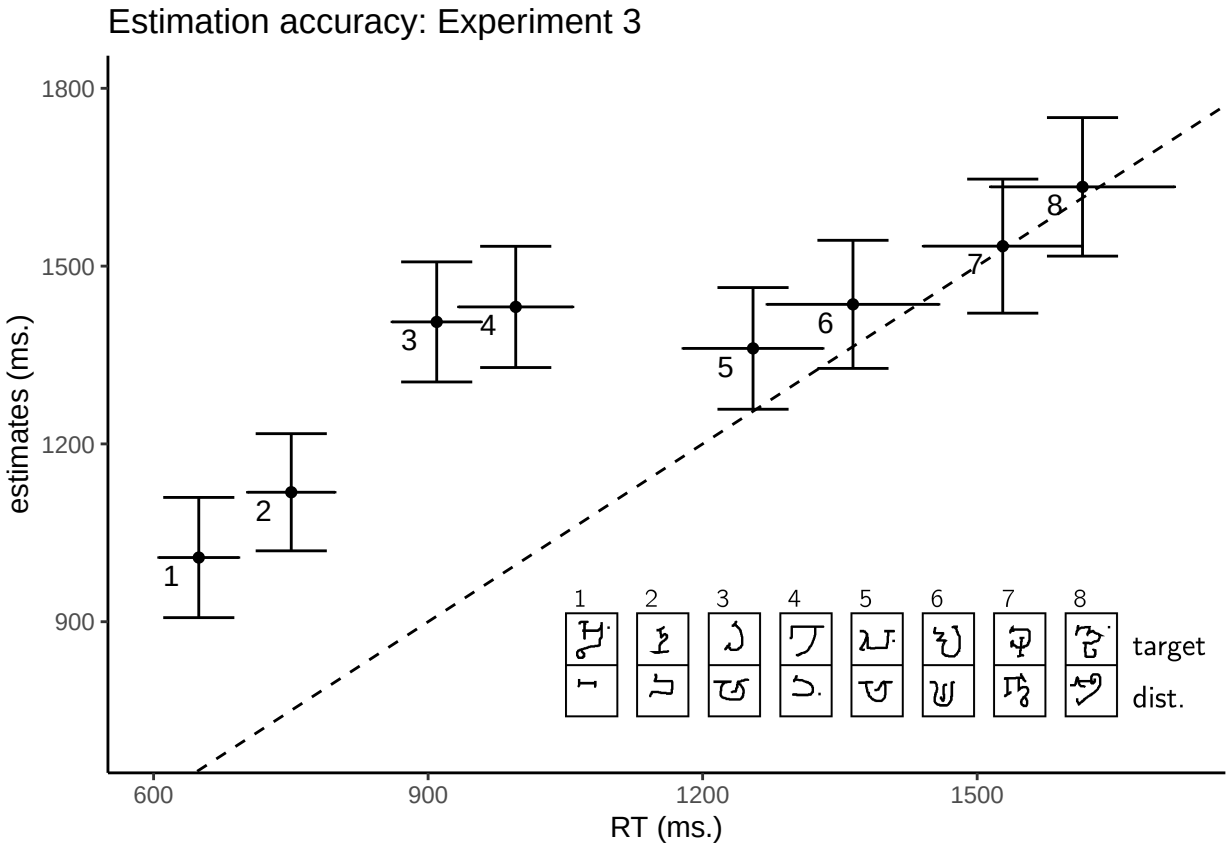


Figure 2.5: Estimated search times plotted against true search times in Experiment 2. The dashed line indicates $y = x$. Legend: each search task involved searching for one Omniglot character (top letter) among ten tokens of a second Omniglot character, drawn by 10 different MTurk workers (bottom letter).

Cross-participant correlations

We chose unfamiliar letters as stimuli for Experiments 3 and 4 in order to make heuristic-based estimation more difficult, and to encourage an introspective estimation process. If participants were using idiosyncratic knowledge about their own attention, we would expect to find higher correlations between their search time estimates and their own search times (*self-self alignment*), compared to with the search times of a random surrogate participant (*self-other alignment*). To test this, we ran a non-parametric permutation test, comparing self-self and self-other alignment in prospective search time estimates. In Exp. 3, a numerical difference between self-self (mean Spearman correlation $M_r = 0.44$) and self-other alignment ($M_r = 0.41$) was marginally significant ($p_{perm} = 0.05$). In Experiment 4, we pre-registered this analysis and found a significant advantage for self-self alignment compared with self-other alignment (see Fig. 2.6; mean Spearman correlations for self-self $M_r = 0.10$ and self-other $M_r = 0.04$, $p_{perm} = 0.01$). We interpret this result as indicating that at least some of participants' internal model of visual search builds on idiosyncratic knowledge about their own attention.

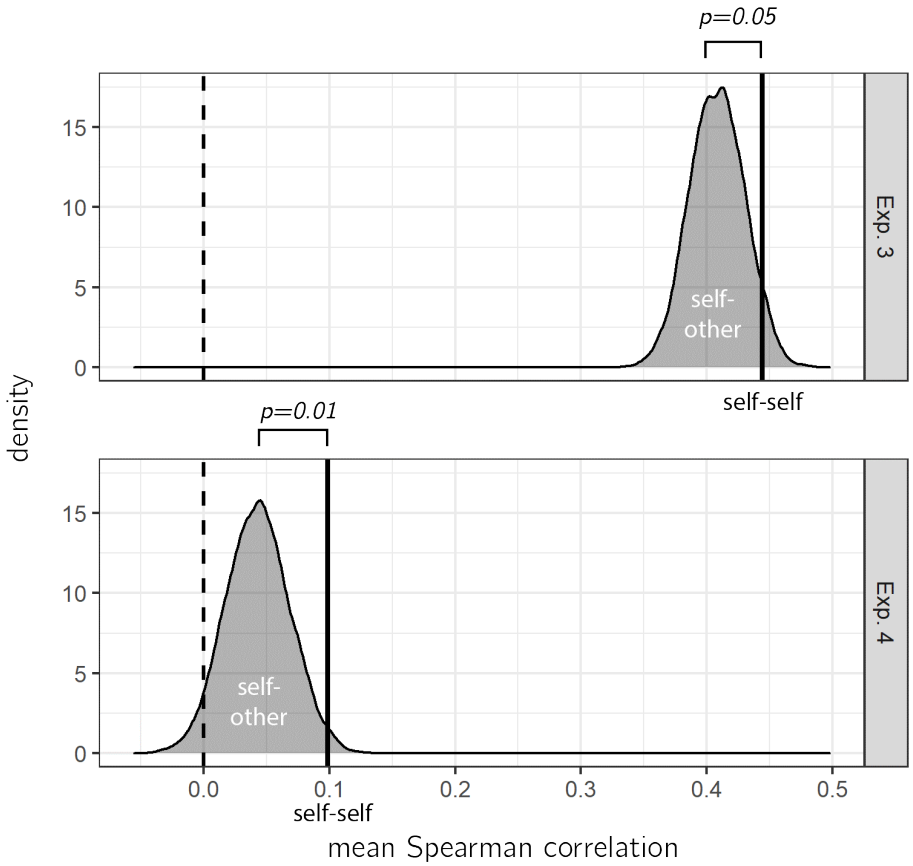


Figure 2.6: True correlation between estimates and search times (self-self alignment, vertical lines) plotted against a null distribution of correlations, when matching the estimates of each participant with the search time of a random surrogate participant (self-other alignment).

Estimation time

We next looked at the time taken to produce search time estimates in the Estimation part. We reasoned that if participants had to mentally simulate searching for the target in order to generate their search time estimates, they would take longer to estimate that a search task will terminate after 1500 compared to 1000 milliseconds. This is similar to how a linear alignment between the degree of rotation and response time in a mental rotation task was taken as support for an internal simulation that evolves over time (Shepard & Metzler, 1971). We find no evidence for within-subject correlation between estimates and the time taken to deliver them, not in Exp. 3 ($t(86) = 0.40$, $p = .692$) and not in Exp. 4 ($t(191) = 0.74$, $p = .458$). However, given that estimation times were three times longer than search time estimates (median time to estimate = 5 seconds in Exp. 3 and 3 seconds in Exp. 4), a simulation-driven correlation may have been masked by other factors that contributed to estimation times, such as motor control over the report slider.

Visual search asymmetry

In Exp. 4, we put to the test an alternative interpretation for the remarkable alignment between search time and search time estimates that we observed in Exp. 3. We considered the possibility that participants were relying on a heuristic: since search time generally inversely scales with the perceived similarity between the target and distractor stimuli, participants could achieve high accuracy in their estimates by basing them not on an intuitive theory of visual search, but on their impressions of similarity between the stimulus pairs. If all participants know about their visual search behaviour is that searches are harder the more similar the target and distractor are, simply being able to rate the similarity between pairs of stimuli would produce a good alignment between search times and their estimates.

To test if this was the case, we leveraged a well-established phenomenon in visual search: subjects are generally faster detecting an unfamiliar stimulus in an array of familiar distractors compared to when the target is familiar and the distractors are not (Malinowski & Hübner, 2001; Shen & Reingold, 2001; Zhang & Onyper, 2020). This asymmetry cannot be captured by a similarity-based heuristic (at least if similarity is represented as a symmetric property, cf. Tversky, 1977). In Exp. 4, participants were presented with pairs of familiar and unfamiliar letters, and estimated their search time for finding the familiar letter among unfamiliar distractors and vice versa. This allowed us to ask if their internal models of visual search were solely based on visual similarity between the target and distractor stimuli.

In addition to extracting correlation between search times and search time estimates, we extracted the same correlations after inverting the identity of the target and distractor stimuli in the estimates, but not in the actual search times. For example, instead of comparing search times for finding the letter v among 10 square spiral letters (stimulus pair 1) with estimates for the same search, we compared it with estimates for finding one square spiral letter among 10 v's (stimulus pair 5). If estimates were affected by the assignment of stimuli to target and distractor, this inversion should

attenuate the correlation, but if visual search estimates reflected a symmetric notion of similarity the correlation should not be affected.

Inverting the target/distractor assignment dropped the correlation between estimates and search time to zero ($M = -0.01$, 95% CI $[-0.06, 0.04]$), significantly lower than the original correlation ($M_d = 0.10$, 95% CI $[0.03, 0.18]$, $t(191) = 2.63$, $p = .009$; see Fig. 2.7B). This is in contrast to what is expected if search time estimates reflected symmetric similarity judgments, and in line with an interpretation of our findings as evidence for a rich internal model of visual search.

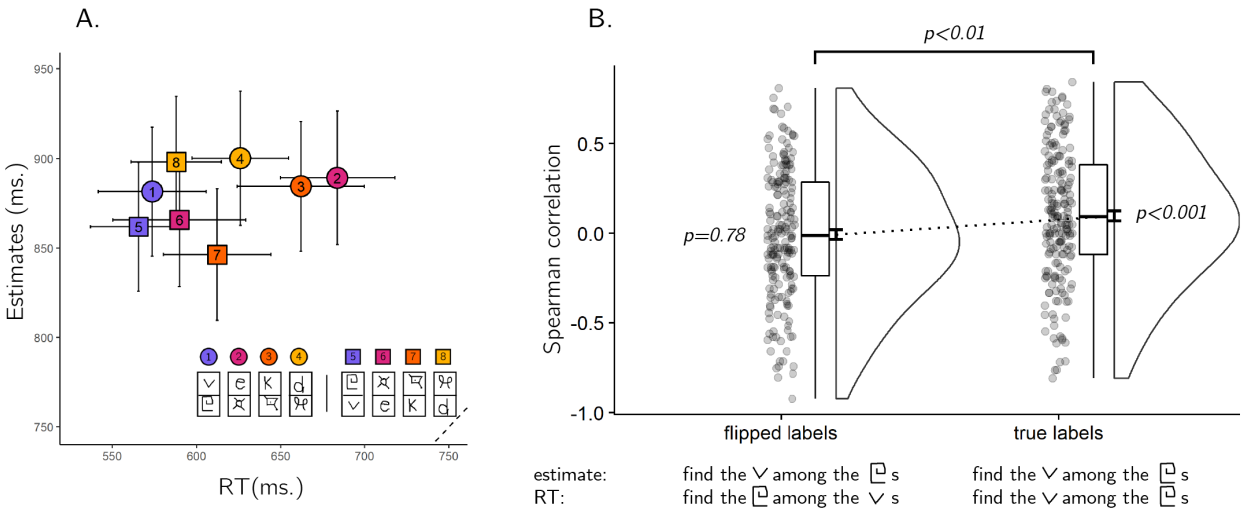


Figure 2.7: A. Median estimated search times plotted against true search times in Exp. 4. The dashed line indicates y=x. Legend: each search task involved searching for one character (top letter) among ten tokens of a different character (bottom letter). In four searches, the target character was from the Latin alphabet (circles), and in the other four from the Futurama alphabet (squares). Search pairs that involved the same pair of stimuli with opposite roles are marked by the same color. B. Spearman correlations between estimates and search times for true labels (upper panel) and target-distractor flipped labels (lower panel) in Exp. 4. Spearman correlations significantly dropped, indicating that participants were aware of the search asymmetry for stimulus familiarity.

2.4 Discussion

Over more than four decades of research on spatial attention, experiments where participants report the presence or absence of a target in a display revealed basic principles such as the set-size effect (A. Treisman, 1986; A. Treisman & Sato, 1990; J. M. Wolfe, 1998), the advantage for feature search over more complicated conjunction and spatial configuration searches (A. Treisman, 1986; A. Treisman & Sato, 1990), and asymmetries in the representations of visual features (Malinowski & Hübner, 2001; Shen & Reingold, 2001; A. Treisman & Souther, 1985). Some of these findings are intuitive, but others are more surprising, suggesting that even without training in psychology, people have a set of expectations and beliefs about their own perception and attention, and about visual search more specifically.

Here we measured these expectations and their alignment with actual visual search behavior. In four experiments, we show that naive participants provide reasonably accurate prospective estimates for their search times. In line with previous reports, prospective search time estimates reflected accurate knowledge of the set size effect and differences in efficiency between feature and conjunction searches (Levin & Angelone, 2008; Miller & Bigi, 1977). We asked whether participants categorically distinguish 'easy' from 'hard' searches, or alternatively represent search efficiency along a continuum. The estimates of single participants revealed a graded representation of search efficiency, indicating metacognitive knowledge that is on par with contemporary theories of visual search such as Guided Search models. Furthermore, participants provided accurate search time estimates for complex stimuli and displays with which they had no prior experience, and had metacognitive insight into the search asymmetry for familiar and unfamiliar stimuli.

In Exp. 4, we show that this internal model of visual search is person-specific: participants' predictions fitted better their own search times compared to the search times of other participants. The fact that this model is not generic suggests that it is learned or calibrated based on first-person experience. Humans accumulate observations not only of external events and objects, but also of their own cognitive and perceptual states. Specifically, subjects have been shown to notice when their attention is captured by a distractor (Adams & Gaspelin, 2021) even in the absence of an overt eye movement (Adams & Gaspelin, 2020). These observations can then be integrated into an internal model or an intuitive theory: which items are more or less likely to capture attention, under what circumstances, etc. Future research into the development of this simplified model and its expansion based on new evidence [for example, by measuring intuitions before and after exposure to some evidence; Bonawitz, Ullman, Bridgers, Gopnik, & Tenenbaum (2019)] is needed to understand the relation between metacognitive monitoring of attention and metacognitive knowledge of attentional processes.

This relates to recent theoretical and empirical advances underscoring the utility of keeping a *mental self-model*, or a *self-schema* for attention control (Wilterson et al., 2020), social cognition (Graziano, 2013) , phenomenal experience (Metzinger, 2003), and inference about absence (Mazor, 2021; Mazor & Fleming, 2021). For example, knowing that a red berry would be easy to find among green leaves, a forager can

quickly decide that a certain bush bears no ripe fruit. Alternatively, knowing that a snake would be difficult to spot in the sand, they might allocate more attentional resources to scanning the ground. Experiments 3 and 4 show that this knowledge is more than a set of heuristics or rules, but reflects an intricate internal model of spatial attention that can be applied to unseen stimuli in novel displays, and is tailored to one's own perceptual and cognitive machinery.

Our final question concerned the structure of this internal model: is it specified as a list of facts and laws [similar to how the acquisition of knowledge about mental states between the ages of 2 and 4 was described as the development of a scientific theory; Gopnik & Meltzoff (1997)], or alternatively as an approximate probabilistic model that can be used to run simulations [similar to the physics engine model of intuitive physics; Battaglia, Hamrick, & Tenenbaum (2013)]? We found no direct evidence for a simulation account in the time taken to produce search time estimates. Nevertheless, participants' ability to provide accurate estimates for displays of unfamiliar stimuli, and the better alignment of their estimates with their own search behavior compared to the search behavior of other participants, provide some indirect support for a simulation account - one that is based on a schematic version of one's own attention. Still, we cannot exclude rule-based implementations of this internal model that are rich in detail and are based on one's first-person experience, without involving a simulation.

One important limitation of our current design is its reliance on explicit estimates, which may have potentially resulted in underestimating the richness and accuracy of these internal models. For example, in Experiments 1 and 2 participants' prospective estimates showed no metacognitive insight into the pop-out effect for color search. This does not necessarily mean that this information was misrepresented in their internal model. Instead, our numeric report scheme may have encouraged participants to adopt an analytical disposition to the problem, rather than relying fully on their intuitions. In support of this, in Chapter 1 a pop-out effect for color absence in the first few trials of a visual search task in interpreted as indicating accurate implicit metacognitive knowledge of the pop-out effect for color presence.

Together, our results reveal an alignment between prospective search time estimates and search times. This alignment places a lower bound on the richness and complexity of participants' internal model of visual search, and of attention more generally, and opens a promising avenue for studying humans' intuitive understanding of their own mental processes.

Chapter 3

Evidence weightings in confidence judgments for detection and discrimination

Matan Mazor, Lucie Charles, Roni Or Maimon Mor & Stephen M. Fleming

In Chapters 1 and 2 I examined inference about absence and its relation to self-modelling in visual search, where a target is present or absent in an array of distractors. In this Chapter, I examine inference about absence in a near-threshold detection setting, where the location of the target is known and no distractors are present. Previous studies of near-threshold discrimination revealed a *positive evidence bias* (PEB) in discrimination confidence: confidence in perceptual decisions is more sensitive to evidence in support of the decision than to conflicting evidence. Recent theoretical proposals suggest that a PEB is due to observers adopting a detection-like strategy when rating their confidence, one that has functional benefits for metacognition in real-world settings where detectability and discriminability often go hand in hand. In three experiments (one lab-based and two online) we first successfully replicate a PEB in discrimination confidence. We then show that a PEB is observed in detection decisions, where participants report the presence or absence of a stimulus, regardless of its identity. We discuss our findings in relation to models that account for a positive evidence bias as emerging from a confidence-specific heuristic, and alternative models where decision and confidence are generated by the same, Bayes-rational process.

3.1 Introduction

When considering two alternative hypotheses, the probability of a chosen hypothesis to be correct is not only a function of the likelihood of observations under the chosen hypothesis, but also under the unchosen one. For example, when deciding that a random dot display was drifting to the right and not to the left, confidence should not only positively weigh motion energy to the right (*positive evidence*), but also negatively weigh motion energy to the left (*negative evidence*). However, when rating their subjective confidence, subjects place disproportional weight on evidence in favour

of the choice, giving rise to a *positive evidence bias* (Koizumi, Maniscalco, & Lau, 2015; Peters et al., 2017; Rollwage et al., 2020; Samaha & Denison, 2020; Sepulveda et al., 2020; Zylberberg, Barttfeld, & Sigman, 2012). Put differently, confidence ratings in discrimination are sensitive not only to the *relative evidence* of the chosen hypothesis compared with the unchosen one (also termed *balance of evidence*), but also to the *sum evidence* for the two hypotheses [which for perceptual decisions is often related to *visibility*; see Fig. 3.1, left panel; Rausch, Hellmann, & Zehetleitner (2018)].

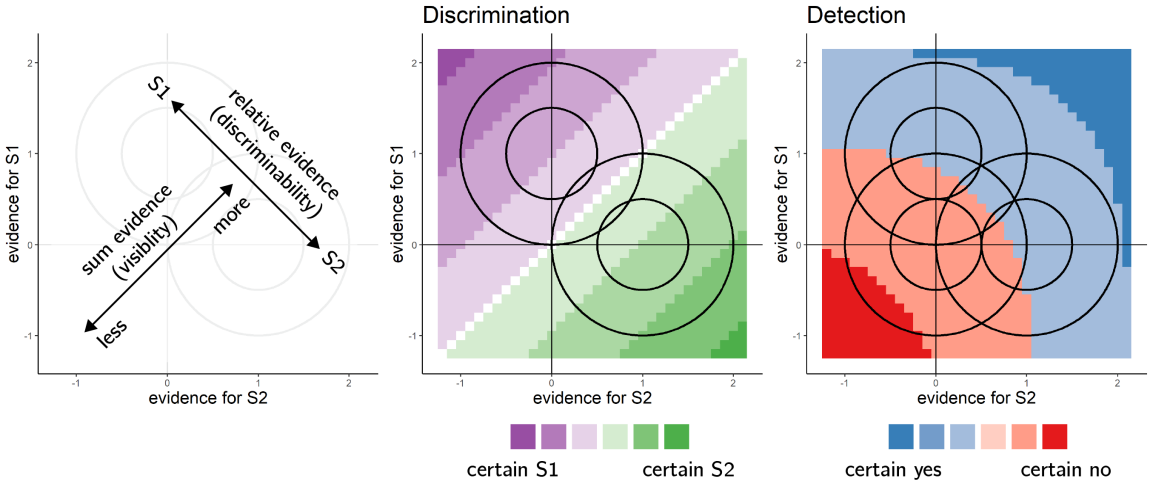


Figure 3.1: Discrimination and detection in a two-dimensional Signal Detection Theory model. Left: in a two-dimensional SDT model, percepts e are sampled from one of two Gaussian distributions (here centered at (0,1) and (1,0)). We define relative evidence as $e_{S1} - e_{S2}$ and sum evidence as $e_{S1} + e_{S2}$. Circles represent cross-sections of two-dimensional distributions. Center and Left: response and confidence accuracy are maximized when based on a log-likelihood ratio for the two stimulus categories. Center: in discrimination, this yields optimal decision and confidence criteria that are based on relative evidence (distance from the main diagonal), irrespective of sum evidence. Right: in detection, percepts in the absence of a stimulus are sampled from a Gaussian distribution centered at (0,0). This yields optimal decision and confidence that are based on a non-linear interaction between relative and sum evidence.

To account for this apparently irrational discounting of incongruent evidence in confidence formation, Maniscalco, Peters, & Lau (2016) point out that outside of a lab setting, representational spaces are so high-dimensional that keeping track of evidence for every possible stimulus category is not feasible. For example, to be confident that an object is an apple, one would have to incorporate evidence for this object not being an orange, a banana, a book and a ferret, among an infinite many other unsupported hypotheses. To resolve this engineering challenge, metacognitive

systems may have evolved to weigh evidence for the chosen hypothesis only, while ignoring conflicting evidence. This is similar to rating confidence not in the identity of a stimulus relative to other hypothetical stimuli, but in the presence of a stimulus relative to absence. Such a strategy is reasonable, as in Signal Detection space, samples that are farther away from the origin (high visibility) are on average farther away from the discrimination criterion (high discriminability). This strategy is then carried over to the lab, where decisions are made in low-dimensional representational spaces, and where keeping track of evidence for the two alternative stimulus categories is in fact feasible.

A more recent model identified the origin of this response-congruent heuristic not in the curse of dimensionality, but in the variance structure of perceptual evidence (Miyoshi & Lau, 2020). In a series of simulations, the authors augmented a bidimensional Signal Detection model with realistic assumptions about the sensory encoding of signal and noise, most importantly that the variance of signal tends to be higher than that of noise. In these settings, a Response Congruent Evidence (RCE) heuristic provided more accurate confidence judgments, meaning ones that are more aligned with objective accuracy, than did a Balance of Evidence (BE) heuristic. Again, this model implies that adopting a detection-like strategy when rating one's confidence might have functional benefits for metacognition.

Notably, both models imply a link between confidence in discrimination, and detection judgments about the presence or absence of a stimulus. In a detection setting with multiple possible targets, the likelihood ratio between stimulus presence and absence is more sensitive to positive evidence for the detected stimulus compared to evidence for the absence of other, undetected stimuli (see Fig. 3.1, right panel). Perhaps surprisingly, however, despite several recent studies finding that discrimination confidence is detection-like, there has been limited focus on the complementary question: do detection decisions share features of discrimination confidence, such as a positive evidence bias? In other words, when faced with a detection task where targets are drawn from two stimulus classes, would detection decisions be sensitive to sum evidence (like discrimination confidence), or to the relative evidence for presence for one category over the other? Moreover, little is known about confidence in these detection responses: would confidence in the presence of a target stimulus be susceptible to the same positive evidence bias as confidence in stimulus type? Finally, we asked whether detection confidence ratings would be sensitive to some form of positive evidence bias not only in decisions about target presence, but also in decisions about target absence.

To examine these questions, we conducted three experiments: one lab-based (N=10, 1800 trials per participant) and two online (N=102/100, 112/168 trials per participant). Participants performed discrimination and detection decisions on noisy stimuli, and rated their confidence in their decisions. Using reverse correlation analysis, we measured the influence of random fluctuations in stimulus energy on both responses and confidence ratings, and tested for the existence of processing asymmetries between detection 'yes' and 'no' responses [in response time, general confidence, and metacognitive sensitivity; Meuwese, Loon, Lamme, & Fahrenfort (2014); Mazor, Friston, & Fleming (2020); Kellij, Fahrenfort, Lau, Peters, & Odegaard (2021); Mazor, Moran, & Fleming (2021)]. In all three experiments, we replicated

previous findings of a positive evidence bias in confidence in discrimination of motion direction and relative luminance (Zylberberg, Barttfeld, & Sigman, 2012). In contrast, our understanding of decision and confidence formation in detection has evolved and changed following each experiment, as evident in our pre-registration documents. When considering the results of all three experiments together, we conclude that, similar to discrimination confidence, detection decisions and confidence ratings are also sensitive to a positive evidence bias (we use the word bias here to mean a deviation from equal weighting of positive and negative evidence, and not in the sense of a deviation from rationality). We discuss our findings with respect to recent theoretical proposals regarding the origin of a positive evidence bias in discrimination confidence.

3.2 Experiment 1

3.2.1 Methods

Participants

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 10 participants were recruited via the UCL's psychology subject pool, and gave their informed consent prior to their participation. Each participant performed four sessions of 600 trials each, in blocks of 100 trials. Sessions took place on different days and consisted of 3 discrimination blocks interleaved with 3 detection blocks.

Experimental procedure

The experimental procedure for Exp. 1 largely followed the procedure described in Zylberberg, Barttfeld, & Sigman (2012), Exp. 1. Participants observed a random-dot kinematogram for a fixed duration of 700 ms. In discrimination trials, the direction of motion was one of two opposite directions with equal probability, and participants reported the observed direction by pressing one of two arrow keys on a standard keyboard. In detection blocks participants reported whether there was coherent motion by pressing one of two arrow keys on a standard keyboard. In half of the detection trials dots moved coherently to one of two opposite directions, and in the other half they moved randomly.

In both detection and discrimination blocks, following a decision participants indicated their confidence in their decision. Confidence was reported on a continuous scale ranging from chance to complete certainty. To avoid systematic response biases affecting confidence reports, the orientation (vertical or horizontal) and polarity (e.g., right or left) of the scale was set to agree with the type 1 response. For example, following a down arrow press, a vertical confidence bar was presented where 'guess' is at the center of the screen and 'certain' appeared at the lower end of the scale (see Fig. 3.2).

To control for response requirements, for five subjects the dots moved to the right or to the left, and for the other five subjects they moved upward or downward. The

first group made discrimination judgments with the right and left keys and detection judgments with the up and down keys, and this mapping was reversed for the second group. The number of coherently moving dots ('motion coherence') was adjusted to maintain performance at around 70% accuracy for detection and discrimination tasks independently. This was achieved by measuring mean accuracy after every 20 trials, and adjusting coherence by a step of 3% if accuracy fell below 60% or went above 80%.

Stimuli for discrimination blocks were generated using the exact same procedure reported in Zylberberg, Barttfeld, & Sigman (2012)[1]. Trials started with a presentation of a fixation cross for one second, immediately followed by stimulus presentation. The stimulus consisted of 152 white dots (diameter = 0.14°), presented within a 6.5° circular aperture centered on the fixation point for 700 milliseconds (42 frames, frame rate = 60 Hz). Dots were grouped in two sets of 56 dots each. Every other frame, the dots of one set were replaced with a new set of randomly positioned dots. For a coherence value of c', a proportion of c' of the dots from the second set moved coherently in one direction by a fixed distance of 0.33°, while the remaining dots in the set moved in random directions by a fixed distance of 0.33°. On the next update, the sets were switched, to prevent participants from tracing the position of specific dots. Frame-specific coherence values were sampled for each screen update from a normal distribution centred around the coherence value c with a standard deviation of 0.07, with the constraint that c' must be a number between 0 and 1.

Stimuli for detection blocks were generated using a similar procedure, with the only difference being that on a random half of the trials coherence was set to 0%, without random sampling of coherence values for different frames (see Fig. 1).

At the end of each experimental block (100 trials), participants estimated the number of correct responses they have made.

3.2.2 Randomization

The order and timing of experimental events was determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized in a way that ensures registration time-locking (Mazor, Mazor, & Mukamel, 2019).

3.2.3 Analysis

Experiment 1 was pre-registered (pre-registration document is available here: https://osf.io/z2s93/). Our full pre-registered analysis of behavioural data is available in Appendix D.

Reverse correlation analysis

For the reverse correlation analysis, we followed a procedure similar to the one described in Zylberberg, Barttfeld, & Sigman (2012). For each of the four directions (right,

[1]We reused the original Matlab code that was used for Exp. 1 in Zylberberg et. al. (2012), kindly shared by Ariel Zylberberg.

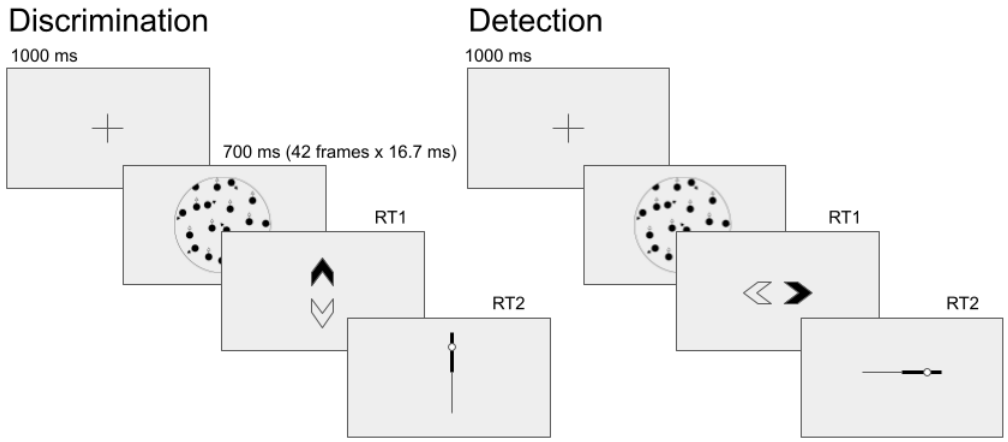


Figure 3.2: Task design for Experiment 1. In both discrimination and detection blocks, participants viewed 700 milliseconds of a random dot motion array, after which they made a keyboard response to indicate their decision (motion direction in discrimination, signal absence or presence in detection), followed by a continuous confidence report using the mouse. 5 participants viewed vertically moving dots and indicated their detection responses on a horizontal scale, and 5 participants viewed horizontally moving dots and indicated their detection responses on a vertical scale.

left, up and down), we applied two spatiotemporal filters to the frames of the dot motion stimuli as described in previous studies (Adelson & Bergen, 1985; Zylberberg, Barttfeld, & Sigman, 2012). The outputs of the two filters were squared and summed, resulting in a three-dimensional matrix with motion energy in a specific direction as a function of x, y, and time. We then took the mean of this matrix across the x and y dimensions to obtain an estimate of the overall temporal fluctuations in motion energy in the selected direction. Additionally, for every time point we extracted the variance along the x and y dimensions, to obtain a measure of temporal fluctuations in spatial variance. Using this filter, we obtained estimates of temporal fluctuations in the mean and variance of motion energy for upward, downward, leftward and rightward motion within each trial. Given a high correlation between our mean and variance estimates, we focused our analysis on the mean motion energy.

In order to distill random fluctuations in motion energy from mean differences between stimulus categories, we subtracted the mean motion energy from trial-specific motion energy vectors. The mean motion energy vectors were extracted at the group level, separately for each motion coherence level and as a function of motion direction. We chose this approach instead of the linear regression approach used by Zylberberg,

Barttfeld, & Sigman (2012) in order to control for nonlinear effects of coherence on motion energy.

Statistical inference

Statistics were extracted separately for each participant, and group-level inference was then performed on the first-order statistics. T-test Bayes factors were used to quantify the evidence for the null when appropriate, using a Jeffrey-Zellner-Siow Prior for the null distribution, with a unit prior scale (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

3.2.4 Results

Response accuracy

Overall proportion correct was 0.74 in the discrimination and 0.72 in the detection task. Performance for discrimination was significantly higher than for detection ($M_d = 0.02$, 95% CI [0.00, 0.04], $t(9) = 2.43$, $p = .038$). This difference in task performance reflected a slower convergence of the staircasing procedure for the discrimination task during the first session. When discarding all data from the first session and analyzing only data from the last three sessions (1800 trials per participant), task performance was equated between the two tasks at the group level ($M_d = 0.00$, 95% CI [−0.02, 0.02], $t(9) = -0.05$, $p = .962$; $\mathrm{BF}_{01} = 3.24$). In order to avoid confounding differences between discrimination and detection decision and confidence profiles with more general task performance effects, the first session was excluded from all subsequent analyses.

Overall properties of response time and confidence distributions

In detection, participants were more likely to respond 'yes' than 'no' (mean proportion of 'yes' responses: $M = 0.59$, 95% CI [0.53, 0.64], $t(9) = 3.45$, $p = .007$). We did not observe a consistent response bias for the discrimination data (mean proportion of 'rightward' or 'upward' responses: $M = 0.52$, 95% CI [0.47, 0.57], $t(9) = 1.00$, $p = .344$).

Replicating previous studies (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2021; Mazor, Friston, & Fleming, 2020; Mazor, Moran, & Fleming, 2021; Meuwese, Loon, Lamme, & Fahrenfort, 2014), we find the typical asymmetries between detection 'yes' and 'no' responses in response time, overall confidence, and the alignment between subjective confidence and objective accuracy (also termed metacognitive sensitivity, here measured as the area under the response-conditional type 2 ROC curve; see Fig. 3.3). 'No' responses were slower compared to 'yes' responses (median difference: 85.37 ms), and accompanied by lower levels of subjective confidence (mean difference of 0.08 on a 0-1 scale). Metacognitive sensitivity was higher for detection 'yes' compared with detection 'no' responses (mean difference in area under the curve units: 0.11). No difference in response time, confidence, or metacognitive sensitivity was found

between the two discrimination responses. For a detailed statistical analysis of these behavioural asymmetries see Appendix D.1.1.

Reverse Correlation

Random fluctuations in motion energy made it possible to apply reverse correlation to test which stimulus features are incorporated into decisions and confidence ratings in detection and discrimination. Following Zylberberg, Barttfeld, & Sigman (2012), our statistical analysis focused on the first 300 milliseconds after stimulus onset.

Discrimination Reverse correlation analysis quantified the effect of random fluctuations in motion energy on the probability of responding 'right' and 'left' (or 'up' and 'down'), and the temporal dynamics of decision formation. Similar to the results obtained by Zylberberg, Barttfeld, & Sigman (2012)., participants' decisions were sensitive to motion energy fluctuations during the first 300 milliseconds of the trial ($t(9) = 7.73$, $p < .001$; see Fig. 3.4, left panels). The symmetry of the two time courses around the x axis does not by itself entail an equal contribution of negative and positive evidence to the final decision, because due to the demeaning procedure, with enough trials negative and positive evidence at each time point should mathematically sum to zero. Instead, we tested the contribution of motion energy in the true and opposite directions of motion (defined with respect to the stimulus, and independently of decision) to discrimination decisions. Fluctuations in motion energy in both directions contributed significantly to discrimination decisions ($t(9) = 8.38$, $p < .001$), with no significant difference between them ($t(9) = -0.65$, $p = .529$). In other words, positive and negative evidence equally contributed to discrimination decisions, even when defined independently of the decision.

We then turned to the contribution of motion energy to subjective confidence ratings. The median confidence rating in each experimental session was used to split all motion energy vectors into four groups, according to decision (chosen or unchosen directions) and confidence level (high or low). Confidence kernels for the chosen and unchosen directions were then extracted by subtracting the mean low confidence vectors from the mean high confidence vectors for both the chosen and unchosen directions. We observed a significant effect of motion energy on confidence within this time window ($t(19) = 2.52$, $p = .021$; see Fig. 3.4, right panels). This effect was significantly stronger for motion energy in the chosen direction, compared to the unchosen direction ($t(9) = 2.81$, $p = .020$). In other words, confidence ratings in the discrimination task were more sensitive to positive evidence than to negative evidence. This is a replication of the Positive Evidence Bias observed in Zylberberg, Barttfeld, & Sigman (2012).

Detection Reverse correlation analysis for detection introduces a challenge: while 'no' responses reflect a belief in the absence of any coherent motion, 'yes' responses can result from detection of any type of coherent motion going in either direction (or both). We chose to have two possible motion directions in the detection task in order to prevent participants from making 'no' responses based on significant motion in an

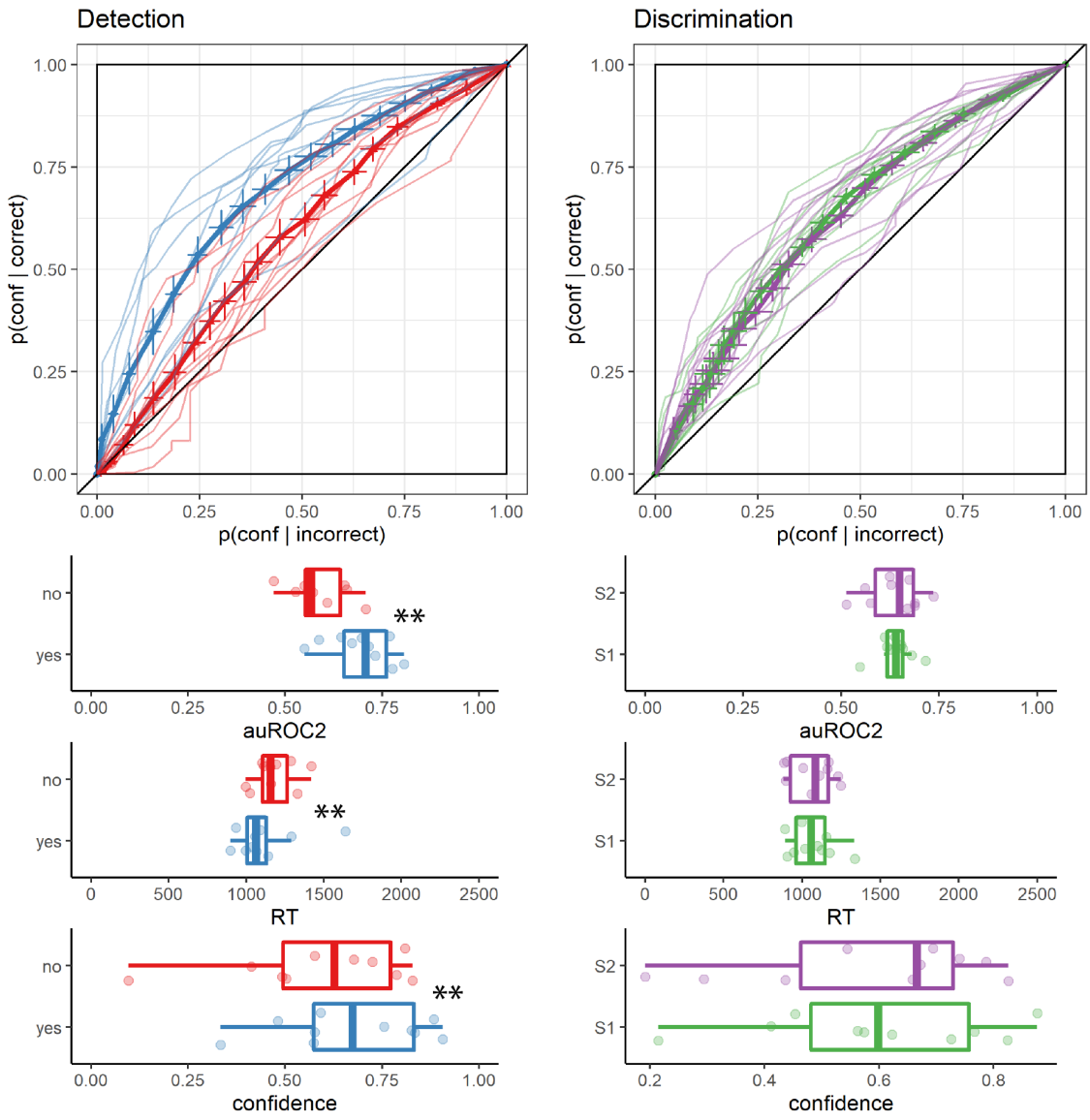


Figure 3.3: Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence, in Exp. 1. Top row: Response conditional type 2 ROC curves for the two tasks and four responses in Exp. 1. The area under the type 2 ROC curve is a measure of metacognitive sensitivity, and the difference in areas between the two responses a measure of metacognitive asymmetry. Single-subject curves are presented in low opacity. Second, third, and fourth rows: distributions of the area under the type 2 ROC curve, median response time, and mean confidence for the four responses, across participants. Box edges and central lines represent the 25, 50 and 75 quantiles. Whiskers cover data points within four inter-quartile ranges around the median. Stars represent significance in a two-sided t-test: **: $p<0.01$, ***: $p<0.001$

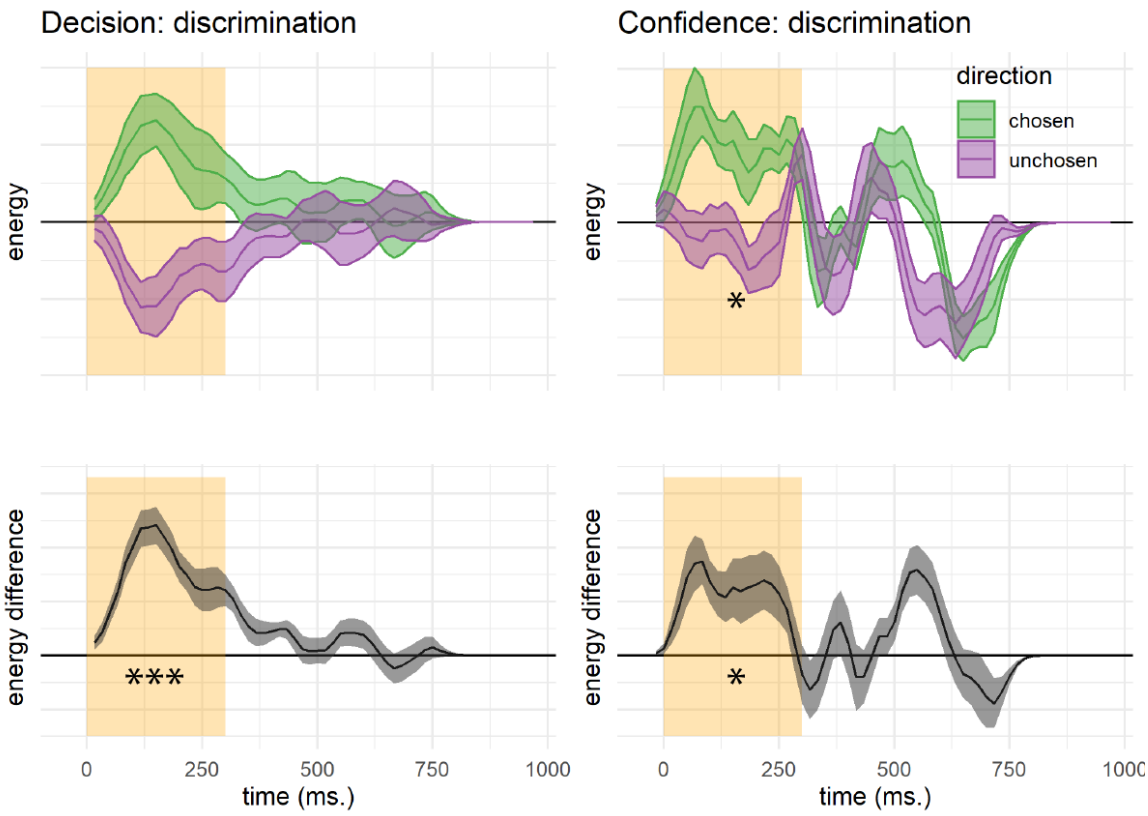


Figure 3.4: Decision and confidence discrimination kernels, Experiment 1. Upper left: motion energy in the chosen (green) and unchosen (purple) direction as a function of time. Lower left: a subtraction between energy in the chosen and unchosen directions. Upper right: confidence effects for motion energy in the chosen (green) and unchosen (purple) directions. Lower right: a subtraction between confidence effects in the chosen and unchosen directions. Shaded areas represent the the mean ± one standard error. The first 300 milliseconds of the trial are marked in yellow. Stars represent significance in a two-sided t-test: *: p<0.05, **: p<0.01, ***: p<0.001. In the upper row, stars represent the significance of a positive evidence bias in evidence weighting.

unexpected direction. While this choice ensured that participants cannot trivially accumulate evidence for absence, it also made the reverse correlation analysis more difficult, as we did not have full access to participants' beliefs about the stimulus when they responded 'yes.'

As a first approximation, we tested whether sum motion energy along the relevant dimension (horizontal or vertical), regardless of direction (up/down or left/right), affected the probability of a 'yes' response. Sum motion energy did not have a significant effect on participants' responses during the first 300 milliseconds ($t(9) = 1.23$, $p = .249$) or at any other time point. The effect of sum motion energy on decision confidence during the first 300 milliseconds was positive and marginally significant ($t(9) = 2.15$,

$p = .060$). Response-specific effects of sum motion energy on decision confidence were not significant for either response.

Detection signal trials

A failure to find significant effects of sum motion energy on detection decisions and confidence may be due to the fact that participants were sensitive to relative evidence (e.g., 'more dots are moving to the right than to the left') rather than to the sum motion along the relevant axis ('many dots are moving to the right and to the left'). However, as we mention above, on any single trial, we cannot tell whether a 'yes' response means 'I perceived coherent motion to the right' or 'I perceived coherent motion to the left.' Instead, in order to approximate participants' belief states during 'yes' responses, we focused only on trials in which coherent motion was presented in one of the two directions (signal trials). In these trials, we reasoned that a 'yes' response is most likely to reflect the detection of the true direction of motion. We therefore asked whether fluctuations in the true and opposite directions of motion contributed to detection decision and confidence. This was done by subtracting the motion energy vectors for 'yes' and 'no' responses in the true and opposite motion directions.

Similar to discrimination decisions, detection decisions were most sensitive to perceptual evidence in the first 300 milliseconds of the trial (see Fig. 3.5, left panels). However, in contrast to discrimination, an asymmetric evidence weighting was apparent in the decision itself: when deciding whether a stimulus contained coherent motion, participants were more sensitive to fluctuations in motion energy that strengthened the true direction of motion, in comparison to fluctuations that weakened motion in the opposite direction ($t(9) = 2.31$, $p = .046$).

Motion fluctuations in the first 300 milliseconds of the trial also contributed to confidence in detection 'yes' responses (contrasting high and low confidence hit trials; $t(9) = 6.13$, $p < .001$). But unlike in the discrimination task here we found no positive evidence bias in confidence ratings ($t(9) = 0.11$, $p = .913$). To reiterate, while detection decisions were mostly sensitive to fluctuations in motion energy toward the true direction of motion, confidence in detection 'yes' responses was equally sensitive to fluctuations in the true and opposite directions of motion. Confidence in 'miss' trials was independent of motion energy ($t(9) = 0.16$, $p = .874$). This was true for motion energy in the true direction of motion ($t(9) = 0.12$, $p = .908$) as well as for motion energy in the opposite direction ($t(9) = -0.08$, $p = .941$). However, and to anticipate the results of Exp. 3 presented below, we note that this equal weighting of positive and negative evidence in detection confidence was not replicated in an experiment designed to directly test this surprising result with an experimental manipulation.

3.3 Experiment 2

In Exp. 1, we replicated previous observations of a positive evidence bias in discrimination confidence, such that evidence in support of a decision was given more weight

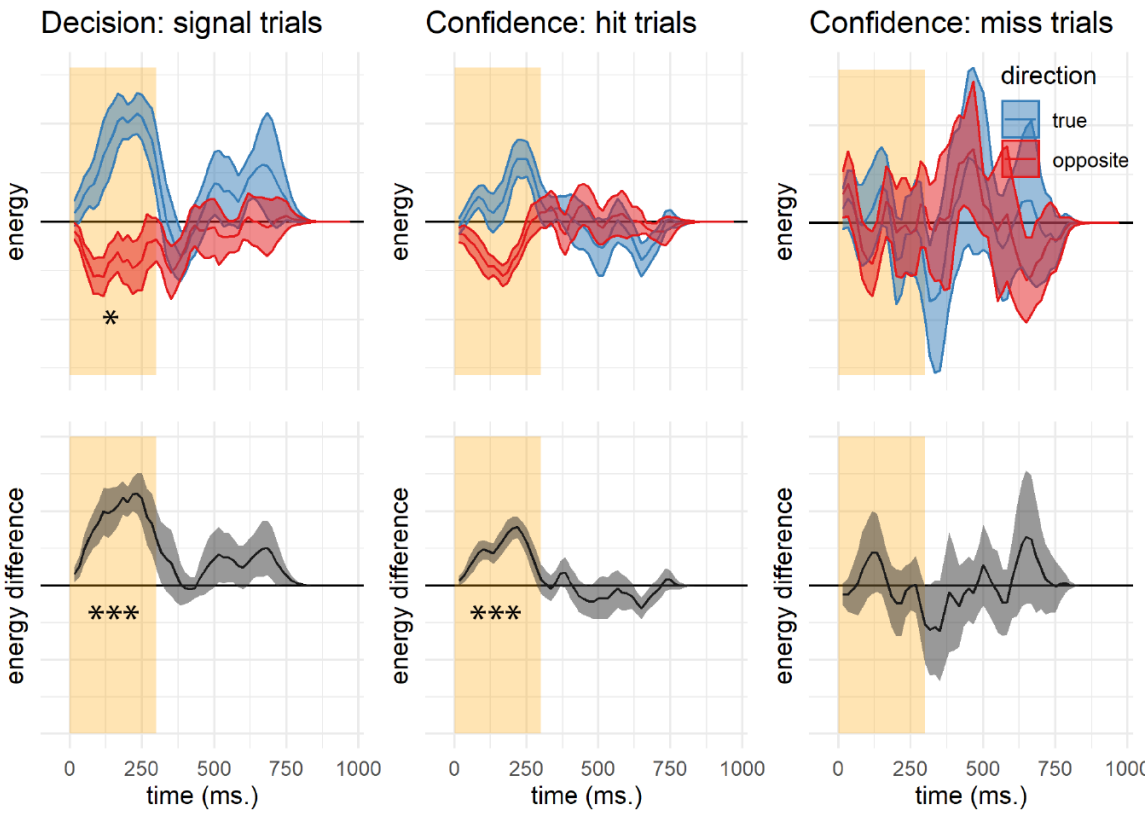


Figure 3.5: Decision and confidence detection kernels in signal trials, Experiment 1. Upper left: difference in motion energy between 'yes' and 'no' responses in the true (blue) and opposite (red) directions as a function of time. Upper middle and right: confidence effects for motion energy in the true and opposite directions for 'yes' and 'no' responses, respectively. Lower panels: the subtraction of decision and confidence kernels for the true and opposite directions. Shaded areas represent the the mean ± one standard error. The first 300 milliseconds of the trial are marked in yellow. Stars represent significance in a two-sided t-test: *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$. In the upper row, stars represent the significance of a positive evidence bias in evidence weighting.

in the construction of confidence than evidence against it. In contrast, in detection a positive evidence bias was apparent for the decision, but not for the confidence kernels. Equal weighting of positive and negative evidence suggests that detection confidence followed not the presence or absence of a signal, but the clarity of its identity. Furthermore, confidence in detection 'no' responses was not at all affected by fluctuations in motion energy.

In Exp. 2 we tested the robustness of these findings by employing a different type of stimuli (flickering patches) and mode of data collection (a ~10 minute online experiment). Our pre-registered objectives (documented here: https://osf.io/8u7dk/) were to first, replicate a positive evidence bias in discrimination, second,

replicate the absence of a positive evidence bias in detection confidence ratings, and third, replicate the absence of an effect for either positive or negative evidence on confidence in 'no' judgments.

3.3.1 Methods

Participants

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 147 participants (median reported age: 32; range: [19-78]) were recruited via Prolific (prolific.co), and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. Following our pre-registration, we aimed to collect data until we had reached 100 included participants based on our pre-specified inclusion criteria (see https://osf.io/8u7dk/). Our final data set includes observations from 102 included participants. The entire experiment took around 10 minutes to complete. Participants were paid £1.25 for their participation, equivalent to an hourly wage of £7.5.

Experimental paradigm

The experiment was programmed using the jsPsych and P5 JavaScript packages (De Leeuw, 2015; McCarthy, 2015), and was hosted on a JATOS server (Lange, Kuhn, & Filevich, 2015). It consisted of two tasks (Detection and Discrimination) presented in separate blocks. A total of 56 trials of each task were delivered in 2 blocks of 28 trials each. The order of experimental blocks was interleaved, starting with discrimination.

The first discrimination block started after an instruction section, which included instructions about the stimuli and confidence scale, four practice trials and four confidence practice trials. Further instructions were presented before the second block. Instruction sections were followed by multiple-choice comprehension questions, to monitor participants' understanding of the main task and confidence reporting interface. To encourage concentration, feedback was delivered at the end of the second and fourth blocks about overall performance and mean confidence in the task.

Importantly, unlike the lab-based experiment, there was no calibration of difficulty for the two tasks. The rationale for this is that in Exp. 1 perceptual thresholds for motion discrimination were highly consistent across participants, and staircasing took a long time to converge. Furthermore, in Exp. 1 we aimed to control for task difficulty, but this introduced differences between the stimulus intensity in detection and discrimination. To complement our findings, here we aimed to match stimulus intensity between the two tasks, and accept that task performance might vary.

Trial structure In discrimination blocks, trial structure closely followed Exp. 2 from Zylberberg, Barttfeld, & Sigman (2012), with a few adaptations. Following a fixation cross (500 ms), two sets of four adjacent vertical gray bars were presented as a rapid serial visual presentation (RSVP; 12 frames, presented at 25Hz), displayed to the

left and right of the fixation cross (see Fig. 3.6). On each frame, the luminance of the bars was randomly sampled from a Gaussian distribution with a standard deviation of 10/255 units in the standard RGB 0-255 coordinate system. The average luminance of one set of bars was that of the background (128/255). The average luminance of the other set was 133/255, making this patch brighter on average. Participants then reported which of the two sets was brighter on average using the 'D' and 'F' keys on the keyboard. After their response, they rated their confidence on a continuous scale, by controlling the size of a colored circle with their mouse. High confidence was mapped to a big, blue circle, and low confidence to a small, red circle. To discourage hasty confidence ratings, the confidence rating scale stayed on the screen for at least 2000 milliseconds. Feedback about response accuracy was delivered after the confidence rating phase. Detection blocks were similar to discrimination blocks,

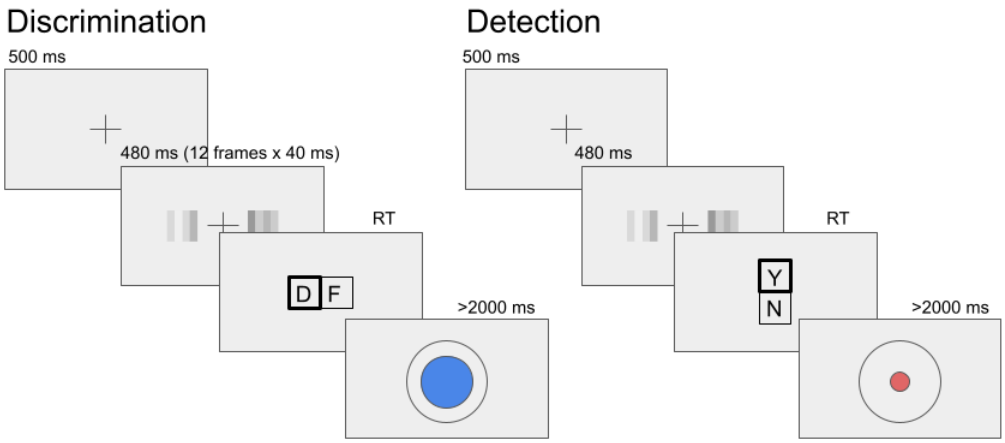


Figure 3.6: Task design for Experiment 2. In both tasks, participants viewed 480 milliseconds of two flicketing patches, after which they made a keyboard response to indicate which of the patches was brighter (discrimination) or whether any of the patches was brighter than the background (detection).

with the exception that decisions were made about whether the average luminance of either of the two sets was brighter than the gray background, or not. In 'different' trials, the luminance of the four bars in one of the sets was sampled from a Gaussian distribution with mean 133/255, and the luminance of the other set from a Gaussian distribution with mean 128/255. In 'same' trials, the luminance of both sets was sampled from a distribution centered at 128/255. Decisions in Detection trials were reported using the 'Y' and 'N' keys. Confidence ratings and feedback were as in the discrimination task.

3.3.2 Randomization

The order and timing of experimental events was determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized in a way that ensures registration time-locking (Mazor, Mazor, & Mukamel, 2019).

3.3.3 Results

Response accuracy

Overall proportion correct was 0.85 in the discrimination and 0.67 in the detection task. Performance for discrimination was significantly higher than for detection ($M_d = 0.18$, 95% CI [0.16, 0.20], $t(101) = 18.01$, $p < .001$). Unlike in Exp. 1, where we aimed to control for task difficulty, here we decided to match stimulus intensity between the two tasks, so a difference between detection and discrimination performance was expected (Wickens, 2002, p. 104).

Overall properties of response and confidence distributions

Similar to Exp. 1, participants were more likely to respond 'yes' than 'no' in the detection task (mean proportion of 'yes' responses: $M = 0.54$, 95% CI [0.53, 0.56], $t(101) = 4.78$, $p < .001$). We did not observe a consistent response bias in discrimination (mean proportion of 'right' responses: $M = 0.50$, 95% CI [0.48, 0.51], $t(101) = -0.62$, $p = .537$).

As in Exp. 1, we also found behavioural asymmetries between the two detection responses (see Fig. 3.7), with 'yes' responses being faster (median difference of 77.12 ms) and accompanied by higher levels of confidence (mean difference of 0.10 on a 0-1 scale). Unlike in Exp. 1, here we found no evidence for a difference in metacognitive sensitivity between 'yes' and 'no' responses (mean difference of 0.02 in AUC units). No asymmetries were observed between the two discrimination responses. For a detailed statistical analysis see Appendix D.2.1.

Reverse Correlation

Stimuli in Exp. 2 consisted of two flickering patches, each comprising 4 gray bars presented for 12 frames. Together, this summed to 96 random luminance values per trial, which we subjected to reverse correlation analysis, following the analysis procedure of Exp 2. in Zylberberg, Barttfeld, & Sigman (2012).

Discrimination decisions First, we asked whether random fluctuations in luminance had an effect on participants' discrimination responses. Similar to the results obtained by Zylberberg et. al., discrimination decisions were sensitive to fluctuations in luminance during the first 300 milliseconds of the trial ($t(101) = 10.98$, $p < .001$; see Fig. 3.8, left panels). As per our approach to the reverse correlation analysis of Exp. 1, in order to test for decision biases we need to divide evidence based on a criterion that is independent of participants' decision. When sorting evidence based on the

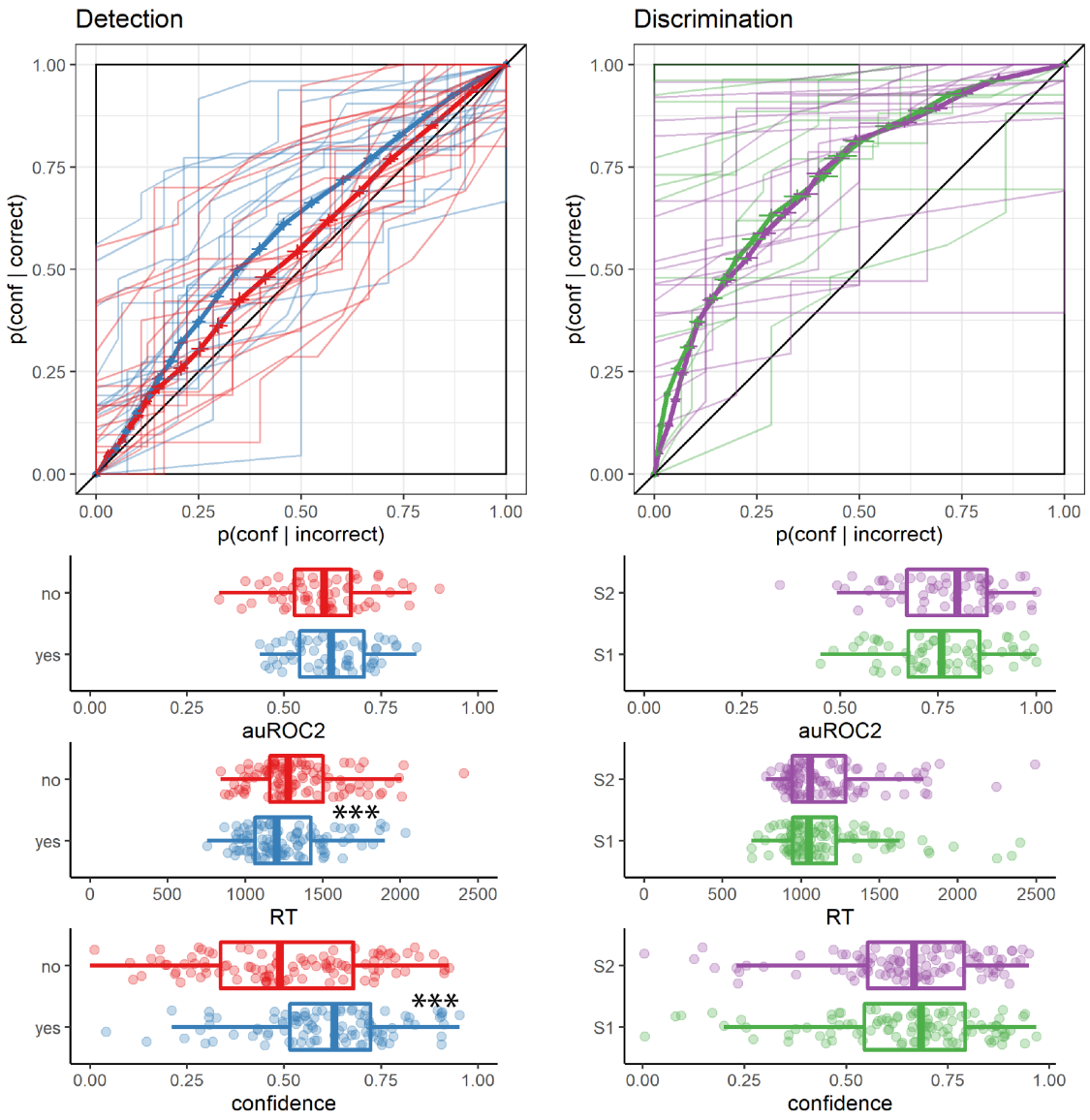


Figure 3.7: Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence, in Exp. 2. Same conventions as in Fig. 3.3.

location of the true signal, participants' decisions were significantly more sensitive to fluctuations in luminance in the non-signal compared with the signal stimulus within the first 300 milliseconds of the trial ($t(100) = -2.29$, $p = .024$). Importantly, this asymmetry (effectively, a negative evidence bias) is in the opposite direction to what we later find in discrimination confidence and detection decisions.

Discrimination confidence We observed a significant effect of luminance on confidence within the first 300 milliseconds of the stimulus ($t(100) = 7.14$, $p < .001$; see Fig. 3.8, right panels). Replicating Zylberberg, Barttfeld, & Sigman (2012), this

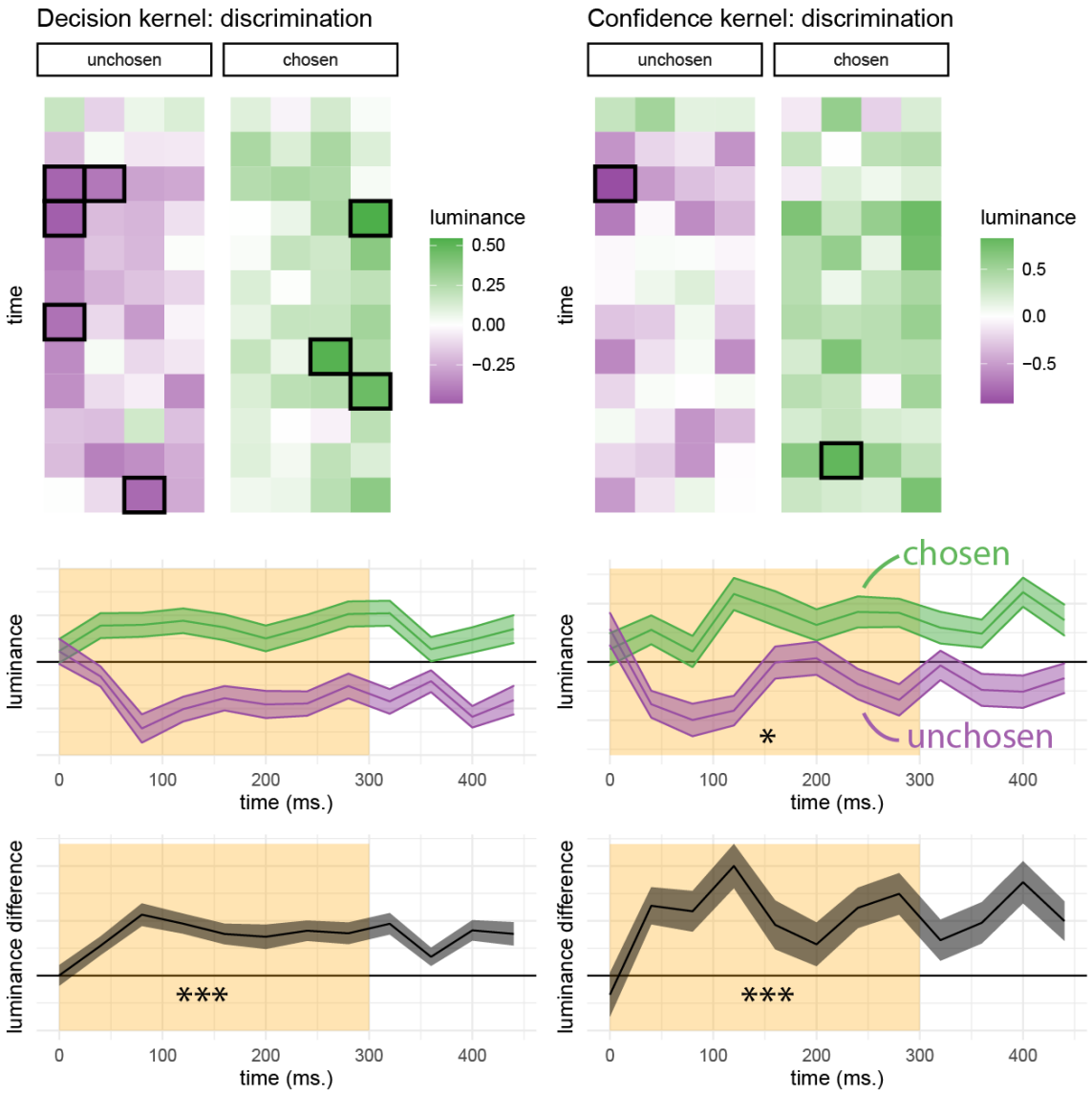


Figure 3.8: Decision and confidence discrimination kernels, Experiment 2. Upper panels: decision (left) and confidence (right) kernels for the flickering patch stimuli. Black frames signify a significant effect at the 0.05 significance level controlling for family-wise error rate across the 48 (12 timepoint x 4 positions) comparisons. middle panels: decision and confidence kernels, averaged across the four bars to yield a single timecourse for the chosen (green) and unchosen (purple) stimuli. Lower panels: subtraction of luminance timecourses for the chosen and unchosen stimuli. Same plotting conventions as Fig. 3.4.

effect was significantly stronger for luminance in the chosen stimulus, compared to the unchosen one ($t(100) = 2.56$, $p = .012$), consistent with a positive evidence bias.

Detection decisions We pooled luminance values from both right and left stimuli and contrasted the resulting values as a function of detection response. Sum luminance had a significant effect on participants' detection responses during the first 300 milliseconds ($t(101) = 6.10$, $p < .001$; see Fig. 3.9, left panel), suggesting that participants were sensitive to sum evidence in their detection responses, as expected from a model in which detection is rationally based on the likelihood ratio between signal presence and absence (see Fig. 3.1).

We then asked if overall luminance had an effect on decision confidence, such that participants are more confident in their 'yes' responses for brighter displays, and more confident in their 'no' responses for darker displays. Interestingly, and in contrast with our hypothesis, sum luminance had no effect on decision confidence in 'yes' responses ($t(99) = -0.02$, $p = .983$), but had a significant negative effect on confidence in 'no' responses ($t(99) = -2.43$, $p = .017$; see Fig. 3.9, middle and right panels). However, to again anticipate our pre-registered Exp. 3, we find an effect of sum luminance on both 'yes' and 'no' responses, suggesting that this surprising absence of an effect for 'yes' responses is likely to be a type-2 error.

3.3.4 Detection signal trials

We next focused on detection signal trials. This analysis diverged from our pre-registered plan; for the pre-registered analysis, please see Appendix section D.4. In these trials, we could separate stimuli to a signal channel (the brighter, target stimulus) and a noise channel (the darker, non-target stimulus), and ask how random variability in luminance in each channel affected detection decisions and confidence. As in Exp. 1, a positive evidence bias effect in detection was apparent in the decision itself: when deciding whether one of the flickering patches was brighter, participants were more sensitive to positive noise in the brighter patch than to negative noise in the darker patch ($t(101) = 6.10$, $p < .001$). Random fluctuations in luminance in the first 300 milliseconds of the trial also contributed to confidence in detection 'yes' responses (hit trials; $t(99) = 5.08$, $p < .001$). In contrast, confidence in 'no' responses was negatively sensitive to the overall luminance of the display. A negative effect of luminance on confidence in 'no' responses was significant for the non-target stimulus ($t(98) = -2.64$, $p = .010$), and marginally significant for the target stimulus ($t(98) = -1.67$, $p = .099$). Consistent with the results of Exp. 1, confidence in 'miss' trials was independent of the contrast in luminance between the right and left stimuli ($t(98) = 1.26$, $p = .210$). Importantly, for both stimuli higher confidence in these trials was associated with lower luminance values, in line with our observation that confidence in detection 'no' responses was based on the overall darkness of the display, rather than on relative evidence. Finally, and similar to the results of Exp. 1, detection confidence was not susceptible to a positive evidence bias ($t(99) = -0.12$, $p = .901$). Exp. 3 below was designed to replicate this surprising symmetric weighting of positive and negative evidence in detection confidence (the absence of a positive evidence bias) in a highly-powered design.

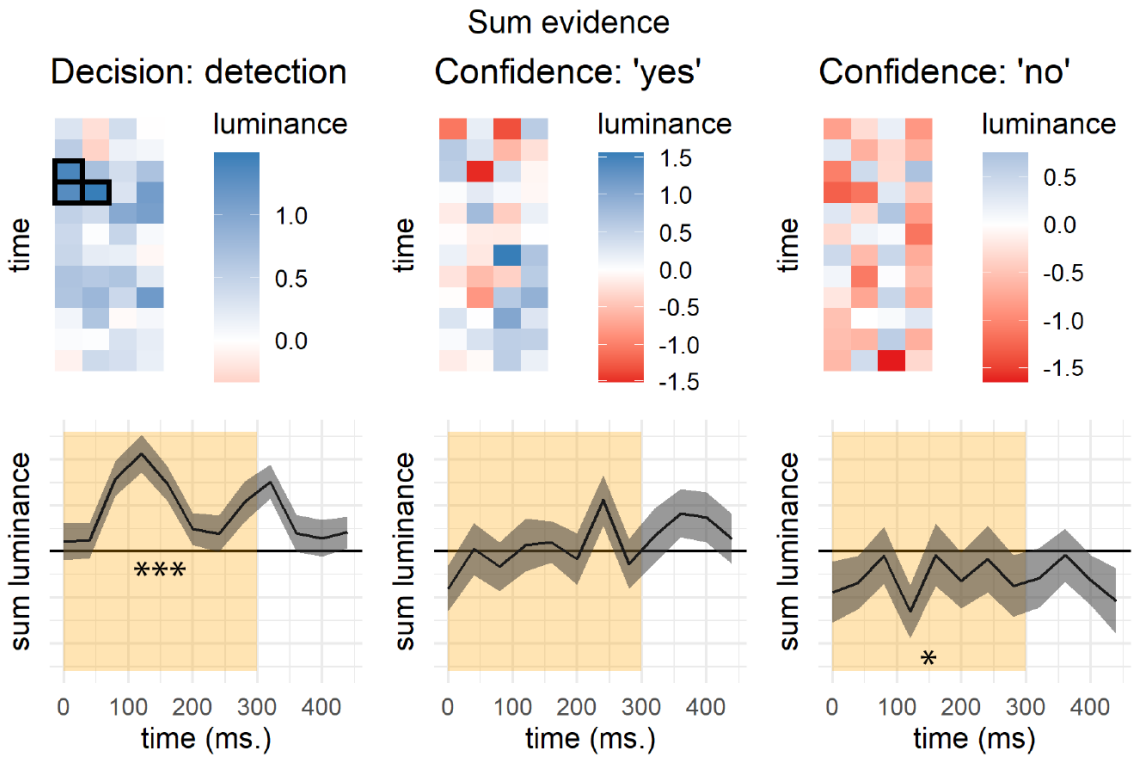


Figure 3.9: Decision and confidence detection kernels, Experiment 2. Upper panels: decision (left) and confidence (right) kernels for the flickering patch stimuli, showing the effect of sum evidence (sum luminance across both stimuli) on decisions and confidence. Black frames signify a significant effect at the 0.05 significance level controlling for family-wise error rate across the 48 (12 timepoint x 4 positions) comparisons. Lower panels: decision and confidence kernels, averaged across the four bars to yield a single timecourse for the difference in luminance effects in 'yes' and 'no' responses. Same conventions as in Fig. 3.8.

3.4 Experiment 3

In Exp. 3 we aimed to replicate our findings using an experimental manipulation, in addition to employing reverse-correlation analysis to random variations between stimuli. Our pre-registered objectives (see our pre-registration document: https://osf.io/hm3fn/) were to first, replicate a positive evidence bias in discrimination, second, replicate a positive evidence bias in detection decisions, and third, replicate the absence of a positive evidence bias in detection confidence.

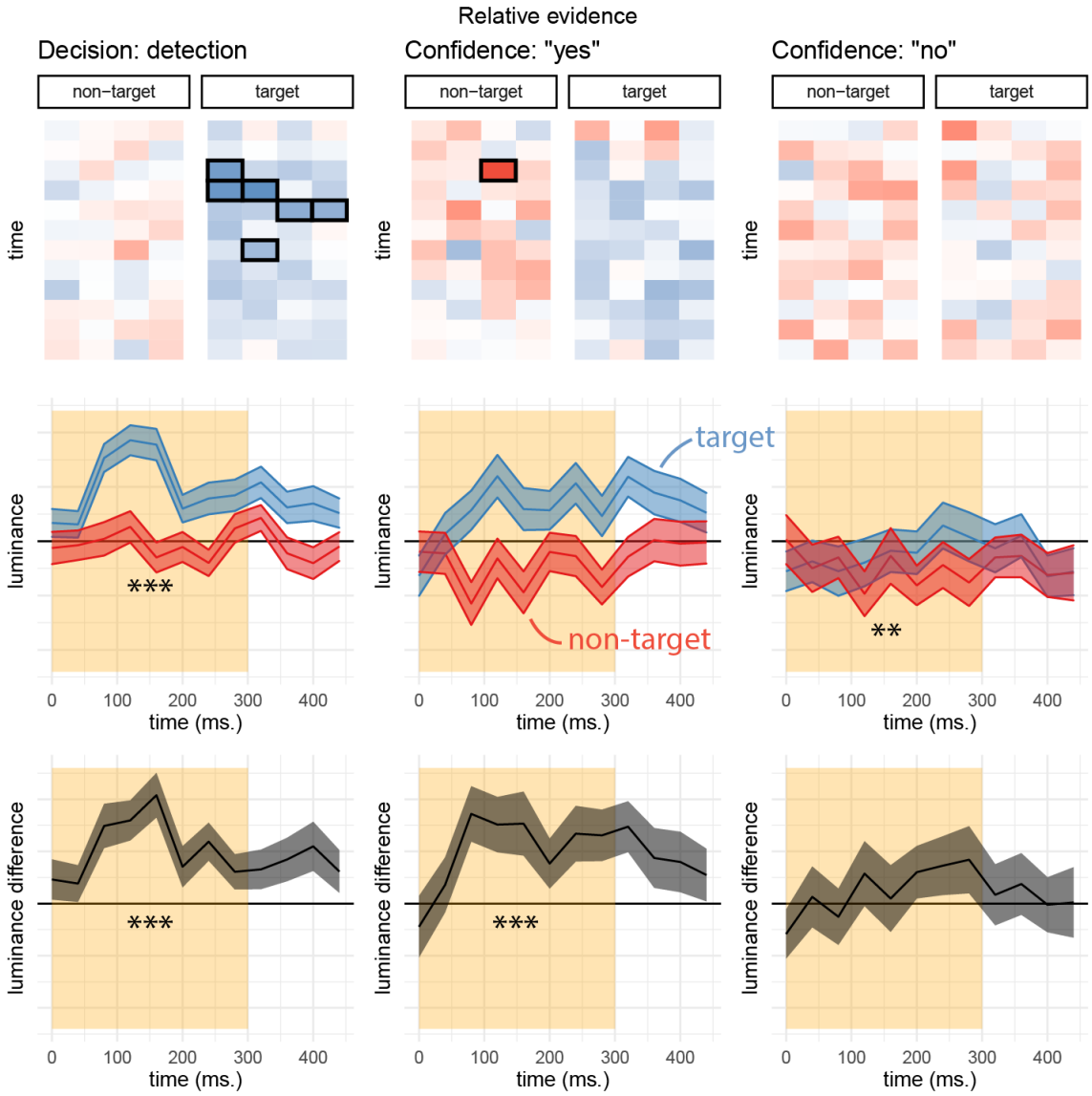


Figure 3.10: Decision and confidence kernels for detection signal trials, Experiment 2. Upper left: mean difference in luminance between 'yes' and 'no' responses for the target and non-target stimuli. Upper middle and right panels: mean effect of luminance on confidence in the target and non-target stimuli, in 'yes' and 'no' responses. Middle panels: the effects of luminance on decision and confidence, averaged across the four spatial locations. Lower panels: a subtraction between the effects of luminance in the target and non-target stimuli. Same conventions as Fig. 3.8

3.4.1 Methods

Participants

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 173 participants (median reported age: 31; range: [18-71]) were recruited via Prolific (prolific.co), and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. Following our pre-registration, we aimed to collect data until we had reached 100 included participants based on our pre-specified inclusion criteria (see https://osf.io/hm3fn/). Our final data set includes observations from 100 included participants. The entire experiment took around 20 minutes to complete. Participants were paid £2.5 for their participation, equivalent to an hourly wage of £7.5.

Experimental paradigm

Experiment 3 was identical to Experiment 2 with two changes. First, on half of the trials (*high-luminance* trials) the luminance of both sets of bars was increased by 2/255 for the entire duration of the display. Second, in order to increase our statistical power for detecting response-specific effects in detection, participants performed four detection blocks and two discrimination blocks. Each block comprised 56 trials. The order of blocks was [detection, discrimination, detection, discrimination, detection, detection] for all participants.

3.4.2 Results

Response accuracy

Overall proportion correct was 0.88 in the discrimination and 0.67 in the detection task. Performance for discrimination was significantly higher than for detection ($M_d = 0.21$, 95% CI [0.19, 0.22], $t(97) = 29.87$, $p < .001$), as expected.

Overall properties of response and confidence distributions

Similar to Experiments 1 and 2, participants were more likely to respond 'yes' than 'no' in the detection task (mean proportion of 'yes' responses: $M = 0.53$, 95% CI [0.51, 0.54], $t(97) = 3.73$, $p < .001$). We did not observe a consistent response bias in discrimination (mean proportion of 'right' responses: $M = 0.50$, 95% CI [0.48, 0.53], $t(97) = 0.46$, $p = .647$).

As in both Experiments 1 and 2, we found behavioural asymmetries between the two detection responses, with 'yes' responses being faster (median difference of 71.81 ms), and accompanied by higher levels of confidence (mean difference of 0.09 on a 0-1 scale). As in Exp. 1, we find a difference in metacognitive sensitivity between 'yes' and 'no' responses (mean difference of 0.03 in AUC units). No asymmetries were observed between the two discrimination responses. For a detailed statistical analysis see Appendix D.3.1.

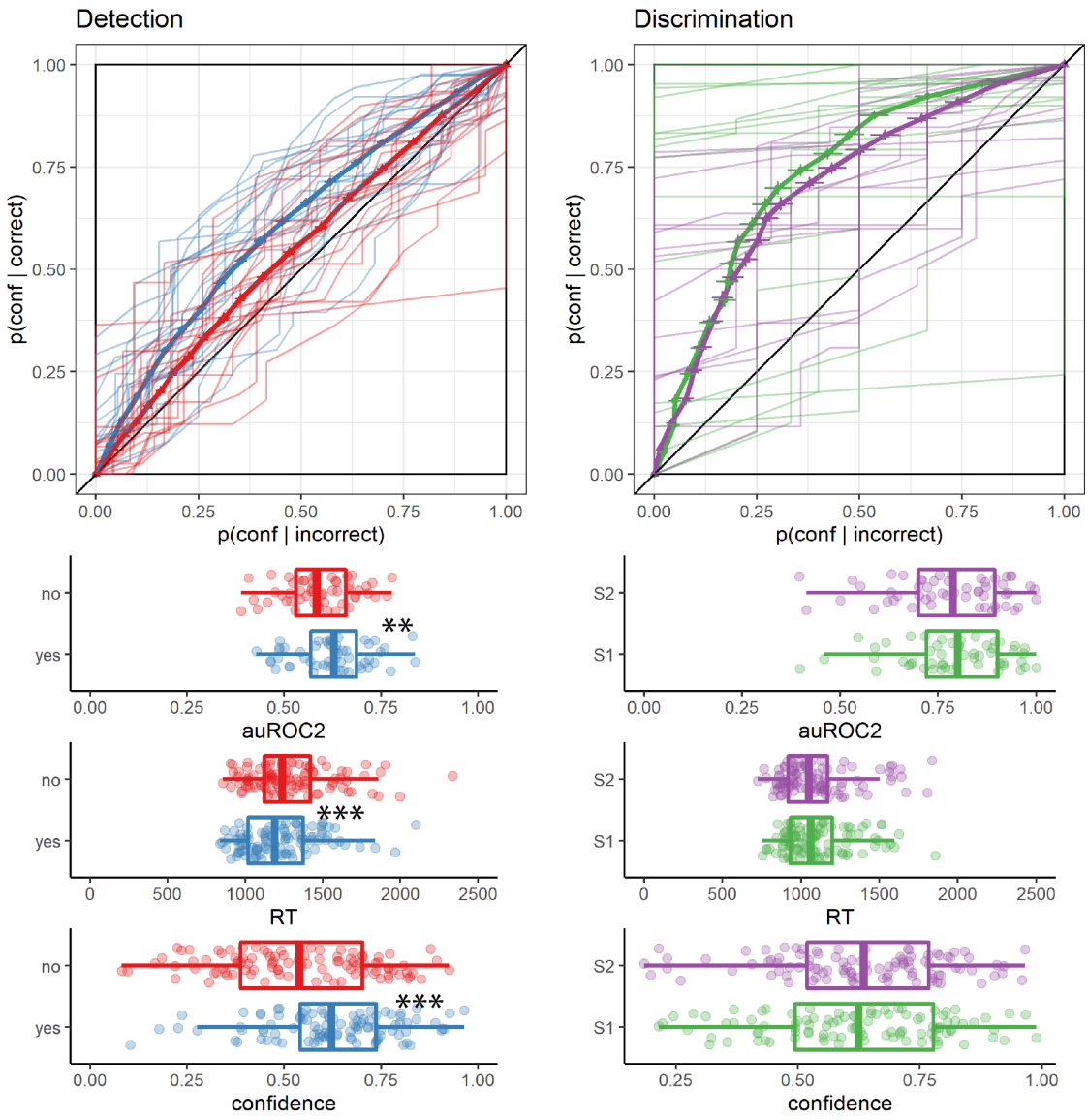


Figure 3.11: Behavioural asymmetries in metacognitive sensitivity, response time, and overall confidence, in Exp. 3. Same conventions as in Fig. 3.3.

Reverse correlation

Discrimination decisions We first focused on reverse correlation analyses that collapsed across high-luminance and standard trials, in order to replicate the same approach used in Exps. 1 and 2. When focusing on standard trials only, the results are qualitatively similar, with the exception of confidence in detection 'no' responses (see Appendix D.3.2). Discrimination decisions were sensitive to fluctuations in luminance during the first 300 milliseconds of the trial ($t(97) = 12.01$, $p < .001$; see Fig. 3.12, left panels). We found no evidence for a positive evidence bias in discrimination decisions, even when grouping evidence based on the location of the true signal rather than

subjects' decisions ($t(97) = 0.83$, $p = .407$).

Discrimination confidence Luminance within the first 300 milliseconds had a significant effect on confidence ratings ($t(97) = 7.23$, $p < .001$; see Fig. 3.12, right panels). A positive evidence bias in discrimination confidence was only marginally significant in this sample ($t(97) = 1.63$, $p = .106$).

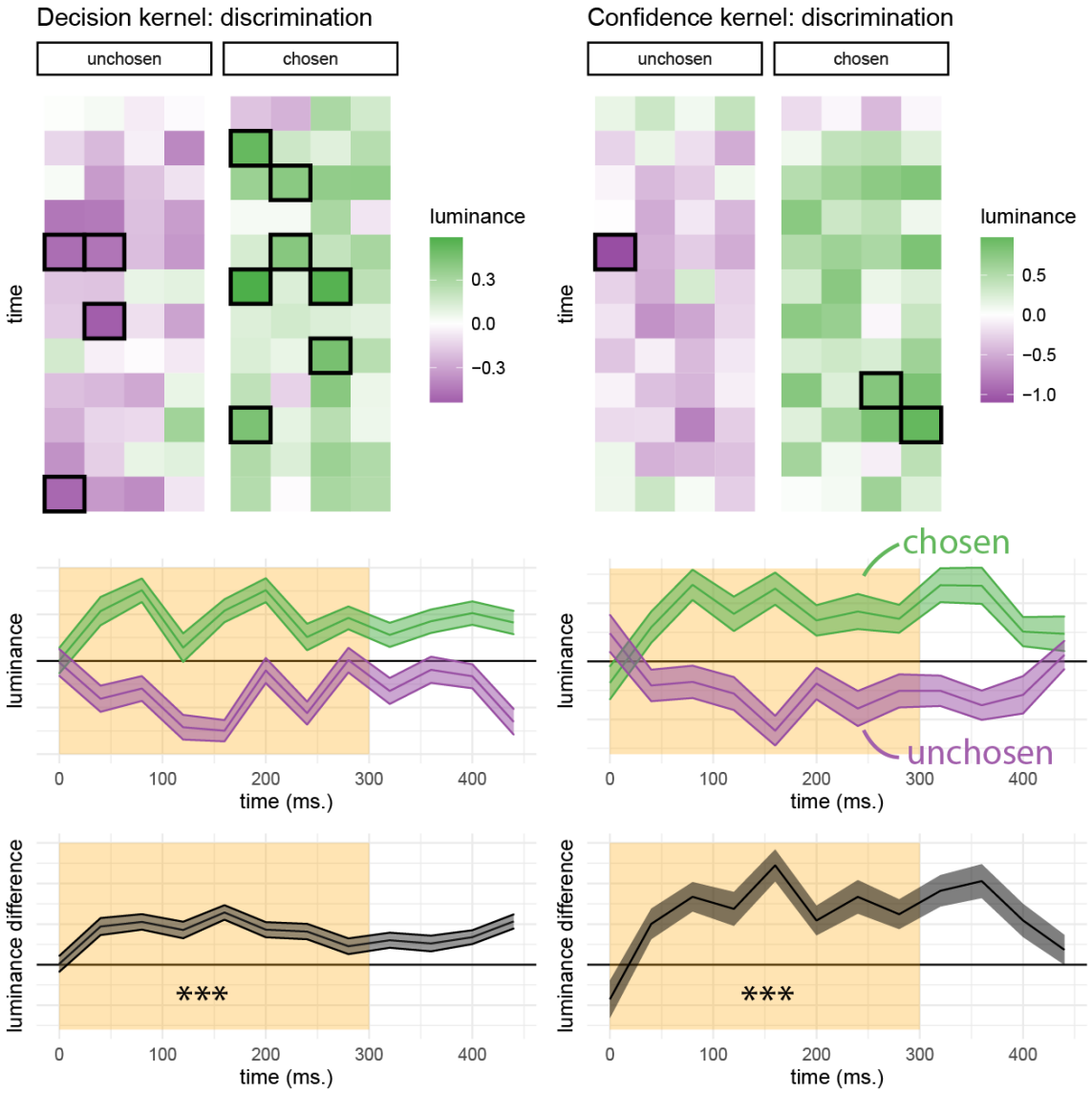


Figure 3.12: Decision and confidence discrimination kernels, Experiment 3. Same conventions as Fig. 3.8.

Detection

Similar to Exp. 2, sum luminance had a significant effect on participants' detection responses during the first 300 milliseconds ($t(97) = 10.94$, $p < .001$; see Fig. 3.13, left

panel). Recall that a surprising finding in Exp. 1 was that sum luminance on detection decisions had no effect on participants' confidence in their judgments of stimulus presence. In contrast, in Exp. 3 sum luminance had a significant positive effect on decision confidence when reporting target presence ('yes' responses; $t(97) = 3.54$, $p = .001$), and a significant negative effect on confidence when reporting target absence ('no' responses; $t(97) = -3.04$, $p = .003$; see Fig. 3.13, middle and right panels).

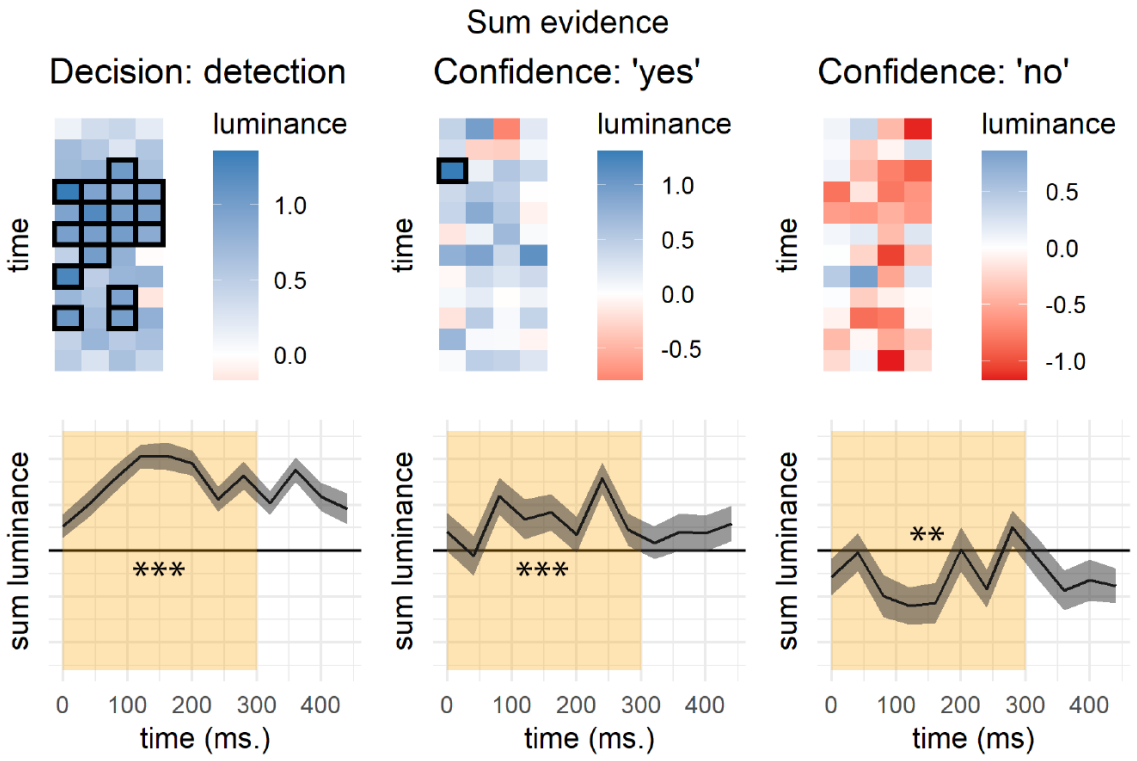


Figure 3.13: Decision and confidence detection kernels, Experiment 3. Same conventions as Fig. 3.9.

Detection signal trials

As in Exp. 2, here we also asked how random variability in luminance in the target (brighter) and non-target (darker) channels affected detection decision and confidence. When deciding whether one of the two flickering patches was brighter than the background, participants were sensitive to positive noise in the target patch more than to negative noise in the non-target patch ($t(97) = 10.94$, $p < .001$), consistent with a positive evidence bias in detection decisions and replicating findings from Exps. 1 and 2. Random fluctuations in luminance in the first 300 milliseconds of the trial also contributed to confidence in detection 'yes' responses (hit trials; $t(97) = 6.07$, $p < .001$). Importantly, however, and in contrast to the results of Exp. 1 and 2, confidence in 'yes' responses was more sensitive to positive evidence than to conflicting evidence ($t(97) = 3.49$, $p = .001$). Together these results are consistent with a positive evidence bias not only for detection decisions, but also for detection confidence.

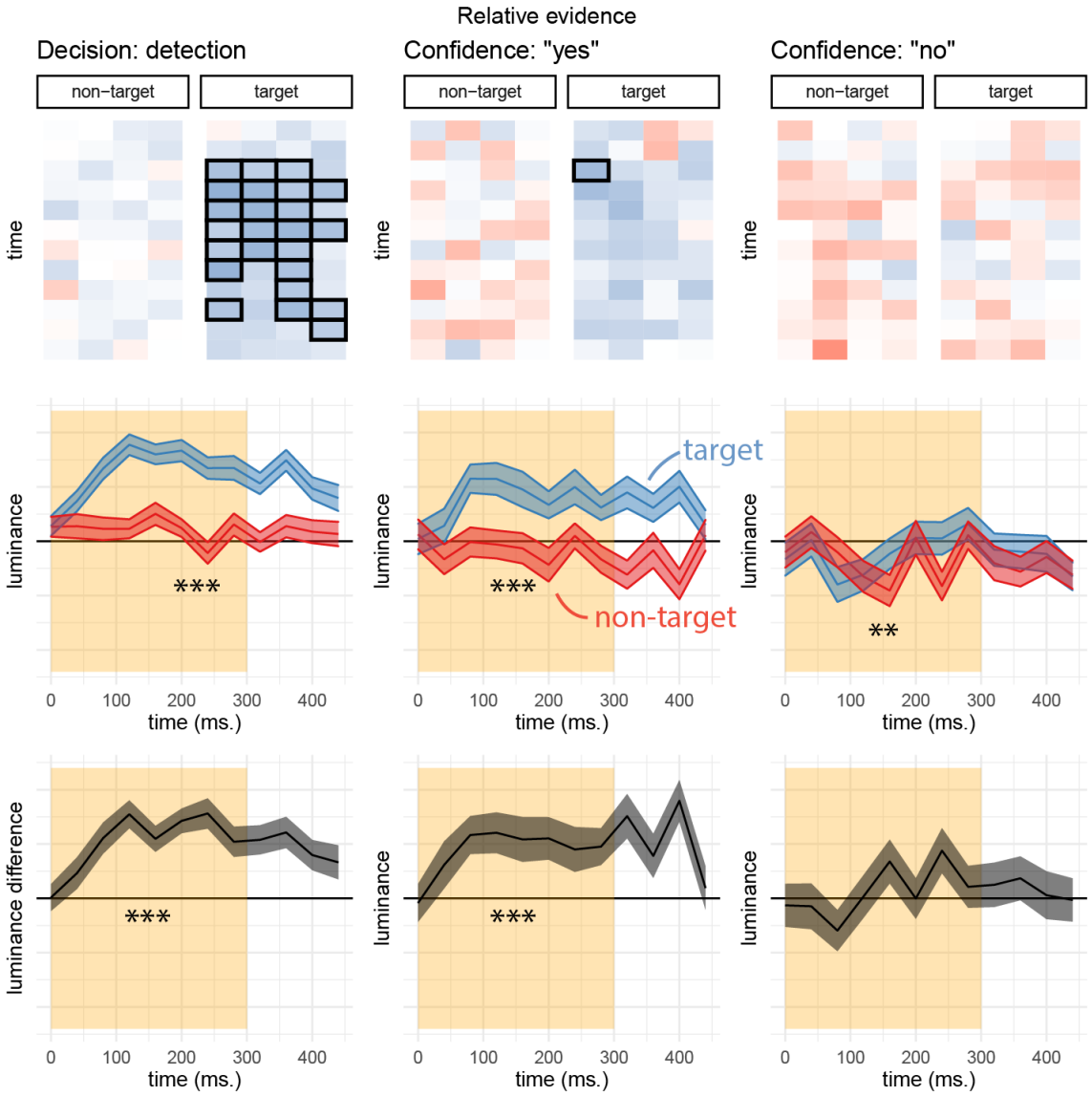


Figure 3.14: Decision and confidence kernels for detection signal trials, Experiment 3. Same conventions as Fig. 3.10.

Confidence in 'miss' trials was independent of the contrast in luminance between the right and left stimuli ($t(96) = 0.89$, $p = .374$) but, as described above, confidence in 'no' responses was sensitive to the overall luminance of the display. This negative effect of luminance on confidence in 'no' responses was significant for the non-target stimulus ($t(96) = -2.91$, $p = .005$), and marginally significant for the target stimulus ($t(96) = -1.67$, $p = .099$). In other words, and similar to our findings in Exp. 2, for both stimuli higher confidence was associated with lower luminance values. This is again consistent with our observation that confidence in judgments about stimulus absence is based on the overall darkness of the display.

Evidence-weighting

In Experiments 1 and 2, confidence in judgments about stimulus presence was invariant to sum evidence (overall motion energy in Exp.1, sum luminance in Exp. 2). This was surprising for two reasons. First, in both cases sum motion energy did have a significant effect on detection decisions. Second, incorporating information about sum evidence into confidence in the presence of a stimulus is rational: a target stimulus is more likely to be present when both stimuli are brighter. As we document above, however, this surprising finding did not replicate in Exp. 3, where detection confidence was now sensitive to the overall brightness of the display. We contrasted the two luminance conditions as a direct experimental test of differential evidence weighting on detection decisions and confidence.

In order to increase statistical power for tests of a positive evidence bias, in Exp. 3 half of the trials consisted of slightly brighter stimuli. In detection, participants were more likely to respond 'yes' on these high-luminance trials ($M = 0.09$, 95% CI [0.07, 0.11], $t(97) = 8.48$, $p < .001$). Overall luminance is a valid cue for signal presence, so relying on it for detection judgments is rational. In discrimination, participants were also more confident in high-luminance trials ($M = 0.02$, 95% CI [0.01, 0.04], $t(97) = 3.28$, $p = .001$), replicating a positive evidence bias for discrimination confidence.

In line with the reverse correlation analysis of Exp. 3 (and in contrast to the findings of Experiments 1 and 2), participants were more confident in their 'yes' responses when overall luminance was higher ($M = 0.02$, 95% CI [0.01, 0.03], $t(97) = 3.01$, $p = .003$). Our pre-registered Bayesian analysis provided strong evidence for the alternative hypothesis that detection confidence is affected by this manipulation ($BF_{10} = 10.90$). Furthermore, this increase in 'yes' response confidence as a function of the brightness manipulation was not significantly different from that observed for discrimination confidence ($M = -0.01$, 95% CI [−0.03, 0.01], $t(97) = -0.55$, $p = .584$).

Finally, and in line with Exp. 2, overall luminance had a significant negative effect on confidence in 'no' responses ($M = -0.02$, 95% CI [−0.03, −0.01], $t(97) = -3.01$, $p = .003$), indicating that participants were more confident in the absence of a target when overall luminance was lower.

3.5 Discussion

In three experiments, we compared the perceptual drivers of decisions and confidence ratings in discrimination and detection, matched for difficulty (Exp. 1) and signal strength (Exp. 2 and 3). In order to measure the contribution of perceptual evidence to confidence in detection and discrimination confidence ratings, we followed Zylberberg, Barttfeld, & Sigman (2012) and applied reverse correlation to noisy stimuli in perceptual decision making tasks. We fully replicated the main results of Zylberberg and colleagues: decisions and confidence were affected by perceptual evidence in the first 300 milliseconds of the trial, peaking at around 200 milliseconds. We also successfully replicated a positive evidence bias for discrimination confidence: confidence in the

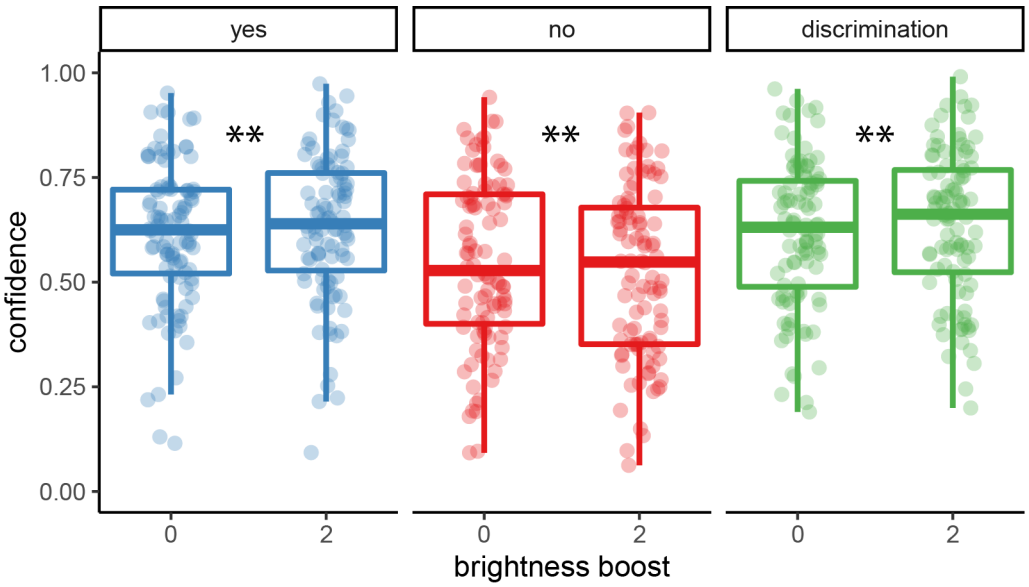


Figure 3.15: Difference in confidence between standard and higher-luminance trials for the three response categories (detection 'yes' and 'no' responses, and discrimination responses) in Exp. 3. Box edges and central lines represent the 25, 50 and 75 quantiles. Whiskers cover data points within four inter-quartile ranges around the median. Stars represent significance in a two-sided t-test: **: $p<0.01$

discrimination task was more affected by supporting than by conflicting evidence. A positive evidence bias in discrimination confidence judgments may indicate that participants adopt a detection-like disposition in their metacognitive judgments, focusing on sum evidence rather than relative evidence when rating their confidence.

In Experiments 1 and 2, detection decisions but not confidence ratings also showed a positive evidence bias: when making a detection response participants mostly ignored random fluctuations in stimulus energy that were not aligned with the true, presented signal, but these fluctuations were later taken into account when rating their confidence. Based on this surprising finding, in Experiment 3 we pre-registered an hypothesis that detection confidence should be equally sensitive to positive and negative evidence. To increase our statistical sensitivity, we doubled the number of detection trials and included a direct manipulation of positive evidence. Results from Experiment 3 provided clear evidence against the hypothesis derived from Experiments 1 and 2, and support an unequal weighting of positive and negative evidence not only for detection decisions, but also for detection confidence judgments.

Previous accounts of the positive evidence bias in discrimination confidence presented it as a heuristic that participants adopt due to cognitive constraints in the face of unreasonably vast representational spaces (Maniscalco, Peters, & Lau, 2016) or due to an asymmetric encoding of signal and noise (Miyoshi & Lau, 2020). A heuristic use of evidence in confidence ratings, but not in the decision itself, in turn implies that different processes are involved in the generation of decisions and confidence ratings, and that participants are in some sense being irrational when discarding relevant

evidence that could be used in constructing confidence.

Similarly, a positive-evidence bias was recently demonstrated in an artificial neural network trained to classify hand-written digits and, in parallel, predict its classification accuracy (Webb, Miyoshi, So, & Lau, 2021). The network was trained on images varying in contrast and visual noise, and later tested on overlays of two digits, varying in contrast only. Under this training regime, the network was more confident for high-contrast images, controlling for classification accuracy. Similar to the heuristic account of Maniscalco, Peters, & Lau (2016), here also a positive evidence bias reflected the application of an inductive bias acquired in real life, or in training, to a test setting where doing so is maladaptive.

An alternative possibility is that a single, Bayes-rational model with a valid prior is governing both choice and confidence ratings, but that the form of that model is yet to be specified. For instance, one possible driver of a positive evidence bias in discrimination confidence ratings is the higher informational value of signal than noise. If the signal channel holds more information about signal identity or stimulus presence, giving more weight to information from this channel is rational. This is the case in unequal-variance SDT settings, where signal is sampled from a wider range of values than noise. As an example, if noise is sampled from a Gaussian distribution with mean 0 and variance 1 and signal from a Gaussian distribution with mean 2 and variance 9, sampling the value 7 (two standard deviations to the right of the signal distribution) is much more informative about the presence or absence of a signal than sampling the value -2 (two standard deviations to the left of the noise distribution), because the first is only likely if sampled from the signal distribution ($\frac{p(x|signal)}{p(x|noise)} > 1,000,000,000$), but the second is likely under both distributions ($\frac{p(x|signal)}{p(x|noise)} = 1.5$).[2] Similarly, if the representation of coherent motion is more variable across trials than the representation of random motion, participants would be rational to give more weight to evidence for coherent motion in one channel than evidence for its absence in the other channel.

Higher variability in the representation of signal is often built into the experiment itself. For example, in our Exp. 1, following Zylberberg, Barttfeld, & Sigman (2012), the number of coherently moving dots was itself randomly determined, sampled from a Gaussian distribution once every four frames. This means that there were two sources of variability for the true direction of motion (variability in the direction of randomly moving dots and variability in the number of coherently moving dots), but only one source of variability for the opposite direction (variability in the direction of randomly moving dots). But even when signal is not made more variable by design, the representation of signal is expected to be more variable due to the Weber-Fechner law (Fechner & Adler, 1860) and the coupling between firing rate mean and variability implied by a Poisson form of neuronal firing rates.

To obtain qualitative predictions for such effects, we simulated a stimulus-dependent noise model (full simulation details, including source code are available in appendix D.5). To model the unequal variance nature of the perception of signal and noise,

[2] In other words, the expected log likelihood ratio when sampling from the signal distribution is higher (in absolute terms) than when sampling from the noise distribution, or $D_{KL}(signal||noise) > D_{KL}(noise||signal)$.

perceptual noise was sampled from a normal distribution with mean 0 and a standard deviation proportional to the exponent of the sensory sample ($x' = x + \epsilon; \epsilon \sim \mathcal{N}(0, 2^x)$). We chose to use the exponent of the sensory sample in order to have positive values only for the standard deviation, but qualitatively similar results are obtained for a linear mapping from sensory samples to sensory noise. A Bayes-rational agent had full knowledge of this generative model for extracting a Log Likelihood Ratio in the process of making a decision and rating their confidence.

This simulation gave rise to a pronounced positive evidence bias in discrimination confidence ratings and in detection decisions (see Fig. 3.16). The agent was more sensitive to variations in the signal channel both for deciding whether a signal was present or not, when rating its confidence in discriminating between two stimulus classes, and when rating its confidence in decisions about stimulus presence. This is in line with our empirical findings. Importantly, in this model decision and confidence ratings are the output of the same Bayes-rational process applied to a situation where perceptual noise scales with signal strength, and do not reflect any suboptimality in evidence weighting.

However, a stimulus-dependent-noise model makes predictions not only for the effect of sum evidence on discrimination and detection confidence ratings, but also for the effect of sum evidence on decision performance (d'). Specifically, if stronger stimuli are also noisier, sum evidence should have a positive effect on confidence, but a negative effect on response accuracy. In contrast with this prediction, in Exp. 3, an increase to the luminance of both stimuli had no effect on accuracy ($M = 0.01$, 95% CI [−0.01, 0.03], $t(97) = 1.06$, $p = .294$), but boosted discrimination confidence nonetheless. Moreover, the effect of overall luminance on confidence was positively correlated with its effect on decision confidence ($r = .32$, 95% CI [.13, .49], $t(96) = 3.33$, $p = .001$) and not negatively correlated as would be expected if degraded accuracy and higher confidence were both driven by higher sum evidence.

Goal-contingent effects also weigh against a Bayes-rational account of the positive evidence bias in discrimination confidence. In a recent study, Sepulveda et al. (2020) presented participants with pairs of dot arrays and asked them to choose the array with more white dots. Participants were more confident when both arrays included more dots, replicating a positive evidence bias. Critically, the experiment also included a second condition, in which participants were asked to choose the array with *fewer* white dots. In this condition, confidence was higher when both arrays included fewer dots: an effect opposite to the positive evidence bias. This effect was also related to participants' information sampling behaviour in the two conditions: they spent more time fixating their gaze at the array containing more dots under typical instructions, but the opposite was the case when they were instructed to select the array with fewer dots. In a Bayes-rational model, evidence weighting should be identical for these two equivalent ways of framing the task instructions. This finding is also inconsistent with heuristic models that are based at variance differences between the encoding of signal and noise, as those are not expected to change as task instructions change.

Instead, our findings, as well as those of Sepulveda et al. (2020), are generally in line with a heuristic account of positive evidence bias that posits limits on cognitive resources when coping with high-dimensional representations (Maniscalco, Peters,

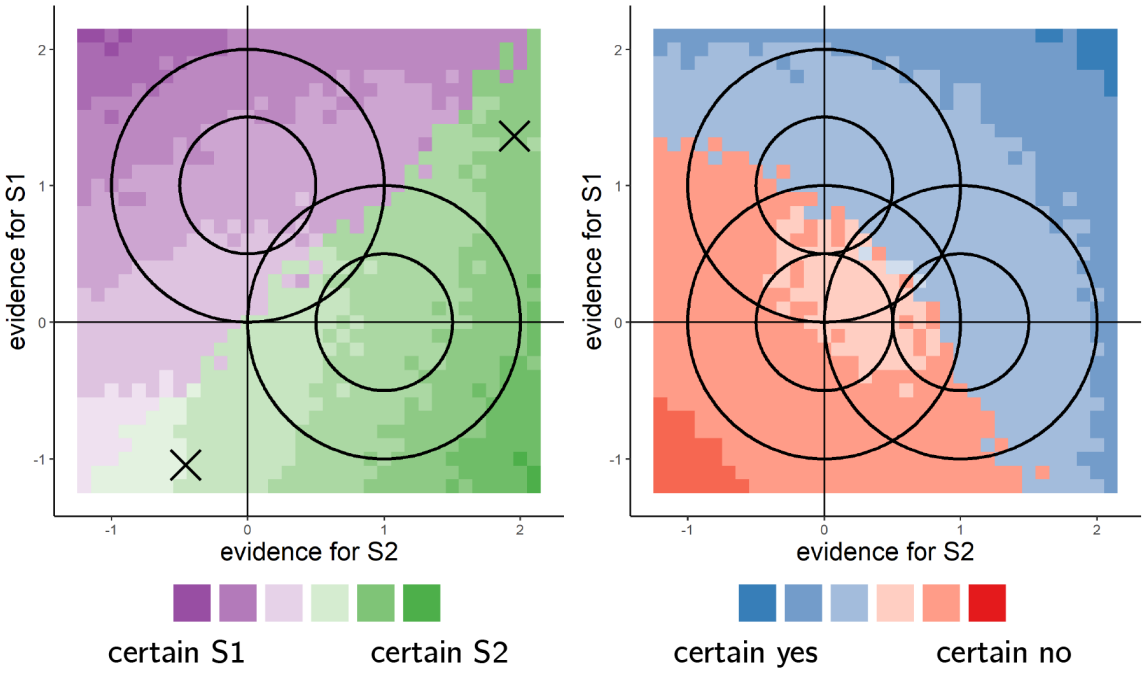


Figure 3.16: Model predictions for a stimulus-dependent noise model. Sensory noise is higher for stronger sensory samples. For each point on the grid, we simulated 200 trials by sampling a sensory sample and extracting a decision and a confidence rating according to the log likelihood ratio of the two hypotheses. Left: two example sensory samples, marked with an X, have the same relative evidence, but absolute evidence is higher for the sample on the upper right. This sample was given a higher-confidence rating, consistent with a positive evidence bias in behaviour, here emerging from a Bayes-rational response and confidence strategy.

& Lau, 2016). Basing confidence on positive evidence in such a world frees agents from the need to consider an infinite number of alternative hypotheses. Similarly, the findings of Sepulveda et al. (2020) can be accounted for by a model in which participants flexibly allocate attention to the choice-consistent dimension of evidence (more vs. fewer dots), while ignoring other dimensions. What constitutes positive evidence is then rationally dependent on an agent's specific goals and attentional set at the time of performing a task.

Chapter 4

Distinct neural contributions to metacognition for detecting (but not discriminating) visual stimuli

Matan Mazor, Karl J. Friston & Stephen M. Fleming

Being confident in whether a stimulus is present or absent (a detection judgment) is qualitatively distinct from being confident in the identity of that stimulus (a discrimination judgment). In particular, in detection, evidence can only be available for the presence, not the absence, of a target object. In accordance with this observation, in Chapter 3 we observed slower reaction times, as well as lower and less reliable subjective confidence ratings for decisions about absence. Here we asked if these conceptual and behavioural differences between decisions about presence and absence are also reflected in brain activation patterns, specifically in parietal and prefrontal brain regions that are typically implicated in higher-order thought. In a within-subject, pre-registered and performance-matched fMRI design, we observed quadratic confidence effects in frontopolar cortex for detection but not discrimination. Furthermore, in the right temporoparietal junction, confidence effects were enhanced for judgments of target absence compared to judgments of target presence. We interpret these findings as reflecting qualitative differences between the neural basis of metacognitive evaluation of detection and discrimination, potentially in line with counterfactual or higher-order models of confidence formation in detection.

4.1 Introduction

When foraging for berries, one first needs to decide whether a certain bush bears fruit or not. Only if berries are detected, can one proceed to examine and classify them into a category - are these raspberries or blackberries? The first is a *detection* task: a decision about whether something is there or not, and the second is a *discrimination* task: a decision about which item is there. For these types of decisions, it is important not only to understand the decision process that leads to deciding present or absent, or raspberries or blackberries, but also our ability to reflect on and estimate the quality

of the decision, known as metacognition. For instance, two foragers working together may want to share their confidence in deciding which bush to tackle next (Bahrami et al., 2010; C. D. Frith, 2012).

There is an increasing understanding of the neural basis of confidence in simple decisions, with a network of prefrontal and parietal regions being identified as important for tracking metacognitive beliefs about the accuracy of both perceptual and value-based decisions (Domenech & Koechlin, 2015; for reviews, see Fleming & Dolan, 2012; Meyniel, Sigman, & Mainen, 2015). Accordingly, neuropsychological data in humans suggests that damage or impairment of prefrontal function can lead to metacognitive impairments such as noisy or inappropriate confidence judgments (for a review, see Rouault, Seow, Gillan, & Fleming, 2018). However, in a majority of these cases, the study of confidence has been restricted to discrimination, or deciding whether a stimulus is from category A or B. Despite their ubiquity and importance in decision-making, much less is known about how confidence is formed in detection settings, in which subjects are asked to make a judgment about whether a target stimulus is present or not.

Computational considerations and behavioural findings suggest that computing confidence in detection judgments may differ from computing confidence in the more commonly studied discrimination tasks. In particular, detection is unique in the landscape of perceptual tasks in that evidence can only be available to support the presence, not the absence, of a target object. This makes confidence ratings in judgments about absence a unique case, where confidence is decoupled from the amount of supporting perceptual evidence. Accordingly, behavioural evidence indicates that metacognitive sensitivity, or the alignment between subjective confidence and objective performance, for judgments about absence is typically impaired compared to metacognitive sensitivity for judgments about presence (Kanai, Walsh, & Tseng, 2010; Meuwese, Loon, Lamme, & Fahrenfort, 2014).

Under one family of models (*first-order models*), confidence in detection judgments is formed in the same way as confidence in discrimination judgments. For example, in evidence-accumulation models, confidence can be evaluated as the distance of the losing accumulator from the threshold at the time of decision (Merkle & Van Zandt, 2006). Similarly, in models of discrimination confidence based on *Signal Detection Theory* (SDT), decision confidence is assumed to be proportional to the strength of the available evidence supporting the decision, which is modeled as the distance of the perceptual sample from the decision criterion on a strength-of-evidence axis (Wickens, 2002, p. 85). While first-order models are traditionally symmetric, they can be adapted to account for the asymmetry between judgments about presence and absence. For example, *unequal-variance* and *multi-dimensional SDT models* account for the inherent difference between presence and absence by making the signal distribution wider than the noise distribution (Wickens, 2002, p. 48), or by assuming a high-dimensional stimulus space, in which the absence of a signal is represented as a distribution centered around the origin (King & Dehaene, 2014; Wickens, 2002, p. 118). Importantly, first-order models treat the process of metacogntive evaluation of detection and discrimination as qualitatively similar, with any differences between detection and discrimination emerging from differences in the underlying distributions

(uv-SDT), or the mapping between stimulus features and responses (two-dimensional SDT).

In contrast with first-order models of detection confidence, *higher-order models* treat confidence in judgments about target absence as emerging from a distinct, higher-order cognitive process. For instance, in one version of the higher-order approach, confidence in judgments about absence is assumed to be based on counterfactual estimation of the likelihood of a hypothetical stimulus to be detected, if presented. In other words, subjects may be more confident in the absence of a target object when they believe they would not have missed it, based on their global estimation of task difficulty, or on their current level of attention. A similar type of modeling has been successfully employed in studies of memory, to explain how participants form judgments that an item was not presented during the preceding learning phase, based on their counterfactual expectations about remembering an item (for example, Glanzer & Adams, 1990). When applied to the comparison of detection and discrimination, this approach predicts that qualitatively distinct cognitive and neural resources will be recruited when judging confidence in detection responses, due to the additional demand on counterfactual and self-monitoring processes, and that this recruitment will be most pronounced for confidence about absence. In particular, the counterfactual account predicts that responses in the frontopolar cortex, a region which has been shown to track counterfactual world states (Boorman, Behrens, Woolrich, & Rushworth, 2009), will show specificity for confidence judgements when inferring the absence of a target.

To test for such qualitative differences, here we set out to directly compare the neural basis of metacognitive evaluation of detection and discrimination responses within two similar low-level perceptual tasks, while controlling for differences in task performance. In a pre-registered design, we asked whether parametric relationships between subjective confidence ratings and the blood-oxygenation-level-dependent (BOLD) signal in a set of predefined prefrontal and parietal regions of interests (ROIs) would show systematic interaction with task (detection/discrimination) and, within detection, type of response (present/absent). To anticipate our results, we observed a quadratic effect of confidence on regional responses in frontopolar cortex for detection, but not for discrimination judgments. In further whole-brain exploratory analyses, we found stronger confidence-related effecs for judgments of absence compared to presence in right temporoparietal junction.

4.2 Methods and Materials

All design and analysis details were pre-registered before data acquisition and time-locked using pre-RNG randomization (Mazor, Mazor, & Mukamel, 2019). The time-locked protocol folder is available in the following GitHub repository. The entire set of preregistered analysis is available in the following OSF Project. Whole-brain imaging results are available in NeuroValut.

4.2.1 Participants

46 participants took part in the study (ages 18-36, mean = 24 ± 4; 29 females). 35 participants met our pre-specified inclusion criteria (ages 18-36, mean=24 ± 4; 20 females). After applying our run-wise exclusion criteria to the data of the remaining 35 participants, our dataset consisted of 5 usable experimental runs from 15 participants, 4 usable experimental runs from 14 participants, 3 usable experimental runs from 5 participants, and 2 usable experimental runs from one participant.

4.2.2 Design and procedure

After a temporally jittered rest period of 500-4000 milliseconds, each trial started with a fixation cross (500 milliseconds), followed by a presentation of a target for 33 milliseconds. In discrimination trials, the target was a circle of diameter 3° containing randomly generated white noise, merged with a sinusoidal grating (2 cycles per degree; oriented 45° or -45°). In half of the detection trials, targets did not contain a sinusoidal grating and consisted of random noise only. After stimulus offset, participants used their right-hand index and middle fingers to make a perceptual decision about the orientation of the grating (discrimination blocks), or about the presence or absence of a grating (detection blocks). The response mapping was counterbalanced between blocks, such that an index finger press was used to indicate a clockwise tilt on half of the trials, and an anticlockwise tilt on the other half. Similarly, in half of the detection trials the index finger was mapped to a 'yes' ('target present') response, and on the other half to a 'no' ('target absent') response.

Immediately after making a decision, participants rated their confidence on a 6-point scale by using two keys to increase and decrease their reported confidence level with their left-hand thumb. Confidence levels were indicated by the size and color of a circle presented at the center of the screen. The initial size and color of the circle was determined randomly at the beginning of the confidence rating phase, to decorrelate the number of button presses and the final confidence rating. The mapping between color and size to confidence was counterbalanced between participants: for half of the participants high confidence was mapped to small, red circles, and for the other half high confidence was mapped to large, blue circles. This counterbalancing was employed to isolate confidence-related activations from activations that originate from the perceptual properties of the confidence scale or from differences in the motor requirement to press the upper and lower buttons. The perceptual decision and the confidence rating phases were restricted to 1500 and 2500 milliseconds, respectively. No feedback was delivered to subjects about their performance.

Participants were acquainted with the task in a preceding behavioural session. During this session, task difficulty was adjusted independently for detection and for discrimination, targeting around 70% accuracy on both tasks. We achieved this by adaptively controlling the stimulus signal-to-noise ratio (SNR) once in every 10 trials: increasing the SNR when accuracy fell below 60%, and decreasing it when accuracy exceeded 80%. Performance on the detection and discrimination task was further calibrated to the scanner environment at the beginning of the scanning session, during

the acquisition of anatomical (MP-RAGE and fieldmap) images. After completing the calibration phase, participants underwent five ten-minute functional scanner runs, each comprising one detection and one discrimination block of 40 trials each, presented in random order.

To avoid stimulus-driven fluctuations in confidence, grating SNR was fixed within each experimental block. Nevertheless, following experimental blocks with markedly bad ($\leq 52.5\%$) or good ($\geq 85\%$) accuracy, grating SNR was adjusted for the next block of the same task (SNR level was divided or multiplied by a factor of 0.9 for bad and good performance, respectively). Finally, grating SNR was adjusted for both tasks following runs in which the difference in performance between the two tasks exceeded 16.25% (SNR level was multiplied by the square root of 0.9 for the easier task and divided by the square root of 0.9 for the more difficult task).

To incentivize participants to do their best at the task and rate their confidence accurately, we offered a bonus payment according to the following payment schedule: bonus = $£\frac{\overrightarrow{accuracy}\cdot\overrightarrow{confidence}}{200}$ Where $\overrightarrow{accuracy}$ is a vector of 1 and -1 for correct and incorrect responses, and $\overrightarrow{confidence}$ is a vector of integers in the range of 1 to 6, representing confidence reports for all trials. We explained the payment structure to participants in the preceding behavioural session. Specifically, we advised participants that to maximize their bonus they should do their best at the main task, rate the confidence higher when they believe they are correct, and rate their confidence lower when they believe they might be wrong.

4.2.3 Scanning parameters

Scanning took place at the Wellcome Centre for Human Neuroimaging, London, using a 3 Tesla Siemens Prisma MRI scanner with a 64-channel head coil. We acquired structural images using an MPRAGE sequence (1x1x1mm voxels, 176 slices, in plane FoV = 256x256 mm^2), followed by a double-echo FLASH (gradient echo) sequence with TE1=10ms and TE2=12.46ms (64 slices, slice thickness = 2mm, gap = 1mm, in plane FoV= 192×192 mm^2, resolution = 3×3 mm^2) that was later used for field inhomogeneity correction. Functional scans were acquired using a 2D EPI sequence, optimized for regions near the orbito-frontal cortex (3.0x3.0x3.0mm voxels, TR=3.36 seconds, TE = 30 ms, 48 slices tilted by -30 degrees with respect to the T>C axis, matrix size = 64x72, Z-shim=-1.4).

4.2.4 Analysis

The preregistered objectives of this study were to:

- Replicate findings of a generic (task-invariant) confidence signal in the activity of medial prefrontal cortex (Fleming & Dolan, 2012; Morales, Lau, & Fleming, 2018).

- Test for an interaction between the parametric effect of confidence level and task (detection/discrimination) in the BOLD response in prefrontal cortex ROIs.

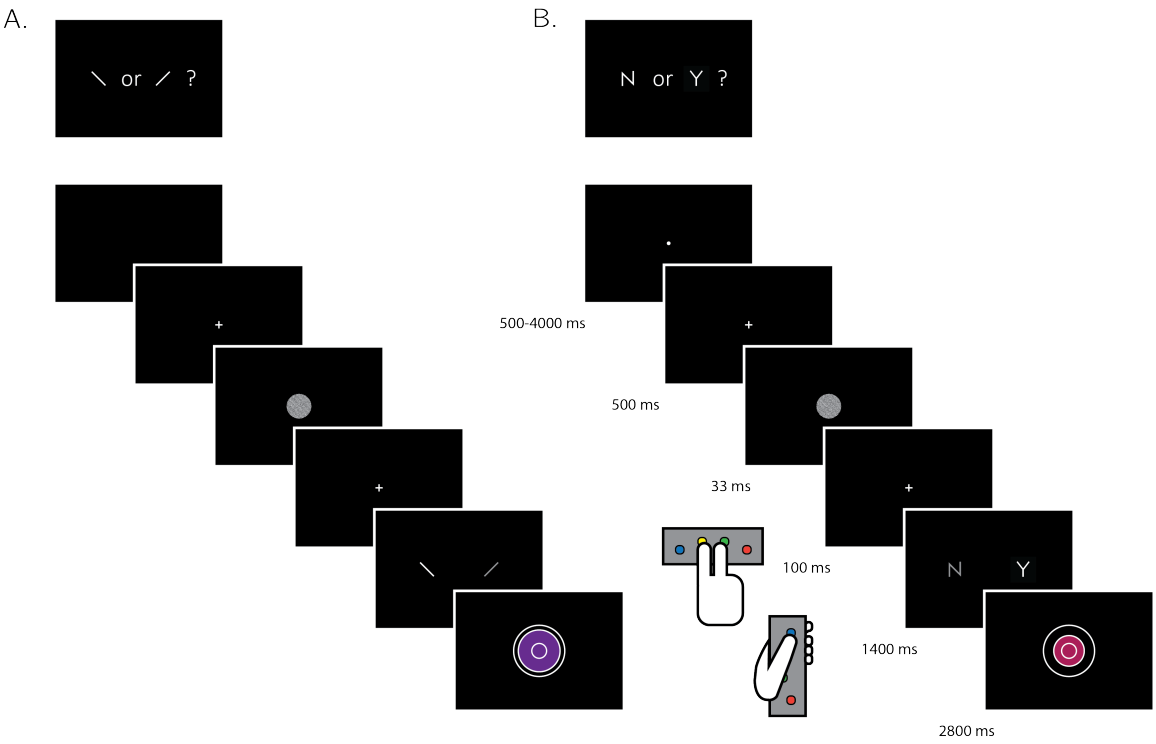


Figure 4.1: Experimental design for discrimination and for detection trials . Perceptual decisions were reported using the right index and middle fingers, and confidence ratings were reported using the left thumb. A) In discrimination blocks, participants indicated the orientation of a visual grating ('clockwise' or 'counterclockwise'). B) In detection blocks, participants indicated whether a grating was embedded in random noise, or not ('yes' or 'no'). Confidence ratings were made by varying the size and color of a circle, with 6 options ranging from small and red to big and blue. For half of the subjects, high confidence was mapped to a small, red circle. For the other half, high confidence was mapped to a big, blue circle. The initial size and color of the circle was determined randomly at the beginning of the confidence rating phase. Participants performed 10 interleaved 40-trial detection and discrimination blocks inside a 3T MRI scanner.

- Within detection trials, test for an interaction between the parametric effect of confidence level and response ('yes'/'no') in the BOLD response, specifically in the prefrontal cortex and in frontopolar regions that have previously been associated with counterfactual reasoning (Boorman, Behrens, Woolrich, & Rushworth, 2009; Donoso, Collins, & Koechlin, 2014).

- Test for relationships between fluctuations in metacognitive adequacy [a trial-by-trial measure of metacognitive sensitivity; Wokke, Cleeremans, & Ridderinkhof (2017)], and the BOLD signal separately for detection and for discrimination, and for 'yes' and 'no' responses within detection.

- Replicate previous findings of between-subject correlations between lateral prefrontal cortex (lPFC) function and metacognitive efficiency [meta-d'/d'; Fleming & Lau (2014)] in discrimination (Yokoyama et al., 2010).
- Identify between-subject functional correlates of metacognitive efficiency in detection. Specifically, ask if metacognitive efficiency in detection is predicted by activity in distinct networks compared to metacognitive efficiency in discrimination.

Exclusion criteria

Subjects were excluded from all analyses for any of the following pre-specified reasons: missing more than 20% of the trials, performing one of the tasks with accuracy below 60%, exceeding the 4 mm affine motion cutoff criterion in more than 2 experimental runs, and showing a consistent response bias (i.e. using the same response in more than 75% of the trials) in at least one task. Individual scan runs were excluded from all analyses if the participant exceeded the affine motion cutoff, if more than 20% of trials were missed, if mean accuracy was below 60% or if the response bias for one of the tasks exceeded 80%.

In addition, we applied a confidence-related exclusion criterion: participants were excluded if they used the same confidence level in more than 80% of all trials globally or for a particular response, and individual scan runs were excluded if the same confidence level was used in more than 95% of the trials, either globally or for particular response types. Our preregistration document specified that the confidence exclusion criterion will be used to exclude participants from confidence-related analyses only, but we subsequently revised this plan in order to use identical design matrices for all participants.

Response conditional type-II ROC curves

Response conditional ROC (Receiver Operating Characteristic) curves were extracted for the two discrimination and two detection responses. This was done by plotting the cumulative distribution of confidence levels in correct responses against the cumulative distribution of confidence levels in incorrect responses. As a measure of response-specific metacognitive sensitivity, we extracted the area under these curves (*AUROC2*). The expected AUROC2 for no metacognitive insight (i.e., the confidence distributions are identical for correct and incorrect responses) is 0.5. Perfect metacognitive insight (i.e., confidence in all correct responses is higher than confidence in all incorrect responses) will result in an AUROC2 of 1.

Imaging analysis

fMRI data preprocessing Data preprocessing followed the procedure described in Morales, Lau, & Fleming (2018):

> "Imaging analysis was performed using SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run

> were discarded to allow for T1 stabilization. Functional images were realigned and unwarped using local field maps (Andersson, Hutton, Ashburner, Turner, & Friston, 2001) and then slice-time corrected (Sladky et al., 2011). Each participant's structural image was segmented into gray matter, white matter, CSF, bone, soft tissue, and air/background images using a nonlinear deformation field to map it onto template tissue probability maps (Ashburner & Friston, 2005). This mapping was applied to both structural and functional images to create normalized images in Montreal Neurological Institute (MNI) space. Normalized images were spatially smoothed using a Gaussian kernel (6 mm FWHM). We set a within-run 4 mm affine motion cutoff criterion."

Preprocessing and construction of first- and second-level models used standardized pipelines and scripts available at the MetaLab GitHub page

Regions of Interest In addition to an exploratory whole-brain analysis (corrected for multiple comparisons at the cluster level), our analysis focused on the following a priori regions of interest, largely following the ROIs used by Fleming, Van Der Putten, & Daw (2018):

- *Frontopolar cortex* (FPC, defined anatomically). We used a connectivity-based parcellation (Neubert, Mars, Thomas, Sallet, & Rushworth, 2014) to define a general FPC region of interest as the total area spanned by areas FPl, FPm and BA46. The right hemisphere mask was mirrored to create a bilateral mask.

- *Ventromedial prefrontal cortex* (vmPFC). The vmPFC ROI was defined as a 8-mm sphere around MNI coordinates [0,46,-7], obtained from a meta-analysis of subjective-value related activations (Bartra, McGuire, & Kable, 2013) and aligned to the cortical midline.

- *Bilateral ventral striatum.* The ventral striatum ROIs was specified anatomically from the Oxford-Imanova Striatal Strctural Atlas included with FSL.

- *Posterior medial frontal cortex* (pMFC). The pMFC ROI was defined as a 8-mm sphere around MNI coordinates [0, 17, 46], obtained from a functional MRI study on decision confidence) and aligned to the cortical midline (Fleming, Huijgen, & Dolan, 2012).

- *Precuneus.* The precuneus ROI was defined as a 8-mm sphere around MNI coordinates [0,-57,18], based on Voxel Based Morphometry studies of metacognitive efficiency (Fleming, Weil, Nagy, Dolan, & Rees, 2010; McCurdy et al., 2013) and aligned to the cortical midline.

For the general FPC ROI, small-volume correction was applied to individual voxels within the ROI for all univariate contrasts. For the multivariate analysis, we used a searchlight approach to scan for spatial patterns within the ROI, followed by a correction for multiple comparisons. For all other ROIs, a GLM was fitted to

the mean time course of voxels within the region, and multivariate analysis was performed on all voxels within the ROI. While our pre-registered analysis defined the frontopolar cortex as a single region, we subsequently decided to separately analyze its 3 separate anatomical subregions identified by Neubert, Mars, Thomas, Sallet, & Rushworth (2014) (FPl, FPm and BA46). The decision to separate the FPC ROI to its subcomponents was made *after* data collection. These anatomical subregions should not be taken as prior ROIs.

Univariate analysis Univariate analysis was based on a design matrix in which different trial types are modeled by different regressors (main design matrix, below). Additionally, to examine the global effect of confidence across trial types, a simpler design matrix was fitted to the data as a first step (global confidence design matrix, below). Experimental runs for each subject were temporally concatenated before estimating the GLM coefficients. This was done in order to maximize sensitivity to response- and task-specific modulations of confidence, given the limited and varying number of trials within each experimental run.

Main Design Matrix (DM-1) The main design matrix for the univariate GLM analysis consisted of 16 regressors of interest. There was a regressor for each of the eight combinations of task x condition x response: For example, a regressor for detection trials where a signal was present and the subject reported seeing a signal with a 'yes' response. The relevant trials were modeled by a boxcar regressor with nonzero entries at the 4300 millisecond interval starting at the offset of the stimulus and ending immediately after the confidence rating phase, convolved with the canonical hemodynamic response function (HRF). Each of these primary regressors was accompanied by a linear parametric modulation of the confidence reported for each trial. Together, the design matrix included 16 regressors of interest (see table 4.1)

Table 4.1: List of regressors in the main design matrix (DM-1).

		Task	Stimulus	Response
1	CW_CW	Discrimination	Clockwise	Clockwise
2	CW_ACW_conf			
3	CW_ACW	Discrimination	Clockwise	anticlockwise
4	CW_ACW_conf			
5	CW_CW	Discrimination	anticlockwise	Clockwise
6	CW_CW_conf			
7	ACW_ACW	Discrimination	anticlockwise	anticlockwise
8	ACW_ACW_conf			
9	Y_Y	Detection	Signal	Yes
10	Y_Y_conf			
11	Y_N	Detection	Signal	No
12	Y_N_conf			

		Task	Stimulus	Response
13	N_Y	Detection	Noise	Yes
14	N_Y_conf			
15	N_N	Detection	Noise	No
16	N_N_conf			

Trials in which the participant did not respond within the 1500 millisecond time frame were modeled by a separate regressor. The design matrix also include a run-wise constant term regressor, an instruction-screen regressor for the beginning of each block, motion regressors (the 6 motion parameters and their first derivatives as extracted by SPM in the head motion correction preprocessing phase) and regressors for physiological measures. Button presses were modeled as stick functions, convolved with the canonical HRF, in three regressors: two regressors for the right and left right-hand buttons, and one regressor for both up and down left-hand presses. We decided to have one regressor for both types of left-hand presses due to the strong positive correlation of the final confidence rating with the number of 'increase confidence' button presses, and the strong negative correlation with the number of 'decrease confidence' button presses.

Global Confidence Design Matrix (GC-DM) The global confidence design matrix consisted of 4 regressors of interest. The first two primary regressors were 'correct trials' (trials in which the participant was correct, across tasks and responses) and 'incorrect trials' (trials in which the participant was incorrect, across tasks and responses). Single events were modeled by a boxcar regressor with nonzero entries at the interval starting at the offset of the stimulus and ending immediately after the confidence rating phase, convolved with the canonical hemodynamic response function (HRF). The duration of this interval was 4300 milliseconds, and not 4000 milliseconds as mistakenly indicated in the preregistration document. Additionally, the design matrix included a confidence parametric modulator for each of the first two regressors. The construction of the regressors and the additional nuisance regressors was handled similarly to the main design.

Quadratic-Confidence Design Matrix (post-hoc analysis; QC-DM) The quadratic-confidence design matrix for the univariate GLM analysis consisted of 12 regressors of interest. There was a regressor for each of the four responses: 'yes,' 'no,' 'clockwise' and 'anticlockwise.' Similar to the main design matrix, the relevant trials were modeled by a boxcar regressor with nonzero entries at the 4300 millisecond interval starting at the offset of the stimulus and ending immediately after the confidence rating phase, convolved with the canonical hemodynamic response function (HRF). Each of these primary regressors was accompanied by two parametric modulators, representing the linear and quadratic effects of confidence. Together, the design matrix included 12 regressors (4 responses + 4 linear confidence regressors + 4 quadratic confidence regressors). The QC-DM included the same set of nuisance regressors as the main design matrix.

Categorical-Confidence Design Matrices (post-hoc analysis; CC-DM) In order to better understand the nature of the linear interaction between confidence in 'yes' and 'no' responses, we specified a pair of design matrices - one for each task - in which confidence level was modeled as a categorical variable. Instead of the 8 primary regressors in the main design matrix, this design matrix consisted of only one regressor of interest for all trials, modeled by a boxcar with nonzero entries at the 4300 millisecond interval starting at the offset of the stimulus and ending immediately after the confidence rating phase, convolved with the canonical hemodynamic response function (HRF). This regressor was in turn modulated by a series of 12 dummy (0/1) parametric modulators - one for every response ('yes' and 'no' for detection and 'clockwise' and 'anticlockwise' for discrimination) and confidence rating (1-6 for both tasks). Using two design matrices instead of one allowed us to set discrimination trials to be the baseline category for detection, and detection trials as the baseline for discrimination. These design matrices included the same set of nuisance regressors as the main design matrix.

For each participant, we used the beta-estimates from the categorical-confidence design matrices as the input to four response-specific multiple linear regression models, with linear confidence and quadratic confidence as predictors, in addition to an intercept term. The subject-specific coefficients were then subjected to ordinary least squares group-level inference, to compare linear and quadratic effects of confidence between responses. The rational for choosing this two-step approach was its ambivalence to differences in the confidence distributions for the four responses, that may bias the estimation of the quadratic and linear terms.

Multivariate analysis Multi-voxel pattern analysis (Norman, Polyn, Detre, & Haxby, 2006) was used to test for consistent spatial patterns in the fMRI data. We used The Decoding Toolbox (Hebart, Görgen, & Haynes, 2015) and followed the procedures described by Morales, Lau, & Fleming (2018). In order to identify brain regions that are implicated in inference about presence and absence, we trained and tested a linear classifier on detection decisions. We classified hits and correct rejections, instead of hits and misses as originally planned, due to an insufficient number of detection misses in some experimental blocks. We then compared the resulting classification accuracy with the cross-classification accuracy of training on detection responses and testing on discrimination confidence and vice versa. The purpose of this comparison was to isolate neural correlates of inference about stimulus absence or presence that should be specific to detection from more general neural correlates of stimulus visibility, that are also expected to affect confidence in discrimination judgements.

The other prespecified multivariate tests were designed to find universal and response-specific spatially multivariate representations of confidence. After conducting this analysis we came to realize that our experimental design was not appropriate for estimating the degree to which the representation of confidence is "response-general." In our experimental design, confidence is confounded with visual feedback during the confidence-rating phase, such that "response-general" representations of confidence could appear if the spatial pattern of activation was sensitive to the visual feedback in

the confidence rating. For completeness, we include the results of this analysis in the appendix (E.7), but do not interpret them further.

Statistical inference

T-test and anova Bayes factors use a Jeffrey-Zellner-Siow Prior for the null distribution, with a unit prior scale (Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Whole-brain fMRI significance was corrected for family-wise error rate at the cluster level ($p<0.05$), with a cluster defining threshold of $p<0.001$.

4.3 Results

35 participants performed two perceptual decision-making tasks while being scanned in a 3T MRI scanner: an orientation discrimination task (*"was the grating tilted clockwise or anticlockwise?"*), and a detection task (*"was any grating presented at all?"*). At the end of each trial, participants rated their confidence in the accuracy of their decision on a 6-point scale. We adjusted the difficulty of the two tasks in a preceding behavioural session to achieve equal performance of around 70% accuracy. At scanning, 10 discrimination and detection blocks were presented in 5 scanner runs.

4.3.1 Behavioural results

Task performance was similar for detection (75% accuracy, d'=1.48) and discrimination blocks (76% accuracy, d'=1.50). Repeated measures t-tests failed to detect a difference between tasks both in mean accuracy ($t(34) = -0.90, p = 0.37$, $BF_{01} = 5.15$), and d' ($t(34) = -0.30, p = 0.76$, $BF_{01} = 7.29$), indicating that performance was well matched. Responses were also balanced for the two tasks. The probability of responding 'yes' (target present) in the detection task was 0.49 ± 0.11, and not significantly different from 0.5 ($t(34) = -0.39, p = 0.70$, $BF_{01} = 7.07$). The probability of responding 'clockwise' in the discrimination task was 0.50 ± 0.08, and not significantly different from 0.5 ($t(34) = 0.22, p = 0.87$, $BF_{01} = 7.43$).

The distribution of confidence ratings was generally similar between the two tasks and four responses. For all four responses, participants were most likely to report the highest confidence rating compared to any other option. Within detection, a significant difference in mean confidence was observed between 'yes' (target present) and 'no' (target absent) responses, such that participants were more confident in their 'yes' responses ($t(34) = -4.85, p < 0.0001$; see Figure 4.2). This difference in mean confidence was mostly driven by the higher proportion of maximum confidence ratings in 'yes' responses compared to 'no' responses (46% of all 'yes' responses compared to 26% of all 'no' responses, $t(34) = 5.63, p < 0.00001$), but persisted even when ignoring the highest ratings ($t(34) = 2.39, p < 0.05$). Metacognitive sensitivity, quantified as the area under the type-II ROC curve, was significantly higher for 'yes' compared to 'no' responses ($t(34) = 7.83, p < 10 - 8$; see Figure 4.2, as expected (Meuwese,

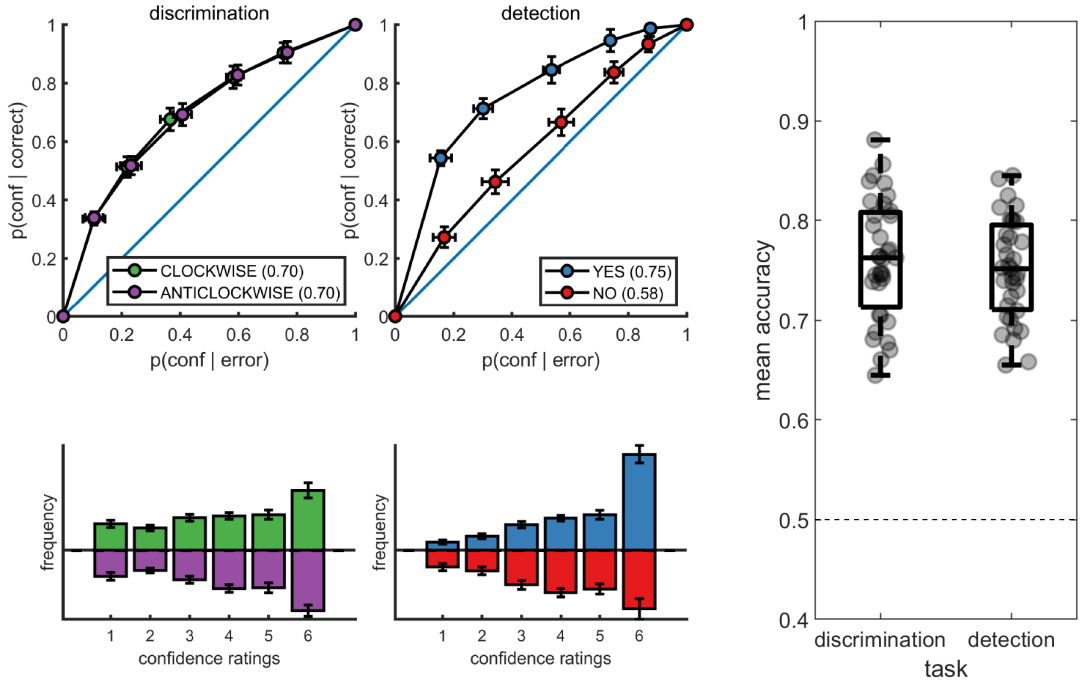


Figure 4.2: Upper panels: response conditional type-2 ROC curves. In parentheses: the mean area under the curve. Lower panels: distribution of confidence ratings for the two tasks and four responses. Right panel: Mean accuracy for both tasks. Error bars represent the standard error of the mean.

Loon, Lamme, & Fahrenfort, 2014). In other words, confidence ratings about the presence of a target stimulus were more diagnostic of accuracy than ratings about target absence, even though both sets of ratings tended to cover the full range of the scale, from low to high confidence. Taking metacognitive sensitivity following discrimination responses as a baseline, we found that this effect was driven by a decrease in metacognitive sensitivity for 'no' responses ($t(34) = -4.89, p < 0.0001$), whereas a quantitative increase in metacognitive sensitivity for 'yes' responses compared to discrimination was not significant ($t(34) = 1.84, p = 0.07$). No difference was observed in metacognitive sensitivity between the two discrimination responses ('clockwise' and 'anticlockwise'; $t(34) = 0.06, p = 0.95$, $BF_{01} = 7.6$). Taken together, these results are consistent with the previously reported selective asymmetry in the fidelity of metacognitive evaluation following judgments about target absence (Kanai, Walsh, & Tseng, 2010; Meuwese, Loon, Lamme, & Fahrenfort, 2014).

Response times were faster on average for correct responses (849 ± 79 milliseconds) compared to incorrect responses (938 ± 95 milliseconds; $t(34) = 10.59, p < 10^{-11}$ for a paired t-test on the log-transformed response times). Within the detection task, 'yes' responses were significantly faster than 'no' responses (850 ± 90 milliseconds and

896 ± 103 milliseconds, respectively; $t(34) = 3.16, p < 0.005$ for a paired t-test on the log-transformed response times).

4.3.2 Imaging results

Parametric effect of confidence

We next turned to our fMRI data to ask whether confidence-related responses were similar or distinct across tasks (detection / discrimination) and response (target present: 'yes' / target absent: 'no'). We first established the presence of linear confidence-related effects in our a priori ROIs, both across tasks and response types and across correct and incorrect responses, in line with previous findings of "generic" or task-invariant confidence signals in these regions (Morales, Lau, & Fleming, 2018). Specifically, high confidence ratings were associated with increased activation in the ventromedial prefrontal cortex (vmPFC), the ventral striatum, and the precuneus. Conversely, activations in the posterior medial frontal cortex (pMFC) were negatively correlated with confidence (see figure 4.3). For the confidence effect pattern obtained from the Global-Confidence Design Matrix (GC-DM), see supplementary figure E.3.

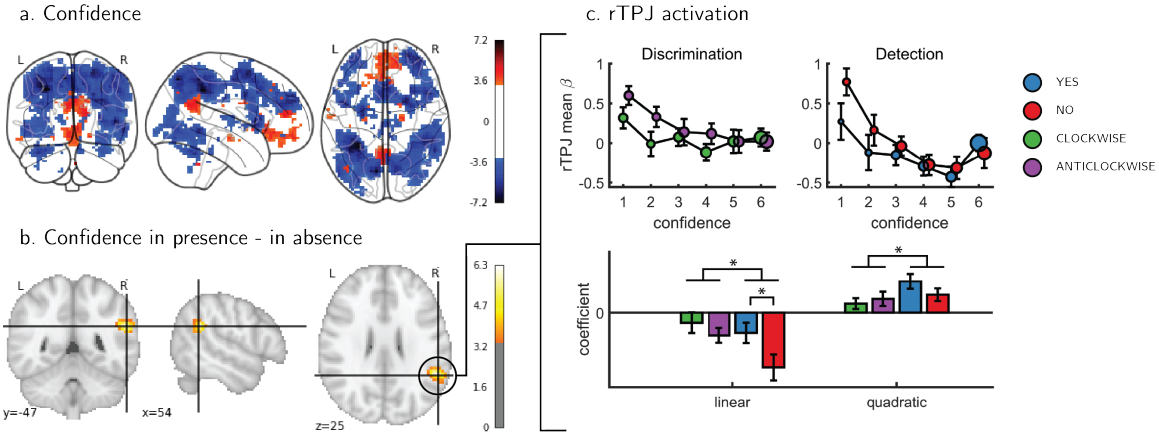


Figure 4.3: Univariate parametric effect of confidence. a) Glass brain visualization of global effect of confidence, thresholded at the single voxel level for visualization (p<0.001, uncorrected). Negative confidence effects appears in blue, and positive effects in red. b) Whole brain contrast between confidence in 'yes' (target-present) and 'no' (target-absent) detection responses, corrected for family-wise error rate at the cluster level (p<0.05) with a cluster defining threshold of p<0.001, uncorrected. c. upper panel: BOLD signal in the rTPJ cluster from panel b as a function of response and confidence. lower panel: mean coefficients of response- and subject-specific multiple linear regression models, predicting rTPJ activation as a linear and quadratic function of confidence. * - p<0.05; uncorrected for multiple comparisons across the four tests.

Interaction of linear confidence effects with task and response

We next asked whether the linear parametric relationship between confidence and BOLD activity differed as a function of task (discrimination vs. detection) and response type ('yes' vs. 'no' in detection). In the pMFC, vmPFC, ventral striatum and precuneus ROIs, the parametric effect of confidence failed to show a significant difference between the two tasks (all p-values > 0.3), between the two discrimination responses (all p-values > 0.24), or between the two detection responses (all p-values > 0.09). Similarly, no cluster within the pre-specified frontopolar ROI showed a differential effect of confidence as a function of task or response. We show below that this absence of a linear interaction should not be taken as evidence of absence of differences between detection and discrimination, due to the presence of nonlinear interaction effects. In the next section we first explain the analysis steps we took to uncover nonlinear effect of confidence.

Interaction of nonlinear confidence effects with task and response

An exploratory whole brain analysis ($p<0.05$, corrected for multiple comparisons at the cluster-level) revealed no differential confidence effect as a function of task anywhere in the brain. However, within detection, whole-brain analysis revealed that the linear effect of confidence was significantly more negative for 'no' compared to 'yes' responses in the right temporo-parietal junction (rTPJ: 101 voxels, peak voxel: [54,-46, 26], $z = 5.10$). To further characterize the nature of the interaction between confidence and response in the rTPJ, we fitted a new design matrix for each task (CC-DM) where confidence was represented as a categorical variable with 6 levels instead of one parametric modulator. In contrast to our original design matrix (DM-1) that assumed a linear effect of confidence, this analysis is agnostic as to the functional form of the confidence effect. We then plotted the mean activation level for each combination of response and confidence level in the rTPJ cluster (see Figure 4.3, panel c).

The categorical-confidence design matrix revealed a positive quadratic effect of confidence on activation levels in the rTPJ, with stronger activation levels for the two extremities of the confidence scale. We confirmed the presence of a significant quadratic effect of confidence in this region by fitting a second-order polynomial to the response-specific confidence curve of each participant (see Methods). This analysis revealed a main quadratic effect of confidence in this region ($t(34) = 5.21, p < 0.00001$), an effect which was stronger in detection compared to discrimination ($t(34) = 2.06, p < 0.05$). Importantly, the linear interaction of confidence with detection responses remained significant for this quadratic model, establishing that this response-specific effect is not explained by an overall quadratic pattern ($t(33) = 2.09, p < 0.05$; see Figure 4.3). More generally, these analyses make clear that linear effects of parametric modulators and their interactions are not exhaustive in their characterization of the confidence-related BOLD response – in this region and potentially in our other ROIs too.

To formally test for such nonlinear differences in the activation profile of other ROIs, we extracted the coefficients from the categorical model for each ROI, and

fitted a second-order polynomial separately for the ensuing confidence-related response. Within our a priori ROIs, no quadratic effect of confidence was observed in the pMFC, the precuneus, the ventral striatum, or the vmPFC (see supplementary figure E.4). In contrast, in all three anatomical subregions of the frontopolar cortex, we found a positive quadratic effect of confidence, with stronger activations for the two extremities of the confidence scale. Strikingly, in both the FPl and the FPm, this positive quadratic effect of confidence was entirely driven by the detection task (FPm: $t(34) = 3.04, p < 0.005$; FPl: $t(34) = 3.90, p < 0.001$; see Figure 4.4). Confidence ratings for the discrimination task however showed a quadratic effect that was not statistically different from zero (FPm: $t(34) = -0.54, p = 0.59$, $BF_{01} = 6.61$; FPl: $t(34) = 1.42, p = 0.16$, $BF_{01} = 2.92$). In the FPm, the linear effect of confidence was more negative for detection than for discrimination ($t(34) = -2.11, p < 0.05$), and within detection, more negative for confidence in judgments about absence ('no' responses; $t(34) = 2.10, p < 0.05$).

Finally, to test for similar quadratic effects of confidence at the whole-brain level, we constructed a new design matrix (in a departure to our pre-registered analysis plan) in which confidence was modeled by a parametric modulator with a polynomial expansion of 2 (QC-DM). Three clusters in the right hemisphere showed a significantly stronger quadratic effect of confidence in detection compared to discrimination (Figure 4.5). These were located in the right superior temporal sulcus (72 voxels, peak voxel: [60,-43,2], Z=3.99), right pre-SMA (130 voxels, peak voxel: [0,35,47], Z=4.07), and right frontopolar cortex, overlapping with our FPl and FPm frontopolar anatomical subregions (51 voxels, peak voxel: [9,65,-10], Z=4.00).

To visualize activity patterns in these regions, we extracted the mean coefficients from the categorical model for these three clusters, and fitted a second-order polynomial separately to each response estimate (see Figure 4.5). In addition to the effect of task on the quadratic effect of confidence in all three clusters, the linear effect of confidence in the right frontopolar cluster was significantly more negative for detection, compared to discrimination ($t(34) = -3.13, p < 0.005$). For both tasks, inter-subject variability in metacognitive efficiency [measured as meta-d'/d'; Maniscalco & Lau (2010)] was not reliably correlated with linear or quadratic parametric effect of confidence in any of the three regions (see Supplementary Figure E.5 in the supplementary materials).

4.3.3 Computational models

We next considered alternative computational-level explanations for the detection-specific quadratic activation profile. Specifically, we evaluated how latent model variables or belief states change non-linearly as a function of confidence in three candidate model architectures (see 4.6): a static 'Signal Detection' model, a 'Dynamic Criterion' model where policy changes as a function of previous perceptual samples, and an 'Attention Monitoring' model in which beliefs about fluctuations in attention inform decisions and confidence judgments. A detailed formal description of the three models is available in the appendix (sections E.8, E.9 and E.10), and Matlab implementations are available in the following page. First, we consider the static Signal Detection Theory (SDT) model. In SDT models of confidence formation, the

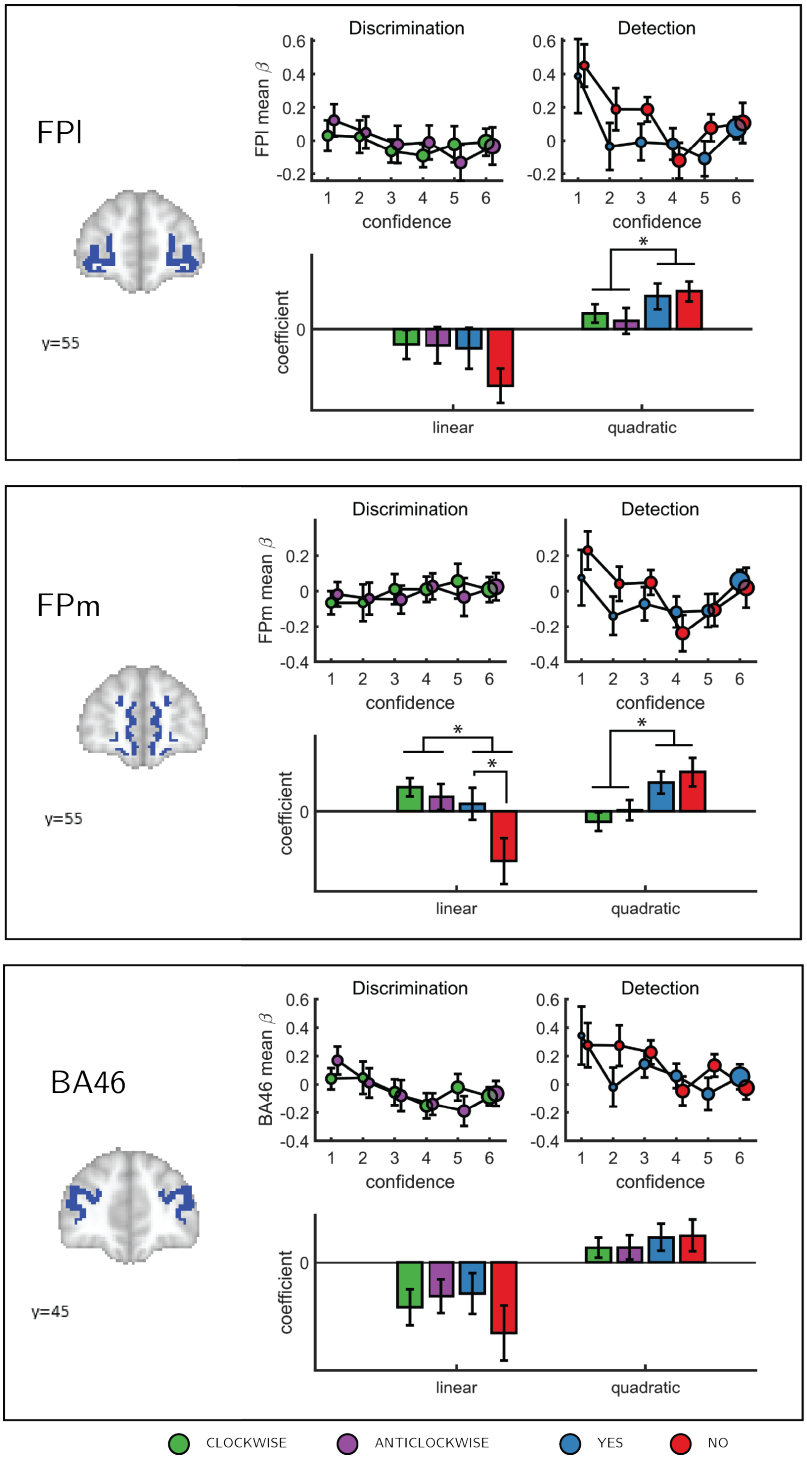


Figure 4.4: Confidence effect as a function of response in the frontopolar cortex separated into its three anatomical subcomponents: FPm, FPl, and BA 46. Same conventions as in Figure 4.4 * - p<0.05; uncorrected for multiple comparisons.

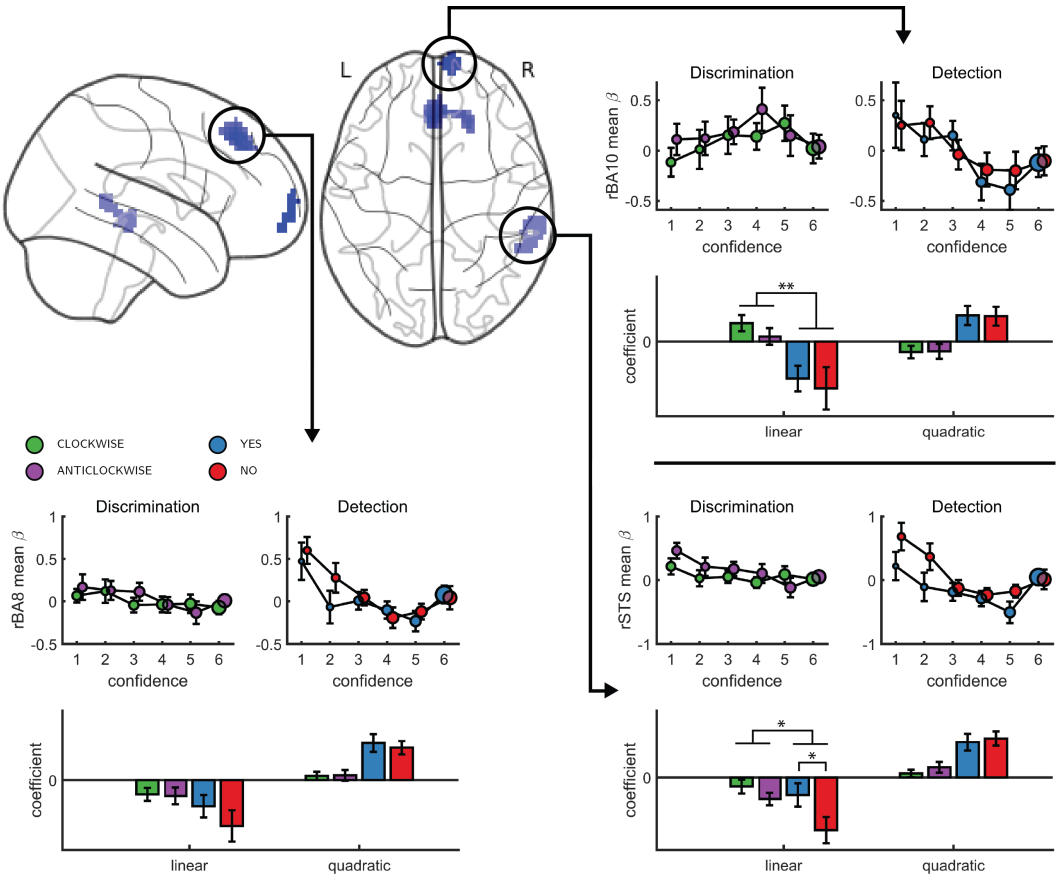


Figure 4.5: Left, top panel: a glass-brain representation of a contrast between the quadratic effects of confidence in detection and in discrimination, whole-brain corrected for family-wise error rate at the cluster-level (p<.05) with a cluster-defining threshold of p<.001, uncorrected). Remaining panels: mean betas from the categorical model for each of the four responses and six confidence ratings, for the three indicated clusters. The second-order polynomial coefficients for these estimates are presented below each plot. Significance is only indicated for the linear effects, which are orthogonal to the quadratic contrast used to select the clusters. * - p<0.05; ** - p<0.01

log likelihood-ratio between the two competing hypotheses ($LLR = log\frac{p(x|S1)}{p(x|S2)}$) is a useful measure for determining the certainty with which one should commit to a choice. The mapping between the perceptual sample x and the LLR is linear for equal-variance SDT, which is often used to model discrimination, but quadratic for unequal-variance SDT, which is often used to model detection. It then follows that if confidence is proportional to the distance of the sample x from the decision criterion, neuronal populations that represent the relative likelihood of a choice being correct (be it LLR or an analogue quantity) will show a quadratic tuning function of confidence in detection and a linear tuning function in discrimination, similar to that observed in FPC, pre-SMA and STS. However, LLR is also expected to scale more strongly

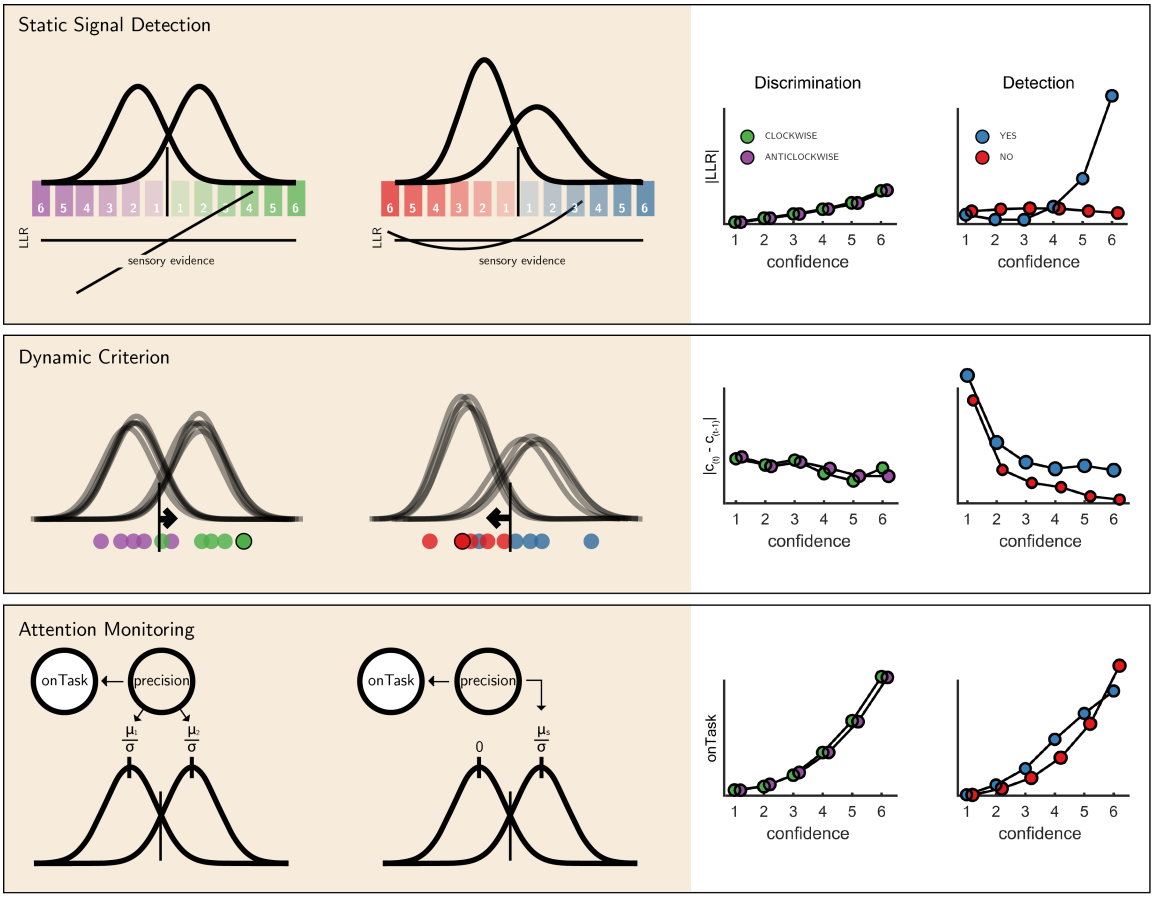


Figure 4.6: The three models (left) and their prediction for confidence effects (right). Top panel: In Signal Detection Theory, perceptual decisions and confidence ratings are generated by comparing the sensory evidence to a fixed set of criteria. In detection the 'signal' distribution is assumed to have higher variance. Plotting the absolute value of the log likelihood ratio as a function of decision and confidence results in a linear curve for discrimination, and a pronounced quadratic effect for 'yes' responses in detection, an effect that is specific to unequal-variance SDT. Middle panel: In a Dynamic Criterion model beliefs about the mean and variance of the perceptual distributions are updated as a function of incoming samples (plotted as circles) and the decision criterion is shifted accordingly. Plotting the absolute change in criterion placement as a function of decision and confidence results in a quadratic effect of confidence for detection responses only. Bottom: In the Attention Monitoring model, beliefs about overall attentiveness ('onTask' node) probabilistically reflect sensory precision. Plotting beliefs about overall attentiveness as a function of decision and confidence results in an overall quadratic effect of confidence, and an interaction between 'yes' and 'no' responses in detection.

with confidence in yes responses (see simulation results in Figure 4.6, upper panel), which was not observed in these brain regions. This model also predicts a stronger quadratic effect of confidence in participants for which the variance ratio between the signal and noise distributions is particularly high. However, the variance ratio was not significantly correlated with the quadratic effect of confidence in any of these regions, as would be expected if they were representing LLR or a similar quantity (see Appendix section E.5).

For the next two models, confidence was assumed to be directly proportional to the LLR, with the measured signal representing internal beliefs about hidden model parameters. In the 'Dynamic Criterion' model, we considered whether a quadratic effect of confidence in detection may reflect the active tuning of decision policy in the absence of explicit feedback (Guggenmos, Wilbertz, Hebart, & Sterzer, 2016; Ko & Lau, 2012). In the model, beliefs about the underlying distributions are updated on a trial-to-trial basis, and in turn affect the placement of decision criterion. The Dynamic Criterion model predicts that the magnitude of shift in decision criterion will display a positive quadratic relation to confidence (LLR) in detection but not discrimination (see simulation results in Figure 4.6, middle panel). This is because the problem is asymmetric in detection, and decision policy should depend on beliefs about both sensory precision (or the relative variance of the noise and signal distribution) and expected signal strength (mean of the signal distribution), which is not the case for a symmetric discrimination problem.

Notably, the pattern of criterion shifts in the Dynamic Criterion model resembled the task-specific effect of confidence in the FPC, STS and pre-SMA. As a post-hoc test of a role for these regions in criterion adjustment, we examined sequential pairs of trials of the same stimulus category (for example, a signal present trial that was followed by a signal present trial), and contrasted 'repeat' trials with 'switch' trials (for example, ['yes,' 'yes'] vs. ['yes,' 'no']). The Dynamic Criterion model predicts stronger activation in switch compared to stay trials in both detection and discrimination. The FPl showed a weak effect in this direction ($t = 2.03, p = 0.05, d = 0.34$), whereas FPm, pre-SMA, right BA10 and STS did not (all p-values>0.15).

Finally, we considered a higher-order 'Attention Monitoring' model in which beliefs about one's current attentional state (precision or inverse variance in SDT) are taken into account when making perceptual decisions and confidence ratings on detection trials. This model formalizes the notion that after not detecting a target the participant may ask 'Given my current attentional state, would I have missed the target?' The Attention Monitoring model thus makes different predictions for confidence in detection 'no' (target absent) responses, where the participant is assumed to reflect on the detection-likelihood of hypothetical targets, compared to 'yes' (target absent) responses, similar to the activation profile observed in the rTPJ. However, this model also predicts a pronounced quadratic confidence profile for all four responses, which we do not see in our data.

4.4 Discussion

Previous studies of the neural basis of human perceptual decision-making have tended to focus on discrimination judgments, such as sorting stimuli into category A or B. The general computational architecture supporting discrimination judgments can be naturally extended to support detection (for instance, within signal detection theory). However, computational considerations and behavioral findings suggest that forming confidence in detection judgments may rest on qualitatively distinct cognitive and neural processes in comparison to generating confidence in discrimination judgments.

To test for such differences, here we acquired functional MRI data from 35 participants who reported their subjective confidence in judgments about stimulus type (discrimination), and target presence or absence (detection). These judgments were given on separate trials that were well-matched for stimulus characteristics, response requirements and task difficulty. Across both tasks, we found the expected linear effects of confidence in our pre-specified regions of interest in the prefrontal and parietal cortex. Specifically, in the precuneus, vmPFC, pMFC and ventral striatum, the effect of confidence was invariant to task and response. In contrast, having adjusted our planned design matrix to be sensitive to non-monotonic effects of confidence, we observed a quadratic effect of confidence in detection judgments in the frontopolar cortex (medial and lateral surfaces of BA10), that was absent for discrimination judgments. Similar quadratic activation profiles were observed for both 'yes' and 'no' responses. Whole-brain analysis revealed a similar effect of task on the quadratic effect of confidence in the right STS and the pre-SMA. Since task performance was matched across the two tasks and since we did not observe overall differences in activation between detection and discrimination, these differences in confidence profiles are unlikely to originate from experimental confounds such as task difficulty, but instead indicate a unique neurocognitive contribution to metacognition of detection judgments. In what follows we will unpack what this contribution might be.

The three regions that showed an interaction of the quadratic expansion of confidence with task in our whole-brain analysis (right frontopolar cortex, right STS, and pre-SMA), as well as two anatomical subcomponents of our frontopolar ROI (FPl and FPm), all shared a very similar activation profile. In detection, the quadratic effect of confidence was positive, but was almost entirely absent for the discrimination task. Follow-up analysis confirmed that this difference was not driven by motor aspects of the confidence rating procedure, such as the number of increase or decrease confidence steps taken to reach the desired confidence level, which was similar for the two tasks (see Appendix E.1). Ours is not the first report of a quadratic relation between activation in prefrontal cortical structures and different subjective ratings. For example, in a study by Christensen, Ramsøy, Lund, Madsen, & Rowe (2006), participants were presented with masked stimuli and gave subjective visibility ratings on a three-point scale. The right frontopolar cortex showed decreased activation for 'clear perception' and 'no perception' categories relative to a middle 'vague perception' category. Similarly, De Martino, Bobadilla-Suarez, Nouguchi, Sharot, & Love (2017) reported a quadratic effect of product desirability in the pMFC. However, for both of the above cases, a quadratic effect can reflect a monotonic relationship with an implicit

representation of subjective confidence (Lebreton, Abitbol, Daunizeau, & Pessiglione, 2015). For example, participants may be more confident in the 'clear perception' and 'no perception' responses compared to the 'vague perception' option, or more confident about liking or not liking a product, compared to when using the middle parts of the liking scale. This explanation cannot account for the observed quadratic trend in our case, where in addition to strong activation levels for the highest confidence ratings in target presence and absence, we also find strong activation levels for the lowest levels of confidence.

We are unable to determine whether this effect originates from one homogeneous population of neurons that shows a quadratic effect of detection confidence, or from two overlapping populations that show nonlinear positive and negative effects of detection confidence – summing to an overall quadratic effect at the voxel level [similar to positive and negative confidence-selective neurons in the human posterior parietal cortex; Rutishauser, Aflalo, Rosario, Pouratian, & Andersen (2018)]. Addressing this question would require higher spatial resolution, for example using single-cell recordings in patients. Furthermore, because confidence judgments were always preceded by perceptual decisions in our design, we cannot determine whether the observed effects reflect an implicit representation of uncertainty, computed in parallel with the perceptual decision itself, or a higher-order representation that emerges at the explicit confidence rating phase. Future studies which use model-based estimates of covert decision confidence (Bang & Fleming, 2018) or EEG-informed fMRI to resolve early and late processing stages (Gherman & Philiastides, 2018) may answer this question.

We considered three alternative computational models that were able to account for asymmetries between detection and discrimination activation profiles. An unequal variance signal detection theory model provided a simple account of the asymmetry between detection and discrimination, but could not account for the similar quadratic profiles observed for 'yes' and 'no' responses. A more direct test of the proposal that a detection-specific quadratic effect of confidence originates from the unequal-variance properties of stimulus distributions in detection would be to test for similar effects in a discrimination task in which one category of stimuli is of higher variance (e.g., Denison, Adler, Carrasco, & Ma, 2018). In contrast, the Dynamic Criterion model provided good qualitative accounts for distinct regional activation profiles, and the Attention Monitoring account predicted an interaction between confidence in judgments about presence and absence. However, the Attention Monitoring model also predicted a quadratic effect in discrimination, which we did not see.

Notably, both of these models share the need to learn (in the Dynamic Criterion model) or estimate (in the Attention Monitoring model) the current level of precision (inverse variance) in detection. Such online precision estimation evinces a profound asymmetry between detection and discrimination tasks: in discrimination tasks, one simply has to evaluate the relative evidence for different causes of sensory samples, under some prior belief about sensory precision; namely, the precision of the likelihood that any particular cause (e.g., clockwise or anticlockwise orientation) would generate sensory samples. In contrast, detection presents a difficult (ill-posed, dual estimation) problem. When assessing the evidence for the absence of a target, there could be no

sensory evidence because the target is not there or because precision is low (or both). This puts pressure on the estimation of precision to resolve conditional dependencies between posterior beliefs about target presence and the precision with which it can be detected. In short, two things have to be estimated; the posterior expectation about the target and posterior beliefs about precision (Clark, 2013; Feldman & Friston, 2010; Haarsma et al., 2018; Palmer, Auksztulewicz, Ondobaka, & Kilner, 2019; Parr, Benrimoh, Vincent, & Friston, 2018).

In line with a role in monitoring of attention or precision, right TPJ showed a negative effect of confidence that was stronger for 'target absent' responses compared to 'target present' responses in detection. This cluster was closest to the posterior subdivision of the right TPJ [TPJp-R; Igelström, Webb, & Graziano (2015)], which is most strongly associated with reasoning about others' beliefs (Igelström, Webb, Kelly, & Graziano, 2016). In addition to its role in Theory of Mind (Lee & McCarthy, 2016; Saxe & Wexler, 2005), previous work has highlighted the importance of the rTPJ in controlling attention (Dugué, Merriam, Heeger, & Carrasco, 2018; Geng & Vossel, 2013; Lee & McCarthy, 2016; Marois, Yi, & Chun, 2004) and filtering distractors in visual search (Shulman, Astafiev, McAvoy, d'Avossa, & Corbetta, 2007). Furthermore, damage to the rTPJ can result in visual hemineglect: a condition in which stimuli in the left visual hemifield fail to reach awareness (Shulman, Astafiev, McAvoy, d'Avossa, & Corbetta, 2007). Together, these observations have led to a proposal (the 'Attention Schema Theory') that the rTPJ is maintaining a simplified representation of one's own and others' attentional states, and that this function makes this region essential for maintaining conscious awareness (Graziano & Webb, 2015).

The current Attention Monitoring model fits well with the Attention Schema Theory. A representation of one's current attentional state is a useful source of information for determining confidence in detection judgments, because stimuli are more likely to be missed when participants are not paying careful attention. This will be specifically useful for judgments about stimulus absence: if a target was not observed, the participant may reason something along the lines of 'given my current state of attention, I was not very likely to miss a target, therefore I can be very confident that a target was not presented.' In support of this idea, the typically poor metacognitive evaluations of decisions about stimulus absence are partially recovered when task difficulty is controlled by manipulating attention rather than stimulus visibility (Kanai, Walsh, & Tseng, 2010; Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018), suggesting that subjects may harness information about their attentional state to inform their confidence judgments. Interestingly, the frontopolar cortex, which showed a detection-specific quadratic effect of confidence in our experiment, has also been implicated in attentional control via the gating of internal and external modes of attention (Burgess, Gilbert, & Dumontheil, 2007) and in discriminating between imagined and externally perceived memory items (Simons, Davis, Gilbert, Frith, & Burgess, 2006; Turner, Simons, Gilbert, Frith, & Burgess, 2008). Together, the engagement of this set of regions in detection confidence hints at a potential role for self-monitoring of attention in metacognition of detection.

To conclude, we find a quadratic effect of confidence in detection judgments in several brain regions, including the frontopolar cortex and rTPJ. In the frontopolar

cortex, this quadratic effect was not seen for discrimination judgments. In the rTPJ, we also found a linear effect of confidence that was more negative for judgments about stimulus absence compared to judgments about stimulus presence. We consider three computational accounts of our results, two of which implicate the learning and estimation of signal-to-noise statistics as promising accounts of the observed detection-specific activation profiles. However, while each of these accounts could explain some of our findings, none of the models could provide a complete account of the data. Further work is needed to decide between these alternatives, or to suggest new ones.

Chapter 5

Metacognitive asymmetries in visual perception

Matan Mazor, Rani Moran & Stephen M. Fleming

In previous chapters we examined inference about absence and its relation to self-modelling, focusing on the absence of entire shapes (Chapter 1), visual gratings (Chapter 4) and non-random patterns in otherwise random displays (Chapter 3). In all cases we find that decisions about absence are slower than decisions about presence, and in chapters 3 and 4 we further replicate findings of higher levels of confidence and improved metacognitive sensitivity for decisions about the presence compared to the absence of objects. in this last chapter, based on a Registered Report, I ask how far can we stretch the definition of 'absence,' focusing on the absence of stimulus features or expectation violations, rather than entire objects or stimuli. Our pre-registered prediction was that differences in the processing of presence and absence reflect a default mode of reasoning: assuming absence unless evidence is available for presence. In a Registered Report, we predicted asymmetries in response time, confidence, and metacognitive sensitivity in discriminating between stimulus categories that vary in the presence or absence of a distinguishing feature, or in their compliance with an expected default state. Six experiments, using six pairs of stimuli, provide evidence that like the presence of entire shapes or gratings, the presence of local and global stimulus features gives rise to faster, more confident responses. Contrary to our hypothesis, however, the presence or absence of a local feature has no effect on metacognitive sensitivity. Our results weigh against our proposal of a link between the detection metacognitive asymmetry and default reasoning, and are instead consistent with a low-level visual origin of the metacognitive asymmetry between detection 'yes' and 'no' responses.

5.1 Introduction

At any given moment, there are many more things that are not there than things that are there. As a result, and in order to efficiently represent the environment, perceptual and cognitive systems have evolved to represent presences, and absence

is implicitly represented as a default state (Oaksford, 2002; Oaksford & Chater, 2001). One corollary of this is that presence can be inferred from bottom-up sensory signals, but absence is never explicitly represented in sensory channels and must instead by inferred based on top-down expectations about the likelihood of detecting a hypothetical signal, had it been present. Experiments on human subjects accordingly suggest that representing absence is more cognitively demanding than representing presence, even in simple perceptual tasks, as is evident in slower reactions to stimulus absence than stimulus presence in near-threshold visual detection (Mazor, Friston, & Fleming, 2020), in a general difficulty to form associations with absence (Newman, Wolff, & Hearst, 1980), and in the late acquisition of explicit representations of absence in development (Coldren & Haaf, 2000; e.g., Sainsbury, 1971; for a review on the representation of nothing see Hearst, 1991).

An overarching difficulty in representing absence may reflect the metacognitive nature of absence representations; to represent something as absent, one must assume that they would have detected it had it been present. In philosophical writings, this form of higher-order, metacognitive inference-about-absence is known as *argument from epistemic closure*, or *argument from self-knowledge* [*If it was true, I would have known it*; Walton (1992); De Cornulier (1988)]. Strikingly, quantitative measures of metacognitive insight are consistently found to be lower for decisions about absence than for decisions about presence. When asked to rate their subjective confidence following near-threshold detection decisions, subjective confidence ratings following 'target absent' judgments are commonly lower, and less aligned with objective accuracy, than following 'target present' judgments [Fig. 5.1; Kanai, Walsh, & Tseng (2010); Meuwese, Loon, Lamme, & Fahrenfort (2014); Kellij, Fahrenfort, Lau, Peters, & Odegaard (2018); Mazor, Friston, & Fleming (2020)].

Metacognitive asymmetries have not only been observed for judgments about the presence or absence of whole physical objects and stimuli, but also for the presence or absence of cognitive variables such as memory traces. For instance, in recognition memory, subjects typically show poor metacognitive sensitivity for judgments about the absence of memories [such as when judging that they haven't seen a study item before; Higham, Perfect, & Bruno (2009)]. Unlike the absence of a visual stimulus, the absence of a memory is not localized in space and does not correspond with a specific representation of 'nothing.'

One way of conceptualizing these findings is that absence asymmetries emerge as a function of default reasoning - absences are considered the 'default,' and information about perceptual or mnemonic presence is accumulated and tested against this default. For instance, an asymmetry may emerge in recognition memory because the presence of memories is actively represented, and the absence of memories is assumed as the default unless evidence is available for the contrary. In the same way, other visual features that are not typically treated as presences or absences may still be coded relative to a default, assuming one state unless evidence is available for the contrary (e.g., assuming that a cookie is sweet rather than salty). However, whether a metacognitive asymmetry in processing presence and absence generalizes to these more abstract violations of default expectations remains unknown. Here we set out to map out the structure of absence representations by testing for metacognitive asymmetries

in the presence and absence of attributes at different levels of representation - from concrete objects, to visual features, to violations of default expectations.

Our choice of stimuli draws inspiration from visual search - a field where asymmetries are observed for a variety of stimulus types and features. In visual search, participants typically take longer to search for a target that is marked by the absence of a distinguishing feature, as compared to searching for a target that is marked by the presence of a feature relative to distractors (A. Treisman & Gormican, 1988; A. Treisman & Souther, 1985). Interestingly, *search asymmetries* have been demonstrated not only for the absence or presence of concrete physical features, but also for the presence or absence of deviations from a more abstract default state, which can be based on experience, culture, and contextual expectations [see methods; Von Grünau & Dubé (1994); U. Frith (1974); Wang, Cavanagh, & Green (1994); Gandolfo & Downing (2020)].

Of special interest for our study are these latter asymmetries due to expectation violations, and their relation with asymmetries induced by the presence or absence of local and global features. Observing a metacognitive asymmetry for expectation violations as well as for the presence and absence of object features would support a strong link between the representation of absence and default reasoning, where differences in metacognitive sensitivity reflect differences in the processing of information that agrees or contrasts with the expected default state.

While traditional accounts interpreted visual search asymmetries as reflecting a qualitative advantage for the cognitive representation of presence [affording a parallel search in the case of feature-present search only; A. Treisman & Gormican (1988)], other models attribute the asymmetry to differences in the distributions of perceptual signals already at the sensory level (Dosher, Han, & Lu, 2004; Vincent, 2011). Similarly, in the case of metacognitive asymmetries, the idea that decisions about absence are qualitatively different from decisions about presence has been challenged by an excellent fit of simple models that assume unequal variance for the signal-present and signal-absent sensory distributions, a model that does not assume any qualitative difference between the two decisions (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018). Deciding between these model families is beyond the scope of this project. However, identifying metacognitive asymmetries for abstract cognitive variables such as familiarity could help refine these models, for instance by revealing that representing deviations from a default state is an overarching principle of cognitive organization, one that goes beyond specific features of visual perception.

5.2 Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. The full registered protocol is available at osf.io/ed8n7.

We ran six experiments, that were identical except for the identity of the two stimuli S_1 and S_2 (and of the stimulus used for backward masking; see section 5.5 for details). Our choice of stimuli for this study was based on the visual search literature.

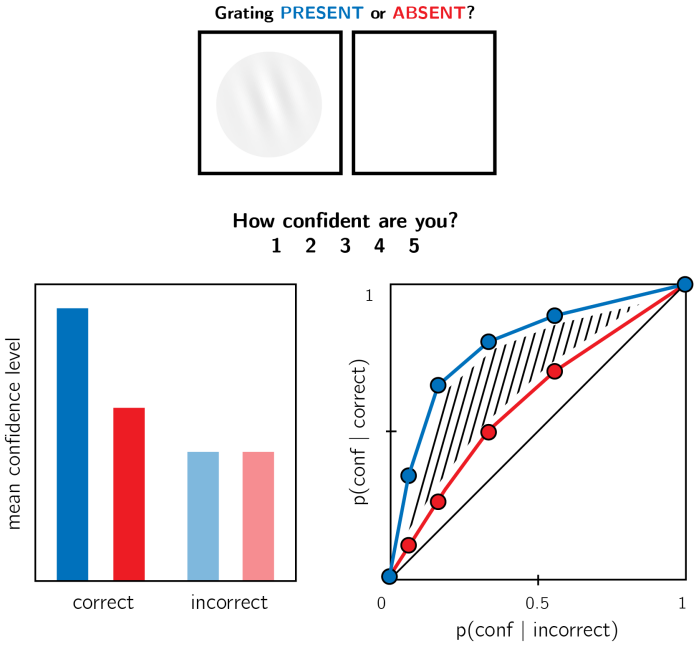


Figure 5.1: In visual detection, subjective confidence ratings following judgments about target absence are typically lower, and less correlated with objective accuracy than following judgments about target presence. Top panel: a typical detection experiment. The participant reports whether a visual grating was present or absent, and then rates their subjective decision confidence. Bottom left: typically, mean confidence in 'yes' responses (blue) is higher than in 'no' responses (red). This effect is much more pronounced in correct trials. Bottom right: the interaction between accuracy and response type on confidence (metacognitive asymmetry) manifests as a lower area under the response conditional type 2 ROC curve (AUROC2) for 'no' responses compared with 'yes' responses. Plots do not directly correspond to a specific dataset, but portray typical results in visual detection.

For some stimulus pairs S_1 and S_2, searching for one S_1 among multiple S_2s is more efficient than searching for one S_2 among multiple S_1s. Such *search asymmetries* have been reported for stimulus pairs that are identical except for the presence and absence of a distinguishing feature. Importantly, distinguishing features vary in their level of abstraction, from concrete *local features* [finding a Q among Os is easier than the inverse search; A. Treisman & Souther (1985)], through *global features* [finding a curved line among straight lines is easier than the inverse search; A. Treisman & Gormican (1988)], and up to the presence or absence of abstract *expectation violations* [searching for an upward-tilted cube among downward-tilted cubes is easier than the inverse search, in line with a general expectation to see objects on the ground rather than floating in space; Von Grünau & Dubé (1994)]. We treat these three types of asymmetries as reflecting a default-reasoning mode of representation, where the

absence of features and/or the adherence of objects to prior expectations is tentatively accepted as a default by the visual system, unless evidence is available for the contrary (A. Treisman & Gormican, 1988; A. Treisman & Souther, 1985). In this study, we test for metacognitive asymmetries for two stimulus features in each category, in six separate experiments with different participants (Fig. 5.2). For each of the following stimulus pairs, searching for S_1 among multiple instances of S_2 has been found to be more efficient than the inverse search:

1. **Local feature: Addition of a stimulus part**. Q and O were used as S_1 and S_2, respectively (A. Treisman & Souther, 1985).
2. **Local feature: Open ends**. C and O were used as S_1 and S_2, respectively (Takeda & Yagi, 2000; A. Treisman & Gormican, 1988; A. Treisman & Souther, 1985).
3. **Global feature: Orientation**. Tilted and vertical lines were used S_1 and S_2, respectively (A. Treisman & Gormican, 1988).
4. **Global feature: Curvature**. Curved and straight lines were used as S_1 and S_2, respectively (A. Treisman & Gormican, 1988).
5. **Expectation violation: Viewing angle**. Upward and Downward tilted cubes were used as S_1 and S_2, respectively (Von Grünau & Dubé, 1994).
6. **Expectation violation: Letter inversion**. Flipped and normal N were used as S_1 and S_2, respectively (U. Frith, 1974; Wang, Cavanagh, & Green, 1994).

The experiments quantified participants' metacognitive sensitivity for discrimination judgments between S_1 and S_2.

5.2.1 Participants

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). Participants were recruited via Prolific, and gave informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. For each of the six experiments, we aimed to collected data until we reached 106 included participants (after applying our pre-registered exclusion criteria). The entire experiment took 10-15 minutes to complete. Participants were paid between £1.25 to £2 for their participation, maintaining a median hourly wage of £6 or higher.

5.2.2 Procedure

Experiments were programmed using the jsPsych and P5 JavaScript packages (De Leeuw, 2015; McCarthy, 2015), and were hosted on a JATOS server (Lange, Kuhn, & Filevich, 2015).

After instructions, a practice phase, and a multiple-choice comprehension check, the main part of the experiment started. It comprised 96 trials separated into 6 blocks. Only the last 5 blocks were analyzed.

On each trial, participants made discrimination judgments on masked stimuli, and rated their subjective decision confidence on a continuous scale. After a fixation cross (500 ms), the target stimulus (S_1 or S_2) was presented in the center of the screen for 50 ms, followed by a mask (100 ms). Stimulus onset asynchrony (SOA) was calibrated online in a 1-up-2-down procedure (Levitt, 1971), with a multiplicative step factor of 0.9, and starting at 30 milliseconds. Participants then used their keyboard to make a discrimination judgment. Stimulus-key mapping was counterbalanced between participants. Following response, subjective confidence ratings were given on an analog scale by controlling the size of a colored circle with the computer mouse. High confidence was mapped to a big, blue circle, and low confidence to a small, red circle. We chose a continuous (rather than a more typical discrete) confidence scale in order to ensure sufficient variation in confidence ratings within the dynamic range of individual participants. This variation is useful for the extraction of the area under response conditional type 2 ROC curves (AUROC2). The confidence rating phase terminated once participants clicked their mouse, but not before 2000 ms. No trial-specific feedback was delivered about accuracy. In order to keep participants motivated and engaged, block-wise feedback was delivered between experimental blocks about overall accuracy, mean confidence in correct responses, and mean confidence in incorrect responses. Online demos the experiments can be accessed at matanmazor.github.io/asymmetry.

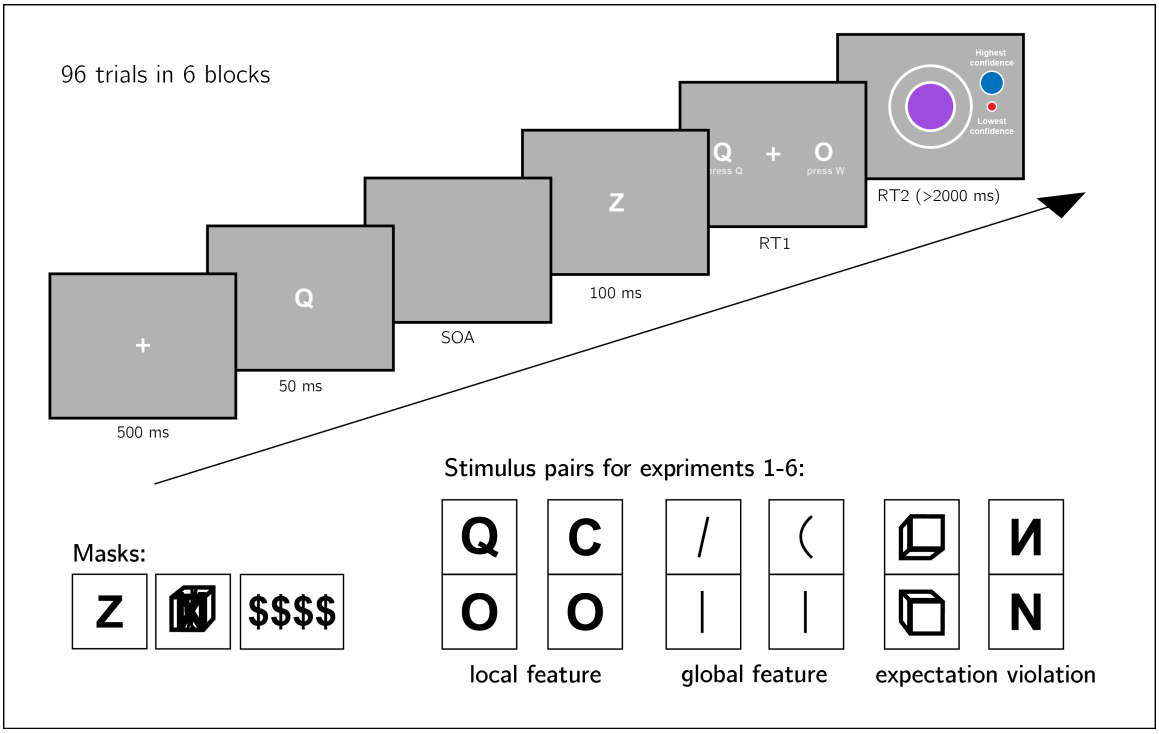


Figure 5.2: Experiment design. Metacognitive asymmetry effects were tested for six stimulus features in six separate experiments, encompassing three levels of abstraction: local features, global features, and expectation violations. The presented trial corresponds to the first stimulus pair, with Q and O as stimuli.

Randomization

The order and timing of experimental events was determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized in a way that ensures registration time-locking (Mazor, Mazor, & Mukamel, 2019).

5.2.3 Data analysis

We used R [Version 4.0.5; R Core Team (2019)] and the R-packages *BayesFactor* [Version 0.9.12.4.2; Richard D. Morey & Rouder (2018)], *broom* [Version 0.7.9; Robinson & Hayes (2020)], *cowplot* [Version 1.1.1; Wilke (2019)], *dplyr* [Version 1.0.7; Wickham, François, Henry, & Müller (2020)], *ggplot2* [Version 3.3.5; Wickham (2016)], *lmerTest* [Version 3.1.3; Kuznetsova, Brockhoff, & Christensen (2017)], *lsr* [Version 0.5; Navarro (2015)], *MESS* [Version 0.5.7; Ekstrøm (2019)], *papaja* [Version 0.1.0.9997; Aust & Barth (2020)], *pracma* [Version 2.3.3; Borchers (2019)], *pwr* [Version 1.3.0; Champely (2020)], and *tidyr* [Version 1.1.3; Wickham & Henry (2020)] for all our analyses.

For each of the six stimulus pairs [S_1, S_2], we tested the following hypotheses:

1. **Hypothesis 1**: Subjective confidence is higher for S_1 responses than for S_2 responses.

For each of the six stimulus pairs, we tested the null hypothesis that subjective confidence for S_1 responses is equal to or lower than subjective confidence for the feature-absent stimulus ($H_o : conf_{S_1} \leq Conf_{S_2}$).

2. **Hypothesis 2**: Metacognitive sensitivity, measured as the area under the response conditional type 2 ROC curves, is higher for S_1 responses than for S_2 responses.

For each of the six stimulus pairs, we tested the null hypothesis that metacognitive sensitivity for S_1 responses is equal to or lower than metacognitive sensitivity for the S_2 responses ($H_o : auROC_{S_1} \leq auROC_{S_2}$).

3. **Hypothesis 3**: Metacognitive sensitivity, measured as the area under the response conditional type 2 ROC curves, is higher for S_1 responses than for S_2 responses, to a greater extent than expected from an equivalent equal-variance SDT model.

For each of the six stimulus pairs, we tested the null hypothesis that difference between metacognitive sensitivities for S_1 and S_2 responses is lower than the difference expected from an equivalent equal-variance SDT model ($H_o : (auROC_{S_1} - auROC_{S_2}) \leq (\widehat{auROC}_{S_1} - \widehat{auROC}_{S_2})$ where $\widehat{auROC}$ is the expected auROC under an equal variance SDT model with equal sensitivity, criterion, and distribution of confidence ratings in incorrect responses).

4. **Hypothesis 4**: S_1 responses are faster on average than S_2 responses.

For each of the six stimulus pairs, we tested the null hypothesis that log-transformed response times for S_1 responses are equal to or higher than log-transformed response times for S_2 responses ($H_o : log(RT_{S_1}) \geq log(RT_{S_2})$).

Hypotheses 1 and 2 correspond to the effects of stimulus type on metacognitive bias and metacognitive sensitivity, respectively. Although these two measures are theoretically independent, both bias and sensitivity are found to vary between detection 'yes' and 'no' responses.

Based on pilot data and previous experiments examining near-threshold perceptual detection and discrimination, we did not expect a response bias (such that the probability of responding S_1 is significantly different from 0.5 across participants). However, such a response bias, if found, may bias metacognitive asymmetry estimates as measured with response conditional type 2 ROC curves. Hypothesis 3 was designed to confirm that metacognitive asymmetry is higher than that expected from an equivalent equal-variance SDT model with the same response bias, sensitivity, and distribution of confidence ratings in incorrect responses as in the actual data. We interpreted conflicting results for Hypotheses 2 and 3 as evidence for a metacognitive asymmetry that is driven or masked by a response bias.

Hypothesis 4 is motivated by two observations from previous studies. First, detection 'yes' responses are faster than detection 'no' responses (Mazor, Friston, & Fleming, 2020). And second, when participants are not under strict time pressure, reaction time inversely scales with confidence (Calder-Travis, Charles, Bogacz, & Yeung, 2020; Henmon, 1911; Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010). Based on these findings, if S_1 and S_2 responses are similar to detection 'yes' and 'no' responses not only in explicit confidence judgments, but also in response times, we should also expect a response time difference for these stimulus pairs.

5.2.4 Dependent variables and analysis plan

Response conditional type 2 ROC (rcROC) curves were extracted by plotting the empirical cumulative distribution of confidence ratings for correct responses against the same cumulative distribution for incorrect responses. This was done separately for the two responses S_1 and S_2, resulting in two curves. The area under the rcROC curve is a measure of metacognitive sensitivity (Fleming & Lau, 2014). The difference between the areas for the two responses is a measure of metacognitive asymmetry (Meuwese, Loon, Lamme, & Fahrenfort, 2014). This difference was used to test Hypothesis 2.

In order to test hypothesis 3, SDT-derived rcROC curves were plotted in the following way. For each response, we plotted the empirical cumulative distribution for incorrect responses on the x axis against the cumulative distribution for correct responses that would be expected in an equal-variance SDT model with matching sensitivity and response bias on the y axis. The difference between the areas of these theoretically derived rcROC curves was compared against the difference between the true rcROC curves.

For visualization purposes only, confidence ratings were divided into 20 bins, tailored for each participant to cover their dynamic range of confidence ratings.

For each of the six experiments, Hypotheses 1-4 were tested using a one tailed

t-test at the group level with $\alpha = 0.05$. The summary statistic at the single subject level was difference in mean confidence between S_1 and S_2 responses for Hypothesis 1, difference in area under the rcROC curve between S_1 and S_2 responses (ΔAUC) for Hypothesis 2, difference in ΔAUC between true confidence distributions and SDT-derived confidence distributions for hypothesis 3, and difference in mean log response time between S_1 and S_2 responses for Hypothesis 4.

In addition, a Bayes factor was computed using the BayesFactor R package (Richard D. Morey, Rouder, Jamil, & Morey, 2015) and using a Jeffrey-Zellner-Siow (Cauchy) Prior with an rscale parameter of 0.65, representative of the similar standardized effect sizes we observe for Hypotheses 1-4 in our pilot data.

We based our inference on the resulting Bayes Factors.

5.2.5 Statistical power

Statistical power calculations were performed using the R-pwr packages pwr (Champely, 2020) and PowerTOST (Labes, Schütz, Lang, & Labes, 2020).

1. Hypothesis 1 (MEAN CONFIDENCE): With 106 participants, we had statistical power of 95% to detect effects of size 0.32, which is less than the standardized effect size we observed for confidence in our pilot sample ($d = 0.66$).

2. Hypothesis 2 (METACOGNITIVE ASYMMETRY): With 106 participants, we had statistical power of 95% to detect effects of size 0.32, which is less than the standardized effect size we observed for metacognitive sensitivity in our pilot sample ($d = 0.73$).

3. Hypothesis 3 (METACOGNITIVE ASYMMETRY: CONTROL): With 106 participants, we had statistical power of 95% to detect effects of size 0.32, which is less than the standardized effect size we observed for metacognitive sensitivity, controlling for response bias, in our pilot sample ($d = 0.81$).

4. Hypothesis 4 (RESPONSE TIME): With 106 participants, we had statistical power of 95% to detect effects of size 0.32, which is less than the standardized effect size we observed for response time in our pilot sample ($d = 0.61$).

Finally, in case that the true effect size equals 0, a Bayes Factor with our chosen prior for the alternative hypothesis will support the null in 95 out of 100 repetitions, and will support the null with a BF_{01} higher than 3 in 79 out of 100 repetitions. In a case where the true effect size is sampled from a Cauchy distribution with a scale factor of 0.65, a Bayes Factor with our chosen prior for the alternative hypothesis will support the alternative hypothesis in 76 out of 100 repetitions, support the alternative hypothesis with a BF_{10} higher than 3 in 70 out of 100 repetitions, and support the null hypothesis with a BF_{01} higher than 3 in 15 out of 100 hypotheses (based on an adaptation of simulation code from Lakens, 2016).

Rejection criteria

Participants were excluded for performing below 60% accuracy, for having extremely fast or slow reaction times (below 250 milliseconds or above 5 seconds in more than 25% of the trials), and for failing the comprehension check. Finally, for type-2 ROC curves to be generated, some responses must be incorrect, and for them to be informative some variability in confidence ratings is necessary. Thus, participants who committed less than two of each error type (for example, mistaking a Q for an O and mistaking an O for a Q), or who reported less than two different confidence levels for each of the two responses were excluded from all analyses.

Trials with response time below 250 milliseconds or above 5 seconds were excluded.

5.3 Data availability

All raw data is fully available on OSF and on the study's GitHub respository: https://github.com/matanmazor/asymmetry.

5.4 Code availability

All analysis code is openly shared on the study's GitHub repository: https://github.com/matanmazor/asymmetry. For complete reproducibility, the RMarkdown file used to generate the final version of the manuscript, including the generation of all figures and extraction of all test statistics, is also available on our GitHub repository.

5.5 Deviations from pre-registration

- *Stimulus used for backward masking*: We planned to use the same stimulus (the letter Z) for backward masking in all six experiments. This mask was effective in Experiments 1 and 2, but in Experiment 3 overly high accuracy levels indicated that for these stimuli the mask was not salient enough. For a subset of participants in Exp. 3, an overlay of all 7 stimuli from experiments 3-6 (vertical, tilted, and curved lines, upward-tilted and downward-tilted cubes, and normal and flipped Ns) was used. For the remaining participants and experiments, we used four dollar signs as our mask. See Fig. 5.2 for depictions of the three masks.

- *Rejection criteria*: In our pre-registration we explain that informative rcROC curves can only be generated if participants make errors. When analyzing the data we came to realize that an additional prerequisite for rcROC curves to be informative is that the variance in confidence ratings is higher than zero, otherwise the curve is diagonal. We therefore required that participants report at least two different confidence levels for each response. Participants that did not meet this additional criterion were excluded from all analyses.

- *Monetary compensation*: For some of the experiments, we noticed that participants completed the experiment more quickly than what we had originally estimated. We therefore reduced our offered payment for some of the experiments, while maintaining a median hourly wage of £6 or higher.

5.6 Results

A summary of the results from all six experiments is available in section 5.6.7 and in Figures 5.3, 5.4 and 5.5.

5.6.1 Experiment 1: *Q* vs. *O*

In Experiment 1, we examined discrimination judgments between the two letters *Q* and *O*. Based on a search asymmetry for these letters [*Q*s are found faster than *O*s than vice-versa; A. Treisman & Souther (1985)], we hypothesized that a similar asymmetry would emerge in subjective confidence judgments, such that metacognitive sensitivity for *Q* responses will be higher than for *O* responses. We used the letter *Z* as our backward mask.

205 participants (median reported age: 32; range: [18-74]) were recruited from Prolific for Experiment 1.

Median completion time was 13.12 minutes. Mean proportion correct was 0.74. Participants reported seeing an *O* on 47% of trials. In a deviation from our preregistration, we excluded 9 participants for having zero variance in their confidence ratings for at least one of the two responses (see Section 5.5). Overall we excluded 71 participants based on our exclusion criteria, leaving 134 participants for the main analysis. Due to a technical error in data collection, this figure is higher than that specified in our preregistration document (N=106). Going forward, only data from included participants is analyzed.

Mean proportion correct among the included participants was $M = 0.74$, 95% CI [0.73, 0.75]. Mean SOA in the last trial was $M = 47.50$, 95% CI [39.39, 55.61]. Participants showed no consistent bias in their responses (quantified as the probability of a 'Q' response minus 0.5; $M = 0.02$, 95% CI [0.00, 0.04]). On a scale of 0 to 1, mean confidence level was $M = 0.49$, 95% CI [0.45, 0.53]. Confidence was higher for correct than for incorrect responses ($M_d = 0.15$, 95% CI [0.13, 0.17], $t(133) = 14.85$, $p < .001$).

Hypothesis 1: In line with our hypothesis, confidence was generally higher for *Q* (feature present) responses than for *O* (feature absent) responses ($t(133) = 7.52$, $p < .001$; Cohen's d = 0.65; $BF_{10} = 1.07 \times 10^9$; see Fig. 5.3, panel 1).

Hypothesis 2: In order to measure metacognitive asymmetry, we extracted the response conditional type 2 ROC (rcROC) curves for the two responses (*Q* and *O*) in the discrimination task. This was done by plotting the cumulative distribution of confidence ratings (high to low) for correct responses against the same distribution for incorrect responses. The area under the rcROC curve (auROC2) was then taken as a measure of metacognitive sensitivity (Kanai, Walsh, & Tseng, 2010; Meuwese, Loon,

Lamme, & Fahrenfort, 2014). In line with our hypothesis, auROC2 for Q responses ($M = 0.72$, 95% CI [0.70, 0.74]) was higher than for O responses ($M = 0.68$, 95% CI [0.66, 0.70]; $t(133) = 2.96$, $p = .002$; Cohen's d = 0.26; $BF_{10} = 6.56$; see Figure 5.4, panel 1), similar to the documented metacognitive asymmetry for detection judgments.

Hypothesis 3: Metacognitive asymmetry was not significantly higher than what is expected based on an equal-variance SDT model with the same response bias and sensitivity as the subjects ($t(133) = 0.97$, $p = .167$; Cohen's d=0.08). A Bayes Factor indicated that our results are more likely under a model that assumes no additional metacognitive asymmetry ($BF_{01} = 6.07$).

Hypothesis 4: In line with our hypothesis, Q responses were faster on average than O responses by 37 ms. ($t(133) = -2.99$, $p = .002$; Cohen's d = 0.26; $BF_{10} = 7.05$; see Fig. 5.3, panel 1).

In summary, in Experiment 1 we found that Q responses were faster and accompanied by higher subjective confidence, in line with a processing advantage for feature-presence. Metacognitive asymmetry however did not go beyond what is expected from an equal-variance SDT model for these stimuli, taking into account response biases.

5.6.2 Experiment 2: C vs. O

In Experiment 2, we looked at discrimination judgments between the two letters C and O. Based on a search asymmetry for these letters [Cs are found faster among Os than vice versa; A. Treisman & Souther (1985); Takeda & Yagi (2000); A. Treisman & Gormican (1988)], we hypothesized that a similar asymmetry would emerge in subjective confidence judgments, such that metacognitive sensitivity for perceiving a C will be higher than for perceiving an O. We used the letter Z as our backward mask.

143 participants (median reported age: 26; range: [20-51]) were recruited from Prolific for Experiment 2.

Median completion time was 12.80 minutes. Mean proportion correct was 0.75, and participants reported seeing an O on 43% of trials. In a deviation from our pre-registration, we excluded 8 participants for having zero variance in their confidence ratings for at least one of the two responses (see Section 5.5). Overall we excluded 37 participants, leaving 106 participants for the main analysis. Going forward, only data from included participants is analyzed.

Mean proportion correct among included participants was $M = 0.74$, 95% CI [0.73, 0.75]. The mean SOA of the last trial was $M = 40.18$, 95% CI [34.37, 46.00]. Participants showed a consistent bias toward reporting a C rather than an O ($M = 0.07$, 95% CI [0.05, 0.08]). On a scale of 0 to 1, mean confidence level was $M = 0.52$, 95% CI [0.48, 0.56]. Confidence was higher for correct than for incorrect responses ($M_d = 0.17$, 95% CI [0.15, 0.19], $t(105) = 15.05$, $p < .001$).

Hypothesis 1: In line with our hypothesis, confidence was generally higher for C (feature present) responses than for O (feature absent) responses ($M_d = 0.05$, 95% CI [0.03, ∞], $t(105) = 3.59$, $p < .001$; Cohen's d = 0.35; $BF_{10} = 42.62$; see Figure 5.3, panel 2).

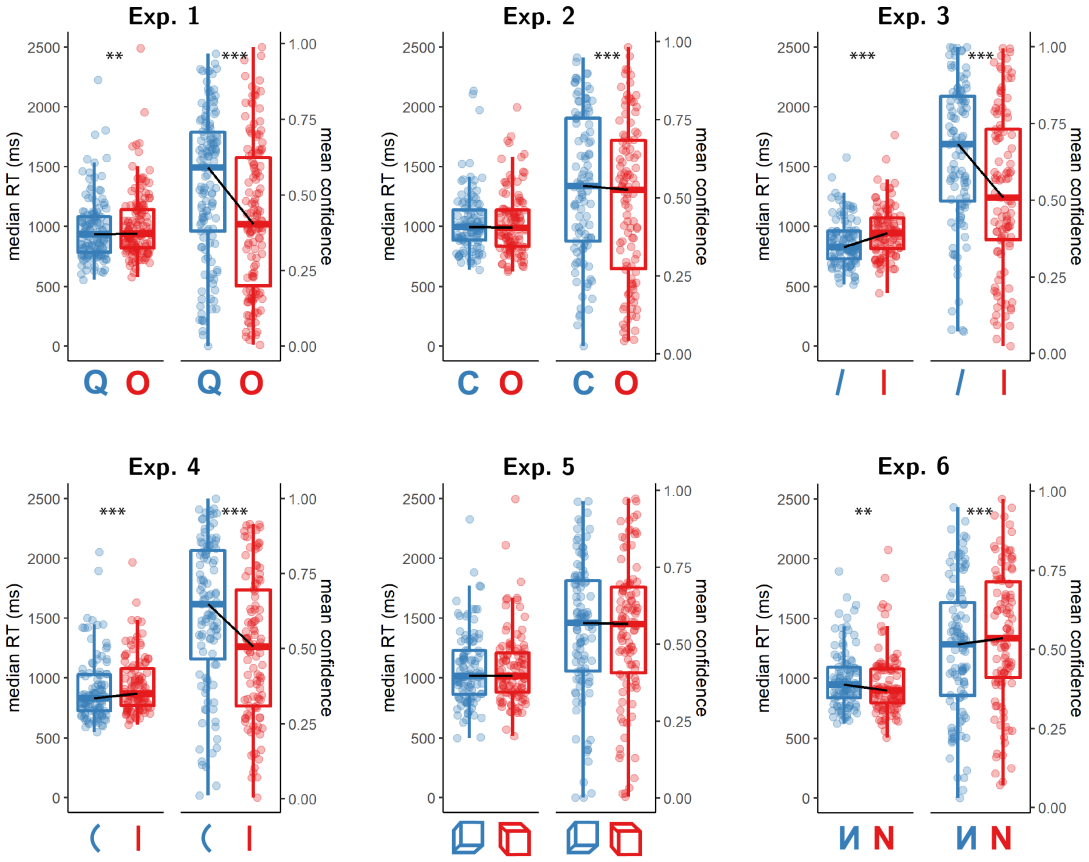


Figure 5.3: Reaction time and confidence distributions for Experiments 1-6. Box edges and central lines represent the 25, 50 and 75 quantiles. Whiskers cover data points within four inter-quartile ranges around the median. Black lines connect the median values for the two responses. Stars represent significance in a two-sided t-test: **: p<0.01, ***: p<0.001

Hypothesis 2: Opposite to our prediction, auROC2 for C responses ($M = 0.70$, 95% CI [0.68, 0.72]) was *lower* than for O responses ($M = 0.75$, 95% CI [0.73, 0.78]; $t(105) = -3.53$, $p > .999$; Cohen's d = 0.34; see Figure 5.4, panel 2.). Bayes Factor strongly supported the alternative ($BF_{10} = 35.19$). Note that our prior on effect sizes was symmetric around zero, such that support for the alternative is obtained for negative, as well as positive effects.

Hypothesis 3: Metacognitive sensitivity for C responses was still higher than for O responses after controlling for bias (Cohen's d=0.49; $BF_{10} = 6.46 \times 10^3$).

Hypothesis 4: Contrary to our hypothesis, response times for C and for O responses were highly similar, with a median difference of 6 ms. ($t(105) = 0.01$, $p = .504$; Cohen's d = 0.00; $BF_{01} = 8.57$; see Figure 5.3, panel 2).

In summary, in Experiment 2 we found a dissociation between our two confidence-related measures. As we hypothesized, participants were generally more confident in

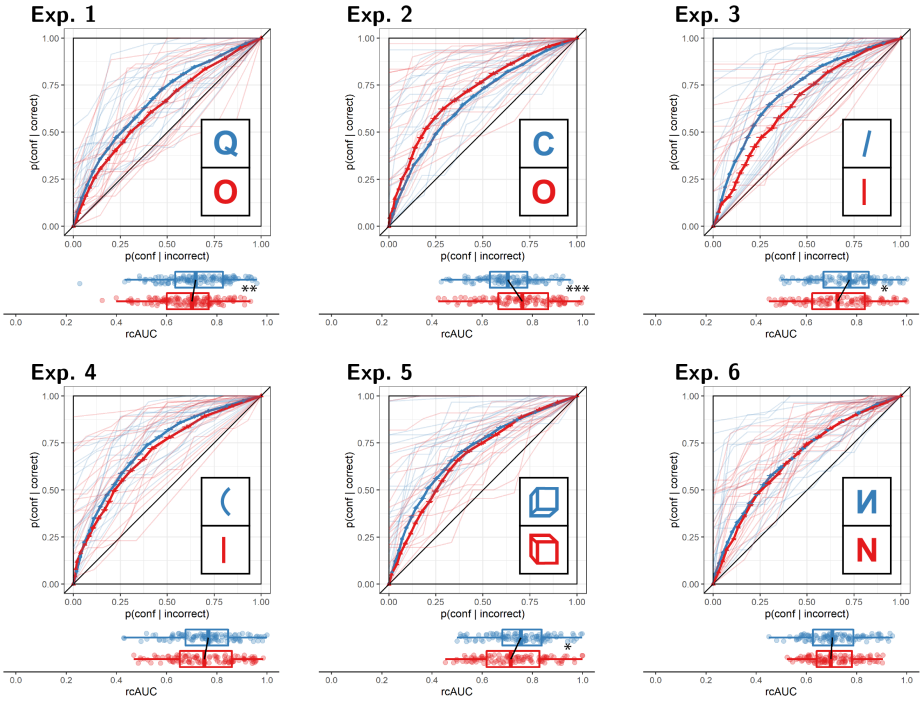


Figure 5.4: Response conditional type 2 ROC (rcROC) curves for Experiments 1-6. The area under the curve is a measure of metacognitive sensitivity. Error bars stand for the standard error of the mean. For illustration, the response conditional ROC (rcROC) curves of the first 20 participants of each Experiment are plotted in low opacity. Below each ROC: distributions of the area under the curve for the two responses, across participants. Same conventions as Fig. 5.3. Stars represent significance in a two-sided t-test: *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$

their C (feature present) responses, but their metacognitive sensitivity was higher following O (feature absent) responses. We found no reliable difference in response times between these two responses.

5.6.3 Experiment 3: tilted vs. vertical lines

In Experiment 3, we looked at discrimination judgments between tilted and vertical lines. Based on a search asymmetry for these stimuli [tilted lines are found faster among vertical lines than vice versa; A. Treisman & Gormican (1988)], we hypothesized that a similar asymmetry would emerge in subjective confidence judgments, such that metacognitive sensitivity for perceiving a tilted line will be higher than for perceiving a vertical line. As described in section 5.5, too high accuracy in the first participants led us to change our masking stimulus, first to an overlay of all stimuli and then to four dollar signs. We present here the combined results from these two cohorts of participants. The results were qualitatively similar in the two cohorts.

304 participants (median reported age: 31; range: [18-66]) were recruited from

Prolific for Experiment 3. Due to shorter than expected completion times in the first 94 participants, the remaining participants were paid £1.25, equivalent to an hourly wage of £6.

Median completion time was 12.43 minutes. Mean proportion correct was 0.86, and participants reported seeing a vertical line on 44% of trials. In a deviation from our pre-registration, we excluded 25 participants for having zero variance in their confidence ratings for at least one of the two responses (see Section 5.5). Overall we excluded 198 participants, leaving 106 participants for the main analysis. Going forward, only data from included participants is analyzed.

Mean proportion correct among included participants was $M = 0.79$, 95% CI [0.78, 0.81]. The mean SOA of the last trial was $M = 30.83$, 95% CI [25.99, 35.68]. Participants showed a consistent bias toward reporting a tilted rather than a vertical line ($M = 0.06$, 95% CI [0.04, 0.08]). On a scale of 0 to 1, mean confidence level was $M = 0.61$, 95% CI [0.56, 0.65]. Confidence was higher for correct than for incorrect responses ($M_d = 0.18$, 95% CI [0.15, 0.20], $t(105) = 13.42$, $p < .001$).

Hypothesis 1: In line with our hypothesis, confidence was generally higher for tilted lines (feature present) responses than for vertical lines (feature absent) responses ($M_d = 0.12$, 95% CI [0.09, ∞], $t(105) = 7.18$, $p < .001$; Cohen's d = 0.70; $BF_{10} = 8.89 \times 10^7$; see Figure 5.3, panel 3).

Hypothesis 2: Contrary to our prediction, Bayes Factor analysis did not provide evidence for or against a difference in auROC2 between reports of seeing a tilted line ($M = 0.76$, 95% CI [0.74, 0.78]) and reports of seeing a vertical line ($M = 0.73$, 95% CI [0.70, 0.75]; Cohen's d = 0.18; $BF_{01} = 1.59$; see Figure 5.4, panel 3.). A difference in metacognitive sensitivity was however significant in a standard t-test ($t(105) = 1.88$, $p = .031$). With a sample size of 106, a one-tailed t-test is significant for observed effect sizes of 0.16 standard deviations or higher. In contrast, for our choice of a scale factor, a Bayes Factor is higher than 3 for observed standardized effect sizes of 0.26 standard deviations or higher. Effect sizes that fall between 0.16 and 0.26 are then significant in a t-test, with no conclusive evidence in a Bayes Factor analysis. A robustness region analysis revealed that no scale factor would have led to the conclusion that auROC2s for the two responses are different with $BF_{10} > 3$. See Supplementary Figure F.1 for a full Robustness Region plot (Dienes, 2019).

Hypothesis 3: A Bayes Factor analysis did not provide evidence for or against metacognitive asymmetry when controlling for response bias and sensitivity ($t(105) = -0.70$, $p = .759$; Cohen's d=0.07; $BF_{01} = 6.74$).

Hypothesis 4: In line with our hypothesis, response times for 'tilted' responses were faster than response times for 'vertical' responses, with a median difference of 68 ms. ($t(105) = -5.82$, $p < .001$; Cohen's d = 0.56; $BF_{10} = 1.83 \times 10^5$; see Figure 5.3, panel 3).

In summary, in Experiment 3 we found that 'tilted' (feature present) responses were faster and accompanied by higher subjective confidence that 'vertical' (feature absent) responses, with no difference in metacognitive sensitivity between the two responses.

5.6.4 Experiment 4: curved vs. straight lines

In Experiment 4, we looked at discrimination judgments between curved and vertical lines. Based on a search asymmetry for these stimuli [curved lines are found faster among vertical lines than vice versa; A. Treisman & Gormican (1988)], we hypothesized that a similar asymmetry would emerge in subjective confidence judgments, such that metacognitive sensitivity for perceiving a tilted line will be higher than for perceiving an vertical line. We used four dollar signs ($ $ $ $) as our mask.

211 participants (median reported age: 31; range: [18-76]) were recruited from Prolific for Experiment 4. Due to shorter than expected completion times in previous experiments, participants were paid £1.25, equivalent to an hourly wage of £6.

Median completion time was 12.08 minutes. Mean proportion correct was 0.84, and participants reported seeing a straight line on 44% of trials. In a deviation from our pre-registration, we excluded 18 participants for having zero variance in their confidence ratings for at least one of the two responses (see Section 5.5). Overall we excluded 104 participants, leaving 107 participants for the main analysis. Going forward, only data from included participants is analyzed.

Mean proportion correct among included participants was $M = 0.79$, 95% CI [0.77, 0.80]. The mean SOA of the last trial was $M = 28.01$, 95% CI [24.22, 31.79]. Participants showed a consistent bias toward reporting a curved rather than a vertical line ($M = 0.06$, 95% CI [0.04, 0.07]). On a scale of 0 to 1, mean confidence level was $M = 0.57$, 95% CI [0.53, 0.61]. Confidence was higher for correct than for incorrect responses ($M_d = 0.21$, 95% CI [0.18, 0.24], $t(106) = 14.96$, $p < .001$).

Hypothesis 1: In line with our hypothesis, confidence was generally higher for curved lines (feature present) responses than for straight lines (feature absent) responses ($M_d = 0.12$, 95% CI [0.09, ∞], $t(106) = 8.25$, $p < .001$; Cohen's d = 0.80; $BF_{10} = 1.61 \times 10^{10}$; see Figure 5.3, panel 4).

Hypothesis 2: Contrary to our prediction, auROC2 for reports of seeing a curved line ($M = 0.76$, 95% CI [0.73, 0.78]) was similar to auROC2 for reports of seeing a straight line ($M = 0.75$, 95% CI [0.73, 0.78]; $t(106) = 0.30$, $p = .382$; Cohen's d = 0.03; $BF_{01} = 8.23$; see Figure 5.4, panel 4.).

Hypothesis 3: (The lack of) metacognitive asymmetry was not different from what would be expected based on an equal-variance SDT model with the same response bias and sensitivity ($t(106) = -1.93$, $p = .972$; Cohen's d=0.19; $BF_{01} = 1.45$).

Hypothesis 4: In line with our hypothesis, response times for 'curved' responses were faster than response times for 'straight' responses, with a median difference of 51 ms ($t(106) = -4.36$, $p < .001$; Cohen's d = 0.42; $BF_{10} = 558.55$; see Figure 5.3, panel 4).

In summary, similar to Experiment 3, 'curved' (feature-present) responses were faster and accompanied by higher subjective confidence than 'straight' (feature absent) responses. However, similar to the results of Experiment 3, here also we did not find a metacognitive asymmetry for these stimuli.

5.6.5 Experiment 5: upward-tilted vs. downward-tilted cubes

In Experiment 5, we looked at discrimination judgments between upward-tilted and downward-tilted cubes. Based on a search asymmetry for these stimuli [upward-tilted cubes are found faster among downward-tilted cubes than vice versa, in line with an expectation to see objects on the ground and not floating in space; Von Grünau & Dubé (1994)], we hypothesized that a similar asymmetry would emerge in subjective confidence judgments, such that metacognitive sensitivity for perceiving an upward-tilted cube will be higher than for perceiving a downward-tilted cube. We used four dollar signs ($ $ $ $) as our mask.

162 participants (median reported age: 32; range: [18-69]) were recruited from Prolific for Experiment 5.

Median completion time was 13.30 minutes. Mean proportion correct was 0.79, and participants reported seeing a downward-tilted cube on 51% of trials. In a deviation from our pre-registration, we excluded 11 participants for having zero variance in their confidence ratings for at least one of the two responses (see Section 5.5). Overall we excluded 56 participants, leaving 106 participants for the main analysis. Going forward, only data from included participants is analyzed.

Mean proportion correct among included participants was $M = 0.77$, 95% CI [0.76, 0.78]. The mean SOA of the last trial was $M = 29.51$, 95% CI [23.20, 35.81]. Participants showed no consistent response bias ($M = -0.01$, 95% CI [−0.03, 0.00]). On a scale of 0 to 1, mean confidence level was $M = 0.55$, 95% CI [0.51, 0.59]. Confidence was higher for correct than for incorrect responses ($M_d = 0.23$, 95% CI [0.20, 0.26], $t(105) = 13.89$, $p < .001$).

Hypothesis 1: Contrary to our hypothesis, confidence was similar for upward-tilted (feature present) responses and downward-tilted (feature absent) responses ($M_d = 0.00$, 95% CI [−0.02, ∞], $t(105) = 0.12$, $p = .452$; Cohen's d = 0.01; $BF_{01} = 8.51$; see Figure 5.3, panel 5).

Hypothesis 2: Contrary to our hypothesis, a Bayes Factor analysis did not provide evidence for or against a difference in auROC2 for reports of seeing an upward-tilted cube ($M = 0.75$, 95% CI [0.73, 0.77]) and reports of seeing a downward-tilted cube ($M = 0.72$, 95% CI [0.70, 0.75]; Cohen's d = 0.22; $BF_{10} = 1.38$; see Figure 5.4, panel 5.). In contrast, a t-test revealed a significant metacognitive asymmetry, with higher metacognitive sensitivity for perceiving an upward-tilted (default-violating) cube ($t(105) = 2.29$, $p = .012$). See Supplementary Figure F.1 for a full Robustness Region plot (Dienes, 2019).

Hypothesis 3: (The lack of) metacognitive asymmetry was not different from what would be expected based on an equal-variance SDT model with the same response bias and sensitivity (Cohen's d=0.22; $BF_{10} = 1.28$). Here also, frequentist and Bayesian analyses conflicted, with a t-test revealing a significant metacogntiive advantage for upward-tilted (default violating) responses when controlling for bias ($t(105) = 2.25$, $p = .013$).

Hypothesis 4: Contrary to our hypothesis, response times for 'upward-tilted' responses were similar to response times for 'downward-tilted' responses with a median difference of 9 ms. ($t(105) = -0.82$, $p = .207$; Cohen's d = 0.08; $BF_{01} = 6.19$; see

Figure 5.3, panel 5).

In summary, in Experiment 5 we found no sign of processing asymmetry between upward and downward-tilted cubes in response-times and confidence. A significant metacognitive asymmetry was observed when using null-hypothesis significance testing, but was not supported by our Bayes Factor analysis. In accordance with our pre-registered plan to commit to the Bayes Factor analysis in interpreting the results, in what follows we interpret these findings as providing no support for a metacognitive asymmetry for upward and downward tilted cubes.

5.6.6 Experiment 6: flipped vs. normal letters

In Experiment 6, we looked at discrimination judgments between flipped and normal Ns. Based on a search asymmetry for these stimuli [flipped Ns are found faster among normal Ns than vice versa; U. Frith (1974); Wang, Cavanagh, & Green (1994)], we hypothesized that a similar asymmetry would emerge in subjective confidence judgments, such that metacognitive sensitivity for perceiving a flipped N will be higher than for perceiving a normal N. We used four dollar signs ($ $ $ $) as our mask.

127 participants (median reported age: 32; range: [18-65]) were recruited from Prolific for Experiment 6. Due to shorter than expected completion times in previous experiments, participants were paid £1.25, equivalent to an hourly wage of £6.

Median completion time was 12.76 minutes. Mean proportion correct was 0.74, and participants reported seeing a normal N on 50% of trials. In a deviation from our pre-registration, we excluded 4 participants for having zero variance in their confidence ratings for at least one of the two responses (see Section 5.5). Overall we excluded 21 participants, leaving 106 participants for the main analysis. Going forward, only data from included participants is analyzed.

Mean proportion correct among included participants was $M = 0.73$, 95% CI [0.72, 0.74]. The mean SOA in the last trial was $M = 37.26$, 95% CI [33.07, 41.46]. Participants showed no consistent response bias ($M = 0.00$, 95% CI [−0.02, 0.02]). On a scale of 0 to 1, mean confidence level was $M = 0.53$, 95% CI [0.49, 0.57]. Confidence was higher for correct than for incorrect responses ($M_d = 0.17$, 95% CI [0.15, 0.20], $t(105) = 16.45$, $p < .001$).

Hypothesis 1: Contrary to our hypothesis, confidence was *lower* for flipped (feature present) responses than for normal (feature absent) responses. This result was in the opposite direction to what we had expected, so was not significant in a one-tailed t-test ($M_d = -0.04$, 95% CI [−0.06, ∞], $t(105) = -3.32$, $p = .999$; Cohen's d = 0.32). However, a Bayes Factor favoured the alternative over the null ($BF_{10} = 18.92$; see Figure 5.3, panel 6).

Hypothesis 2: Contrary to our hypothesis, auROC2 for reports of seeing a flipped N ($M = 0.71$, 95% CI [0.69, 0.73]) was similar to auROC2 for reports of seeing a normal N ($M = 0.71$, 95% CI [0.69, 0.73]; $t(105) = 0.08$, $p = .468$; Cohen's d = 0.01; $BF_{01} = 8.54$; see Figure 5.4, panel 6.).

Hypothesis 3: (The lack of) metacognitive asymmetry was not different from what would be expected based on an equal-variance SDT model with the same response bias and sensitivity ($t(105) = 0.26$, $p = .396$; Cohen's d=0.03; $BF_{01} = 8.28$).

Hypothesis 4: Contrary to our hypothesis, response times for 'flipped' responses were *slower* than response times for 'normal' responses, with a median difference of 30 ms. ($t(105) = 2.81$, $p = .997$; Cohen's d = 0.27; $BF_{10} = 4.66$; see Figure 5.3, panel 6).

In summary, in Experiment 6 we found a difference in response speed and subjective confidence in the opposite direction to what we expected, with a processing advantage for the default-complying stimulus (N) compared to the default-violating stimulus (flipped N). We found no metacognitive asymmetry for these stimuli.

5.6.7 Experiments 1-6: summary

Overall, the pattern of results from Experiments 1-6 only partly matched our hypotheses in some cases, and stood in direct contrast to them in other cases (see fig. 5.5). A reliable metacognitive asymmetry was observed only in Experiment 2, and this asymmetry was in the opposite direction to what we had predicted, with a metacognitive advantage for O (feature absent) over C (feature present) responses. A metacognitive advantage for reporting Q over Os (Exp. 1) was not reliably above what is expected based on an equal-variance signal detection model.

For both local and global visual features (Experiments 1-4) we observed differences in mean confidence and response times that were consistent with our hypothesis of a processing advantage for the representation of the presence compared to the absence of visual features. In Experiments 5 and 6, we tested more abstract expectation violations. In Experiment 5, discrimination between upward-tilted and downward-tilted cubes showed no asymmetry in response time and confidence. In Experiment 6, participants were less confident and slower in their reports of seeing a flipped N, contrary to our prediction that default-violating signals should be easier to perceive. We found no evidence for or against a metacognitive asymmetry in either of the experiments.

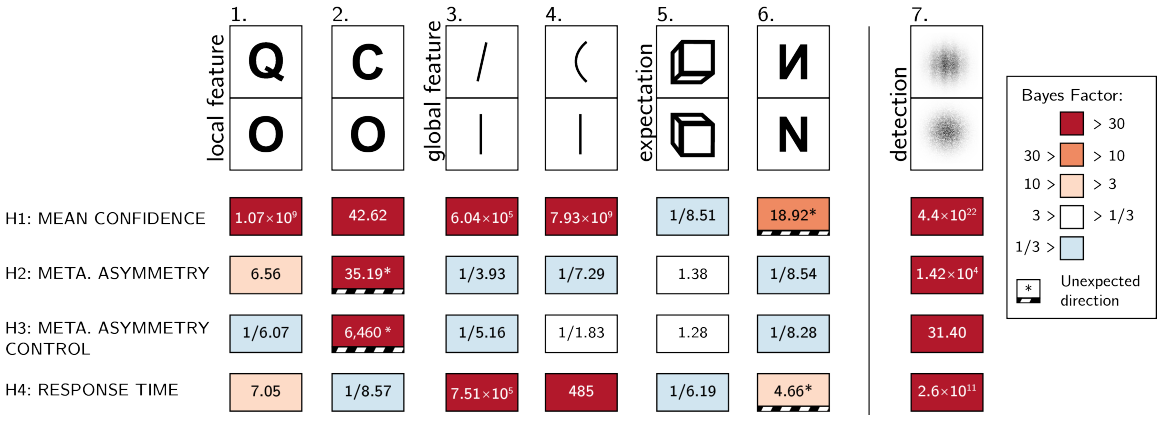


Figure 5.5: Summary of results from Experiments 1-6, and from the positive-control Experiment 7

5.6.8 Experiment 7 (exploratory): grating vs. noise

Results from Experiments 1-6 revealed that search asymmetry is not always accompanied by an asymmetry in metacognitive sensitivity. Given that we did not observe a true metacognitive asymmetry in the expected direction for any of our stimulus pairs, we were concerned that our experimental design may have been unsuitable for detecting classical metacognitive asymmetries in detection, for example due to an insufficient number of trials, the masking procedure, or the confidence report scheme. As a positive control, we collected data for an additional experiment that more closely resembled typical detection experiments. In this experiment, participants discriminated between two stimuli: random noise and a noisy grating (presented to participants as a 'zebra' stimulus; see Fig. 5.6). In Chapter 4, similar stimuli produced a robust metacognitive asymmetry between target absent (noise) and target present (noisy grating) responses (Mazor, Friston, & Fleming, 2020). We used black and white concentric circles as a mask. Apart from the choice of stimuli and mask, the procedure was identical to that of our pre-registered experiments.

141 participants (median reported age: 31; range: [21-39]) were recruited from Prolific for exploratory Experiment 7. For this positive control, all four hypotheses were fulfilled.

Median completion time was 10.70 minutes. Mean proportion correct was 0.73, and participants reported seeing a grating on 48% of trials. Overall we excluded 36 participants, leaving 105 participants for the main analysis. Going forward, only data from included participants is analyzed.

Mean proportion correct among included participants was $M = 0.76$, 95% CI [0.74, 0.77]. The mean SOA of the last trial was $M = 53.87$, 95% CI [38.85, 68.89]. Participants showed no consistent response bias ($M = 0.01$, 95% CI [0.00, 0.03]). On a scale of 0 to 1, mean confidence level was $M = 0.55$, 95% CI [0.51, 0.59]. Confidence was higher for correct than for incorrect responses ($M_d = 0.15$, 95% CI [0.13, 0.17], $t(104) = 12.58$, $p < .001$).

Hypothesis 1: In line with our hypothesis, confidence was higher for reports of target presence than for reports of target absence ($M_d = 0.20$, 95% CI [0.17, ∞], $t(104) = 14.07$, $p < .001$; Cohen's d = 1.37; $BF_{10} = 4.39 \times 10^{22}$; see Figure 5.6, right panel).

Hypothesis 2: In line with our hypothesis, auROC2 for reports of target presence ($M = 0.75$, 95% CI [0.73, 0.77]) was higher than for reports of target absence ($M = 0.68$, 95% CI [0.66, 0.70]; $t(104) = 5.20$, $p < .001$; Cohen's d = 0.51; $BF_{10} = 1.42 \times 10^4$; see Figure 5.6, left panel).

Hypothesis 3: In line with our hypothesis, this metacognitive asymmetry was stronger than what is expected based on an equal-variance SDT model with the same response bias and sensitivity ($t(104) = 3.49$, $p < .001$; Cohen's d=0.34; $BF_{10} = 31.40$).

Hypothesis 4: In line with our hypothesis, reports of target presence were faster than reports of target absence, with a median difference of 124 ms. ($t(104) = -8.84$, $p < .001$; Cohen's d = 0.86; $BF_{10} = 2.63 \times 10^{11}$; see Figure 5.6, right panel).

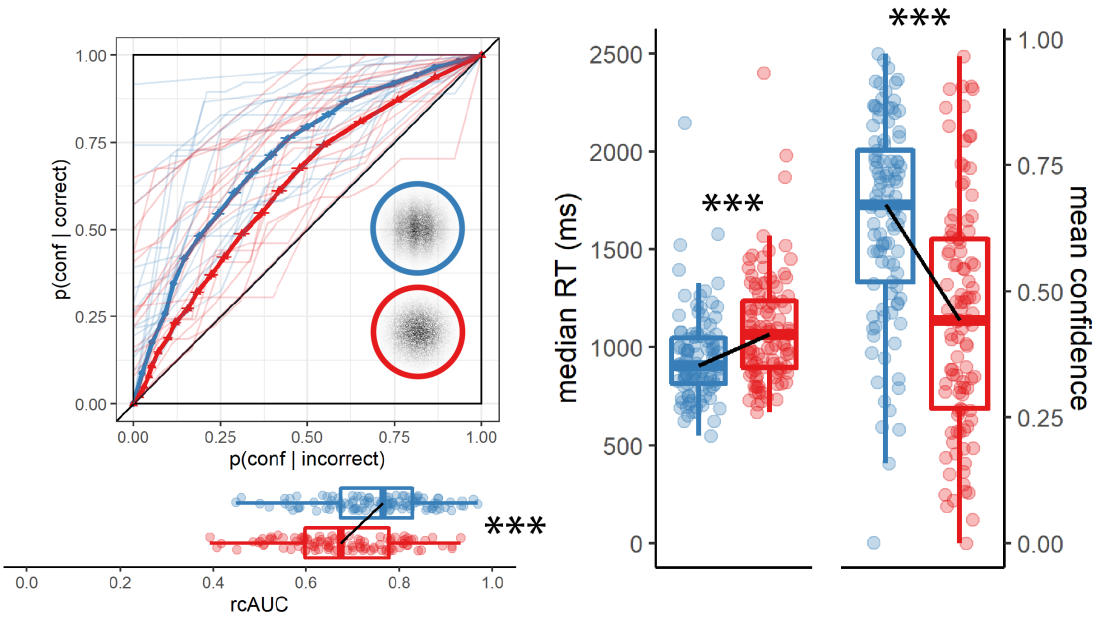


Figure 5.6: rcROC curves (left panel) and confidence and reaction time distributions (right panel) for Exp. 7 (detection positive control)

5.6.9 Exploratory analysis

zROC analysis

In a signal-detection framework, metacognitive asymmetry appears when the signal distribution has both a higher mean and higher variance than that of the noise distribution. This unequal variance setting produces higher metacognitive sensitivity for judgments of signal presence, compared to judgments of signal absence. A direct measure for the ratio between the variances of the two distributions is the slope of the *type-1 zROC curve*. A zROC curve is constructed by applying the inverse of the normal cumulative density function to false alarm and hit rates for different confidence thresholds. The slope of the zROC curve equals 1 exactly when the variance of the signal and noise distributions are equal. In detection experiments, the slope is often shallower than 1, indicating a wider signal distribution. Indeed, in our positive control experiment (Exp. 7), the median zROC slope was 0.86 and significantly shallower than 1 ($t(103) = -5.08$, $p < .001$ for a t-test on the log-slope against zero). Measuring the slope of the zROC curve in our six pre-registered experiments, we asked whether our 'feature-present' distributions had higher variance than our 'feature-absent' distributions. We used the standardized effect size obtained from Experiment 7 as a scaling factor for the prior distribution over effect sizes, reflecting a belief that a difference in slopes should be similar in magnitude to what is observed in a detection task.

zROC slopes were numerically shallower than one in Experiments 1 (Q vs. O; median slope = 0.95), 3 (line tilt; median slope = 0.94), 4 (line curvature; 0.97) and 5 (cube orientation; 0.95). This was significant only in Experiment 5 ($t(101) = -2.09$,

$p = .039$). In agreement with the results of our rcROC analysis, the zROC slope in Exp. 2 (C vs. O) was significantly higher than one, suggesting that the representation of the letter 'O' was more variable than that of the letter 'C' (median slope = 1.09; $t(104) = 2.29$, $p = .024$). A Bayes Factor analysis did not provide support for or against the null hypothesis for any of the six experiments (all Bayes Factors between 1/3 and 3).

Previous studies reported similar variance structures for these stimuli when presented in visual search arrays. For example, confidence in a vertical/tilted visual search task revealed higher variance in the representation of tilted (feature positive) compared to vertical (feature negative) stimuli (Vincent, 2011). Similarly, reverse correlation analysis revealed higher variance in the representation of Q (feature positive) compared to O (feature negative) stimuli (Saiki, 2008). Finally, and in agreement with our results, variance in the representation of O (feature negative) was found to be higher than in the representation of C (feature positive) (Dosher, Han, & Lu, 2004). Note that for the case of line tilt and Q vs. O, finding a high-variance target among low-variance distractors is easier than finding a low-variance target among high-variance distractors. However, the opposite is true for C vs. O, where a low-variance target (C) renders the search easier. This last observation challenges the suggestion that variance structure is the determining factor for visual search asymmetries (Dosher, Han, & Lu, 2004; Saiki, 2008; A. Treisman & Gormican, 1988; Vincent, 2011). #### Inter-subject correlations {-}

Across experiments, asymmetry in mean confidence (Hypothesis 1) and in response time (Hypothesis 4) were mostly aligned. This is consistent with previous reports of a negative correlation between response times and confidence across trials within participants (Calder-Travis, Charles, Bogacz, & Yeung, 2020; Henmon, 1911; Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010). To test if this was the case across participants too, and not only across experiments, we fitted a mixed-effects regression model to data from all seven experiments with experiment as a random effect ($\Delta RT \sim \Delta conf + (1 + \Delta conf|exp)$). The association between confidence and RT effects was significant in this model ($p < 0.001$; see Fig. 5.7; upper panel). In contrast, metacognitive asymmetry (difference between the area under the rcROC curves, controlling for response bias) was not significantly associated with asymmetry in either confidence ratings ($p = 0.41$; see Fig. 5.7; lower panel) or reaction time ($p = 0.54$).

5.7 Discussion

In perceptual detection, judgments about the presence or absence of a target stimulus differ in several ways. First, participants are more confident in stimulus presence than in stimulus absence (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018; e.g., Meuwese, Loon, Lamme, & Fahrenfort, 2014). Second, confidence ratings in judgments of stimulus presence are more aligned with objective accuracy (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018; Mazor, Friston, & Fleming, 2020; Meuwese, Loon, Lamme, & Fahrenfort, 2014). Finally, participants are faster to report stimulus presence

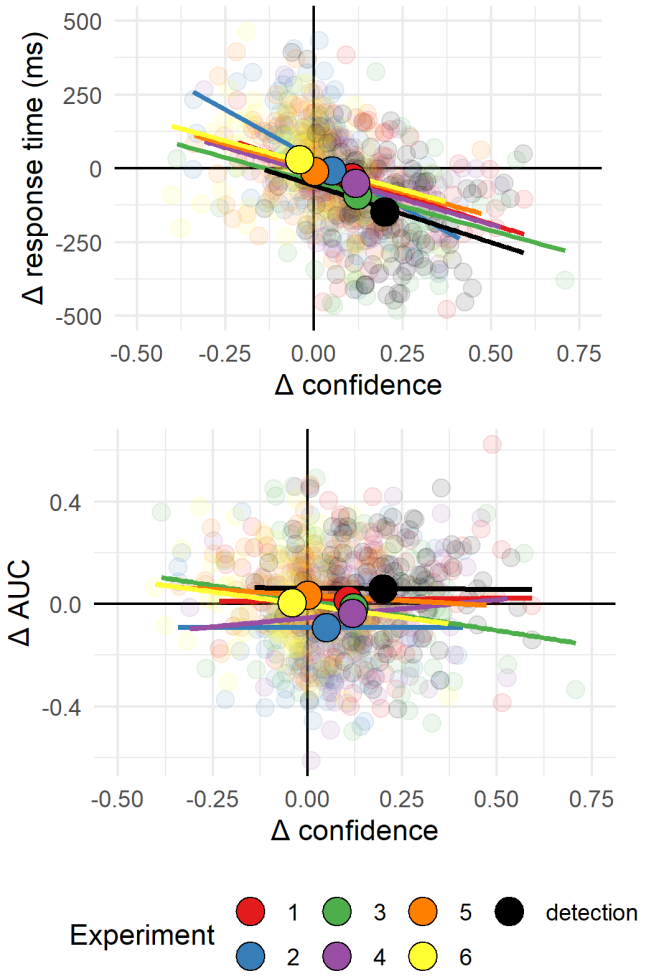


Figure 5.7: Upper panel: Difference in mean confidence between S1 and S2 responses plotted against differnce in mean response time between S1 and S2 responses across the seven experiments. Lower panel: Difference in mean confidence between S1 and S2 responses plotted against difference in metacognitive sensitivity, controlling for response bias, across the seven experiments. Semi-transparent circles represent individual subjects. Opaque circles are the means for each of the seven experiments, across participants. Lines indicate the best-fitting linear regression line for experiments 1-7.

(Mazor, Friston, & Fleming, 2020). In our positive control detection experiment (Experiment 7) we replicated these detection asymmetries. We found a mean difference of 20% confidence between decisions about the presence or absence of a grating, a metacognitive asymmetry of 0.07 in AUC units (ranging from 0 to 1), and a median difference of 124 milliseconds in response time between reports of target presence and absence.

In six pre-registered experiments, we focused on these three behavioural signatures

of decisions about the presence and absence of a stimulus, and asked whether they extend to discrimination tasks where stimuli are distinct in the presence or absence of sub-stimulus features such as the presence of an additional line in a letter, the curvature of a line, or more abstractly, the presence of a surprising default-violating signal. Our six stimulus pairs have been shown in previous studies to produce asymmetries in visual search, potentially reflecting differences in the processing of presences and absences of visual features, and of default-complying versus default-violating stimuli. If detection asymmetries also reflect differences in the abstract processing of presences and absences, or of default-complying versus default-violating sensory input, one would expect to find detection-like asymmetries in response time, confidence, and metacognitive sensitivity for discrimination between stimuli that produce asymmetries in a visual search task.

Starting from the end, Experiments 5 and 6 provide evidence against the proposal that asymmetries in confidence, reaction time and metacognitive sensitivity emerge for default-violating signals at all levels of representation. Stimulus pairs in Exp. 5 (cube orientation) and 6 (letter inversion) produced rcROC curves that were more consistent with the absence of metacognitive asymmetry than with our prior distribution on effect sizes (see section 5.2.4 for the specifics of our Bayesian hypothesis testing, including our prior on effect sizes). Given that these stimuli have been shown to produce reliable asymmetries in visual search (U. Frith, 1974; Malinowski & Hübner, 2001; Shen & Reingold, 2001; Von Grünau & Dubé, 1994; Wang, Cavanagh, & Green, 1994), we can safely conclude that not all default violations that produce an asymmetry in visual search also produce an asymmetry in metacognitive sensitivity.

Moreover, in Exp. 6, default-complying N responses were faster, and accompanied by higher levels of subjective confidence, than default-violating flipped-N responses. This is in contrast to our prediction of a processing advantage for default-violating signals, and in line with previous reports of a processing advantage for familiar over unfamiliar stimuli in the context of face perception and reading. For example, in a breaking continuous flash suppression (bCFS) paradigm, inverted faces took longer to break into awareness than upright faces (Stein & Peelen, 2021). A similar processing advantage for familiar stimuli has been documented for the perception of words (Albonico, Furubacke, Barton, & Oruc, 2018) and Chinese letters (Xue, Chen, Jin, & Dong, 2006). One possibility is that the perception of highly familiar stimuli such as letters and faces is supported by specific expert brain systems, affording a processing advantage beyond the general superior processing of default-violating signals (Xue, Chen, Jin, & Dong, 2006; Yovel & Kanwisher, 2005). Indeed, Exp. 6 was the only experiment in which we observed a processing advantage for familiar over unfamiliar stimuli.

Next, in Experiments 3 and 4 we looked at two features that have a global effect on stimulus appearance: tilt and curvature. Based on visual search asymmetries, A. Treisman & Gormican (1988) proposed that tilt and curvature are represented as positive features in the visual system. This takes us one step closer to typical detection experiments: participants now detect the presence or absence of a basic visual feature. In agreement with our Hypotheses 1 and 4, participants were more confident in identifying tilted and curved lines (mean differences of 0.12 and 0.12 on a

0-1 confidence scale), and were faster in giving these responses (mean differences of 67.67 and 50.57 ms). However, we did not find evidence for or against a metacognitive asymmetry for these global visual features. Our strongest candidate for a stimulus pair for which we expected to find a presence-absence asymmetry was *Q* vs. *O* (Exp. 1). The difference between these two letters is the presence of an additional line stroke: a concrete stimulus part that is localized in space and is independent of the rest of the stimulus. Theoretically, participants could approach this task as a detection task: ignore the common denominator (*O*) and focus on the presence or absence of the distinctive feature (‘,’). As we hypothesized, participants were more confident in their *Q* responses (mean difference of 0.11 on a 0-1 confidence scale). Participants were also faster in their *Q* responses (median difference of 37 ms). However, unlike stimulus-level detection, a small difference of 0.04 units in the area under the rcROC ROC curves was not different to what is expected based on a null SDT model.

Finally, In Experiment 2 we looked at discrimination between *C* and *O*s based on evidence from visual search that open edges are represented as a positive feature in the visual system (Takeda & Yagi, 2000; A. Treisman & Gormican, 1988; A. Treisman & Souther, 1985). As we hypothesized, *C* responses were accompanied by higher levels of subjective confidence (mean difference of 0.05 on a 0-1 confidence scale). However, in striking contrast to our original hypothesis, metacognitive sensitivity was *lower* for *C* responses (mean difference of 0.05 AUC units), even when controlling for response bias. This result strongly supports different underlying mechanisms behind search and metacognitive asymmetries. Furthermore, the results of Experiment 2 suggest distinct factors mediate the processing advantage for presence over absence (as reflected in shorter response times and higher confidence for *C* responses), and the metacognitive asymmetry between presence and absence (as reflected in improved metacognitive sensitivity for *O* responses).

C and *O* are unique in that the difference between them corresponds to two contrasting notions of presence and absence. On the one hand, *C* is marked by the presence of one additional feature - open edges (A. Treisman & Gormican, 1988; A. Treisman & Souther, 1985). On the other hand, it is marked by the absence of a piece: there is simply less of it relative to *O*. These two notions of presence and absence are typically coupled in detection. For example, the presence of a grating on a screen corresponds to the presence of additional features (such as orientation, contrast, and phase) as well as of more ‘visual stuff,’ relative to the blank background. A compelling interpretation of the results of Exp. 2 is that it is the presence or absence of visual features such as open edges that is driving the difference in confidence and response time, whereas a more quantitative notion of presence or absence (the amount of ‘visual stuff’ presented) is driving the metacognitive asymmetry between these two responses. We note however that based on this interpretation, we would expect a metacognitive sensitivity to operate also in Experiment 1, where *O* is missing a piece relative to *Q*. As described above, Experiment 1 provided no evidence for such a metacognitive asymmetry beyond what is expected from an equal-variance signal-detection model.

When interpreting our findings in a broader context, it is useful to note that in all six experiments we used backward masking for controlling the visibility level of our stimuli. Different visibility manipulations have been shown to affect detection

metacognitive sensitivity in different ways. For example, whereas metacognitive sensitivity in detection 'no' responses is at chance when backward masking is used, it is significantly higher than chance when the attentional blink is used to control stimulus visibility (Kanai, Walsh, & Tseng, 2010). Similarly, phase scrambling but not attentional blink produces a metacognitive advantage for 'yes' responses (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018). While our positive control (Exp. 7) produced a reliable metacognitive asymmetry between judgments of target presence and absence, it was also the only experiment where stimulus visibility was controlled with low contrast, in addition to backward masking (for the purpose of compatibility with previous experiments; see Fig. 5.6). Based on our findings alone, we cannot rule out the possibility that using other visibility manipulations may reveal metacognitive asymmetries for the presence or absence of abstract default violations. Furthermore, it is possible that some of the observed asymmetries for low-level features may reflect asymmetries in the joint perception of target stimulus and backward mask, rather than in the perception of the target stimulus by itself (Jannati & Di Lollo, 2012; Kahneman, 1968).

Together, our findings weigh against our original proposal that metacognitive asymmetries in perceptual detection are a signature of higher-order default reasoning. Unlike search asymmetries that extend to abstract levels of representations such as familiarity (Wang, Cavanagh, & Green, 1994; J. M. Wolfe, 2001) and even social features such as ethnicity and gender (Gandolfo & Downing, 2020; Levin & Angelone, 2001), metacognitive asymmetries in visual discrimination are grounded in concrete visual processing. Furthermore, we provide evidence for a dissociation between asymmetries in metacognition and in response time and confidence, where the latter is linked to activation of basic feature-detectors, for example of orientation, open ends, or curvature.

5.8 Conclusion

In a set of six experiments, we sought to test the proposal that a metacognitive asymmetry between the representation of stimulus presence and absence is one instance of a more general asymmetry between the representation of default states and default-violating surprises. Our findings provide evidence against this idea. A metacognitive asymmetry was not observed in near-threshold discrimination between stimulus pairs that differ in their alignment with default expectations. This was the case even in pairs that showed substantial asymmetries in response time and overall confidence levels. Results from our pre-registered set of analyses are most consistent with a narrow interpretation of the presence/absence metacognitive asymmetry in visual detection, that is limited to concrete, spatially localized presences. Furthermore, a metacognitive asymmetry between *C*s and *O*s in the opposite direction to what is observed in visual search indicates that different cognitive and perceptual mechanisms underlie these two apparently similar phenomena.

General Discussion

In this thesis I investigated inference about absence in visual perception, and its relation with self-modeling and default-mode reasoning. In chapters 1 and 2 I focused on visual search, and asked what people know about their visual search behaviour, and how this knowledge related to their ability to efficiently terminate a search in the absence of a target. In Chapter 3 I used reverse correlation to ask what information is incorporated into confidence judgments in decisions about the presence and absence of a stimulus. Then, in chapter 4 I used functional imaging to compare the neural processes governing metacognitive evaluation of decisions about stimulus type and stimulus presence or absence. Finally, in chapter 5 I borrowed ideas from the visual search literature to ask at what cognitive level does the metacognitive asymmetry between judgments of presence and absence emerge.

In what follows I evaluate my original proposal, that inference about absence critically relies on self-knowledge, in light of my findings. Specifically, I list observations that don't fit with this idea. Before concluding, I critically review two approaches to inference about absence that obviate the need in self-modelling, and briefly describe two directions for future research that build on and extend my work here.

What I didn't find

The theoretical proposal put forward in this thesis is that inference about absence is unique in that it requires relying on a self-model. In previous chapters I tried to make sense of my data in light of this proposal. However, some patterns that I expected to find were missing in the data, and some patterns that I did find were difficult to reconcile with this overarching idea. In the following, and following Charles Darwin's advice to make an effort to remember observations that don't fit with one's theory (Darwin & Darwin, 1958, p. 123), I list some things that I thought I should find, but didn't.

Chapter 1: no correlation with explicit metacognition

In Chapter 1, I show that participants can immediately recognize the absence of a salient target in a display, and that they can do so even before having experience with the task. My interpretation of the results abscribes this 'absence pop-out' to pre-existing metacognitive knowledge of the parallel nature of feature search. I further show that even participants whose explicit metacognitive reports show no insight

into the parallel nature of feature search exhibit the same pattern of search time in target-absent trials - a finding that we interpret as indicating a dissociation between explicit and implicit metacognitive knowledge.

A much stronger support for the proposal I put forward here would be a correlation between explicit metacognitive knowledge of search efficiency and search slopes in target-absent trials, such that participants who rate feature searches as easier also quit them earlier in the absence of a target. Instead, I found no correlation between explicit metacognition and behaviour on target-absent trials. A full dissociation between implicit and explicit metacognition is one possible interpretation for my results. An alternative interpretation is that the immediate recognition of target absence in these first trials is not at all dependent on metacognitive knowledge - explicit or implicit. In the discussion of Chapter 1 I list some alternative accounts, such as immediate recognition of absence via ensemble perception, and explain how they depend on implicit metacognitive knowledge, for example in the form of a firing threshold on neurons in the visual pathway. One concern is that a notion of self-modelling that encompasses the firing thresholds of visual neurons is too permissive to be scientifically useful. The visual system translates incoming sensory signals into beliefs about the external world. To do that, neurons in associative visual areas implicitly represent a likelihood function going from world states into firing patterns, and invert this function to approximate a representation of the world state, given observed firing patterns in primary visual areas. This likelihood function is a form of self-knowledge, in the sense that it is knowledge about how the brain responds to incoming signals. But its importance is not specific to the representation of absence (although representations of absence may be more sensitive to these internal models compared with other decisions).

In the pre-registration documents for Experiments 1 and 2 I focused on a contrast between blocks 1 and 3 (before and after experience with target-present trials), with the hypothesis that task experience will affect the search slope for target-absent trials. A learning effect between blocks 1 and 3 would have allowed me to further ask how generalizable this new knowledge is, and for how long is it retained in the system, giving us a better hold of this mental self-model. In a later section I expand on how focusing on model failures (such as mismatches between target-present and target-absent search efficiency) can be useful for understanding the structure of the mental self model.

Chapter 3: no effect of confidence in signal presence

In Chapter 3, I asked what drives confidence in decisions about target absence. Since decisions about target absence are based on the absence of perceptual evidence, I hypothesized that subjective confidence in such decisions may rely on other factors. For example, if participants are using a counterfactual heuristic, their confidence in previous 'target present' trials may inform their confidence in 'target absent' decisions ("When a target was present it was highly visible, so I would have seen the target if it were present"). This effect should be stronger than the effect of previous confidence in target-absence decisions on confidence in presence, because confidence in presence can be based on perceptual evidence. To test this, I looked at the correlation

between confidence in absence and confidence in the last target-presence decision. This correlation was not significantly different from 0 in Experiment 1 ($t(9) = 0.86$, $p = .412$). In Experiment 2, this correlation was significantly higher than 0 at the group level ($M = 0.17$, 95% CI [0.10, 0.23], $t(100) = 5.37$, $p < .001$), but a similarly high correlation between confidence in 'yes' responses and in the last decision about target absence suggests that this was not specific to decisions about absence ($\Delta M = 0.01$, 95% CI [−0.07, 0.09], $t(197.32) = 0.25$, $p = .803$).

Similarly, a counterfactual heuristic predicts lower confidence in absence when local target prevalence (for example, the number of targets presented in the last 5 trials) is low (Hsu, Horng, Griffiths, & Chater, 2017). To see this, compare two participants that are presented with random noise: one that hasn't seen a target for 4 trials in a row, and one that just saw a target in the previous trial. The first participant may doubt their perceptual sensitivity and give a low confidence rating, but the second can be more confident that they would not have missed a target. Contrary to this prediction, local target prevalence had no effect on confidence in 'no' responses in Experiment 2 (quantified as the distance in number of trials from the last encounter with a target; $t(100) = -1.22$, $p = .224$). In Experiment 1, a significant correlation in the opposite direction was observed, such that participants were more confident in their 'no' responses when a target hasn't been observed for a longer series of trials ($t(9) = 3.11$, $p = .013$). Overall, we found no evidence for higher susceptibility of confidence in absence to confidence and stimulus prevalence in previous trials.

Chapter 4: only minor differences in brain activity between inference about absence and presence

In Chapter 4, I compared brain activity in discrimination and detection. Within detection, we compared decisions about signal presence and absence. Our pre-registration document largely focused on the behavioural differences between 'yes' and 'no' responses, and their possible neural underpinning, with a focus on the lateral prefrontal cortex and regions that have been associated with counterfactual reasoning ["I would have seen it if it were there"; Boorman, Behrens, Woolrich, & Rushworth (2009)]. To my surprise, I found no significant difference in overall activity between 'yes' and 'no' responses (for an uncorrected map, see here). A significant difference in the parametric modulation of confidence was found not in our pre-defined regions of interests, but in the right temporo-parietal junction (rTPJ).

Given the reliable behavioural differences in reaction time, overall confidence, and metacognitive sensitivity, a similar profile of BOLD activation for 'yes' and 'no' responses was unexpected. In language comprehension, for example, the processing of negation shows distinct neurobiological markers that are overlapping with those of response inhibition (Papeo & Vega, 2020). In visual search, the right lateral prefrontal cortex was more engaged in target-absent trials (Vallesi, 2014), and the right temporo-parietal junction showed differential activation in visual search hit and miss trials (Shulman, Astafiev, McAvoy, d'Avossa, & Corbetta, 2007). Surprisingly, however, despite the robust behavioural differences, BOLD activations for 'yes' and 'no' responses

gave rise to indistinguishable baseline activation, differing only in the modulation of confidence. When focusing on the frontopolar cortex, confidence modulation was different between detection and discrimination, but remarkably similar for detection 'yes' and 'no' responses. Again, this is in contrast to my original hypothesis, that counterfactual reasoning should play a major role in decisions about target absence, more so than in decisions about target presence.

Chapter 5: no metacognitive asymmetry between default-complying and default-violating signals

In Chapter 5, I focused on three behavioural asymmetries between detection 'yes' (stimulus present) and 'no' (stimulus absent) responses: in reaction time, global confidence, and metacognitive sensitivity. Using stimulus pairs that generate asymmetries in visual search, I asked whether detection-like asymmetries would emerge in discrimination tasks that can be described as the detection of sub-stimulus presences: local stimulus features (such as the line that distinguishes a *Q* from an *O*), global features (such as the presence or absence of curvature), and expectation violations (such as the presence or absence of letter inversion). My reasoning was the following: if presence/absence asymmetries emerge because participants assume absence as default unless they have evidence for presence, similar asymmetries should emerge for other things that we take as default (e.g., letters are not mirrored, objects are not floating in space). This default-reasoning framework also provided a conceptual link to metacognitive asymmetries in recognition memory: there also, participants assume as default that an item is new (i.e., hasn't been presented in the study phase), unless they have evidence for that it is old.

I found no evidence for a metacognitive asymmetry between default-complying and default-violating signals. Furthermore, in Experiment 6 participants were *slower* and gave *lower* confidence ratings when they reported seeing a flipped letter: a significant finding that stood in direct contrast to the default-reasoning proposal. We interpreted our findings as placing the metacognitive asymmetry for detection judgments at lower levels of the cognitive hierarchy, potentially in early visual processing.

A similar behavioural profile for the identification of default-complying and default violating stimuli does not stand in contrast to the proposal that inference about absence does involve a default-reasoning component, and that as a result it requires reliance on a mental self model (see Section 0.2.1). Nevertheless, finding a metacognitive asymmetry for expectation violations would have provided strong support for this framework, which we did not get from our findings.

Inference about absence without self-modelling

In this thesis, I focused on the role of self-modelling in inference about absence. My investigation was guided by a conceptual analysis, based on default-reasoning (Section 0.2). In Chapter 1 I also considered alternative accounts of inference about absence in visual search, where decisions about the absence of a target object are guided not by

counterfactual reasoning based on a self-model, but by a model-free heuristic based on success in previous trials, or by an immediate perception of ensemble statistics of a visual scene. In the following, I describe two approaches to inference about absence that do not involve self-modelling: patch-leaving heuristics in foraging, and philosophical accounts of absence perception. I unpack some of their strengths and limitations.

Patch-leaving in foraging

Chapter 4 opens with a foraging example: an agent deciding whether a bush bears ripe fruit or not. In this example, detecting berries is an instance of inference about presence, and deciding that a bush bears no fruit is an instance of inference about absence. Since evidence can only be available for the presence but not for the absence of berries, decisions about the absence of berries must rely on some form of counterfactual reasoning ("I would be seeing the berries if they were present"), that in turn relies on a self-model. However, when considering the behaviour of foragers, explicitly deciding that a bush bears no ripe fruit is mostly unnecessary. Instead, a decision to move to the next bush can be motivated by an explicit or implicit belief that leaving the current bush will be more rewarding than staying.

Heuristics for approximating when is the optimal time to leave a patch in search for other sources of food have been formalised and tested against animal behaviour. For example, in Charnov's *Marginal Value Theorem* [MVT; Charnov (1976)], a decision to leave the current patch (a spatially defined source of food, like a bush of berries) is optimal when the instantanous rate of return (e.g., berries found per minute) falls below the mean rate for the environment as a whole. Thus, MVS predicts that agents would exploit patches for longer under conditions in which patches yield less returns on average, or in which patches are physically farther away from each other. Foraging behaviour that is consistent with these qualitative predictions has been observed in birds (R. J. Cowie, 1977; Krebs, Ryan, & Charnov, 1974), armadillos and guinea pigs (Cassini, Kacelnik, & Segura, 1990). Similarly, when searching for an unspecified number of visual targets in an array (e.g., gas stations in satellite images), online participants set their giving up times in accordance with MVT (Ehinger & Wolfe, 2016).

As MVS shows, a decision to move on to the next bush can be made without a self-model (or any model other than knowledge of basic properties of the environment). Importantly, however, it does not show that absence can be inferred without a self-model, but that patch-leaving is not necessarily an instance of inference about absence. This is because search in natural foraging tasks is not exhaustive: the task is rarely to find *all* berries on a bush, but to find as many berries as possible, considering the cost of search itself in energy and exposure to threats.

In ecological settings outside of controlled experiments, instances of exhaustive search are usually ones where the cost of missing a target are considerable, such as when scanning a lake for predators before approaching to drink, or checking a memograpm for potential indicators of a tumor. In these cases, basing decisions on inference about absence is crucial, rendering MVS-like approaches dangerous. Still,

patch-leaving algorithms reveal that for many behavioural functions, including foraging for food, strict inference about absence is not necessary.

Direct perception

According to some contemporary philosophers absence need not be inferred because it is directly perceived. For example, philosopher Anna Farennikova explains the perception of absence as a perception of a mismatch between sensory input and expectations of presence: "The phenomenology of absence is the experience of incongruity" (Farennikova, 2013, 2015). Farennikova presents the following example of absence perception:

> "You've been working on your laptop in the cafe for a few hours and have decided to take a break. You step outside, leaving your laptop temporarily unattended on the table. After a few minutes, you walk back inside. Your eyes fall upon the table. The laptop is gone! This experience has striking phenomenology. You do not infer that the laptop is missing through reasoning; you have an immediate impression of its absence."

According to this account, the absence of a laptop is directly perceived, instantaneously and without any conscious effort, as a mismatch of sensory input relative to a perceptual template of a laptop on a table. This seems to contrast with the account presented here in several ways.

First, according to this account, absence is perceived, whereas in the account I defend it is inferred. On closer inspection, this is not in fact a point of disagreement. Perception is widely held to involve, and depend on, inference from noisy sensory data about unknown world states (Friston, 2010; Gershman, Vul, & Tenenbaum, 2012; Helmholtz, 1948). Therefore, that absence is inferred does not mean that is cannot also be perceived. Indeed, Gow (2021) proposes that absence is perceived via "intellectual seeming": a form of inference that results not in beliefs or judgments, but in perceptual states.

The next point of potential disagreement concerns what knowledge is necessary to infer absence. According to the template-mismatch account, any sensory mismatch relative to an expected template immediately results in a perception of, or inference about, absence. In the account defended here, absence can only be inferred when one believes that they would have perceived the missing object if it were present. Consider, for example, returning from a break and finding a waiter occluding some of the table. As in Farennikova's example, the sensory input is not consistent with your expectation to find your laptop on the table, but this time you are not inferring that it is absent, because you know that the waiter might be occluding it. Similarly, if you believe the laptop would be difficult to see (for example, if your forgot your glasses inside), you will not infer absence until you check the table more closely. In both cases, inference about absence depends on much more than a comparison to a sensory template: it depends on sophisticated inference based on sensory and metacognitive cues. In support of this more elaborate account of inference about absence, in Chapter

1 I show that participants take longer to infer absence in displays that make finding the target more difficult.

In defense of a template-mismatch account, one may argue that the difference between seeing the absence of a laptop in Farennikova's example and not seeing it in my occluding-waiter or missing-glasses variants is not in post-perceptual inferences, but in the sensory templates against which the sensory input is compared. For example, my sensory template of a laptop on a table may itself become less clear when I know the lighting has changes. Critically, this flexible updating of sensory templates based on changing environmental and internal conditions is a model-based process, one that involves not only modelling of objects and other agents, but of my own perception and attention too.

Finally, in support of the template-mismatch account, Farennikova mentions that many experiences of absence feel instantaneous and lacking in conscious effort, indicating some automaticity of absence processing. However, introspection can be misleading. Using different tasks and stimuli, in Chapters 1, 4, 3 and 5 I show that inference about absence is significantly slower than inference about presence or stimulus type, even when controlling for response requirements (Chapter 3), and when presenting the decision as a discrimination task between two stimuli (Chapter 5, Exp. 7). The difference in response times between inferences about presence and absence ranged from 46 ms in Chapter 4 to 124 ms in Chapter 5. These are neurally and psychologically significant differences, that are comparable in size to congruency effects in Stroop Flanker tasks, and to perceptual priming effects (Semmelmann & Weigelt, 2017). This strongly suggest that, at least in the context of a detection task, inference about absence is slower than inference about presence.

To conclude, a template-mismatch account of inference about absence as the one put forward by Farennikova (2013) either includes implicit self- and world-modelling in the generation of context-sensitive templates, or fails to account for the flexibility with which subjects infer absence in dynamic environments and internal conditions.

Future directions

As the list in the section "what I didn't find" makes clear, the investigation of a link between inference about absence and self-modeling is far from complete. The data are telling us that there is more to the story than default reasoning and reliance on self knowledge, or that my particular formulation of these concepts is lacking. More work is needed to further investigate the mechanisms that allow humans to form explicit representation of absence based on the absence of evidence, and how this relates to their generative models of their own perception and cognition. In the following, I list two avenues for future studies: leveraging failures of a self-model, and using inference about absence in more naturalistic settings.

Failures of a self-model

The structure of models is best revealed when they fail to faithfully represent their object. For example, a failure of the body-schema to correctly identify the position of one's arm following synchronous touch (Botvinick & Cohen, 1998) has advanced our understanding of how humans represent their own bodies, and how they update these representations based on sensory evidence (D. Cowie, Makin, & Bremner, 2013; Kammers, Vignemont, Verhagen, & Dijkerman, 2009; Tsakiris & Haggard, 2005). Systematic errors in participants' predictions of the physical effects of collisions informed theories of people's intuitive understanding of physics (A. N. Sanborn, Mansinghka, & Griffiths, 2013). Lastly, children's failure to ascribe a false-belief to an agent provided cognitive scientists with an experimental handle on the development of a Theory of Mind between the ages of 2.5 and 4 (Gopnik & Wellman, 1992).

Similarly, a scientific investigation of the mental self model can make use of cases in which this model fails to accurately represent the mental self. In the introduction, I provided two examples for misrepresentations that were revealed by suboptimal inference about absence: in near-threshold detection, participants overestimate the effect of eccentricity on perceptual sensitivity (Odegaard, Chang, Lau, & Cheung, 2018; Solovey, Graney, & Lau, 2015), and in visual search participants fail to fully represent the search advantage for unfamiliar targets [Wang, Cavanagh, & Green (1994); Zhang & Onyper (2020); but see Chapter 2 for evidence that intuitive theories of visual search are sensitive to this advantage, at least to some extent]. Focusing on these mismatches between the mental self (including perception, attention, and higher cognition) and the mental self-model has the potential to uncover the boundaries of the self model, the simplifications it makes, and its computational building blocks.

For example, large-scale online data collection with adaptive (Cavagnaro, Pitt, & Myung, 2011; He, Chen, & Li, 2020) and sequential designs (Hsu, Martin, Sanborn, & Griffiths, 2019; Langlois, Jacoby, Suchow, & Griffiths, 2021; A. Sanborn & Griffiths, 2008) now affords to identify stimulus features that affect visual search in target-present trials more than in target-absent trials or vice versa, indicating a mismatch between visual attention and participants' model of their own attention. Beside directly contributing to our knowledge of the contents and structure of the mental self model (e.g., the mental self model is better calibrated for basic visual features than for experience-based ones), this systematic mapping of model misrepresentations can then be used to ask how is the mental self model expanded and adjusted based on experience, by measuring how these biases change in light of task experience. Collecting data from multiple subjects on multiple test items has been successful in investigating factors that contribute to image memorability, to subjective memorability scores, and to mismatches between the two (Isola, Xiao, Parikh, Torralba, & Oliva, 2013; Rust & Mehrpour, 2020).

Inference about asbence in multi-dimensional and hierarchical representational spaces

in order to achieve high levels of experimental control, experiments in this thesis have mostly used simple stimuli: random dot kinematograms, visual gratings, flickering patches, and simple geometrical shapes. Using low-level visual stimuli has made is possible to precisely control the input to participants' perceptual system (e.g., in the form of signal to noise ratio). However, restricting our focus to the representation of presence and absence in low-dimensional representational spaces has potentially masked some crucial properties of inference about absence that are revealed in high-dimensional, hierarchically structured representational spaces (see Section 0.2.1 in the introduction).

Consider, for example, the results of the imaging experiment in Chapter 4, where to our surprise we found no univariate difference in activation between detection 'yes' and 'no' responses. One explanation is that our limited stimulus set (noise and right- and left-tilted gratings embedded in noise) has allowed subjects to form active representations of stimulus absence (in this case, noise), bypassing the need for counterfactual reasoning for inferring absence. In Chapter 3, Experiment 2, reverse correlation analysis revealed an active accumulation of evidence for absence (in the form of overall darkness).

In future studies, using high-dimensional stimuli such as photographs of animals (Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018), faces, or words (Kay & Yeatman, 2017) may render it impossible to accumulate evidence for absence, instead pushing participants to adopt a counterfacutal reasoning heuristic. Combinatorically, the number of all possible stimuli is exponential in the number of dimensions, making an exhaustive search impossible in high-dimensional stimulus spaces. In such highly asymmetric spaces, a counterfactual heuristic (would I have seen a target stimulus if it were present?) may be more advantageous.

Conclusion

In five studies I investigated inference about absence, self-modeling, and the relation between the two. Using visual search and perceptual detection and discrimination, I asked what separates decisions about the presence of a signal from decisions about signal absence, and to what extent is the difference between these two types of decisions related to the reliance of the latter on counterfactual reasoning on the basis of a self-model. Overall, I observed mixed results for and against this proposal. In visual search, participants' rich and accurate explicit theory of their own visual search behaviour played no role in deciding that a target was absent from a display. In near-threshold detection, a parametric modulation of confidence on brain activation was similar for decisions about the presence and the absence of a stimulus, except for in the right temporoparietal junction - a brain region that has been associated with monitoring one's own attention, as well as the attention of other agents. Finally, discrimination tasks with stimuli that varied in the presence or absence of a local

feature, a global feature, or a default violation, allowed a dissociation of the factors that independently contribute to behavioural differences between decisions about the presence and absence of a stimulus. Overall, my findings suggest that decisions about presence and absence differ in more than one way. Self-modeling and counterfactual thinking may account for some of these differences, but not for all of them.

Appendix A

Signal Detection Theory

"Signal Detection Theory" is a conceptual framework for the description of decision making between two alternatives in the presence of uncertainty. Examples include deciding whether a presented word has been studied before or not, to which of two groups does a noisy stimulus belong, or whether a stimulus was presented on the screen or not (Stanislaw & Todorov, 1999; Tanner Jr & Swets, 1954). Under this framework, on each experimental trial a "decision variable" is sampled from one of two distributions. I will refer to these distributions here as the *signal* and *noise* distributions, although depending on context they can have different labels, such as *old* and *new* distributions in recognition memory task or *right* and *left* in a movement discrimination task. On trials in which the decision variable exceeds a criterion c, a 'yes' response is executed, otherwise a 'no' response is executed (see Fig. A.1).

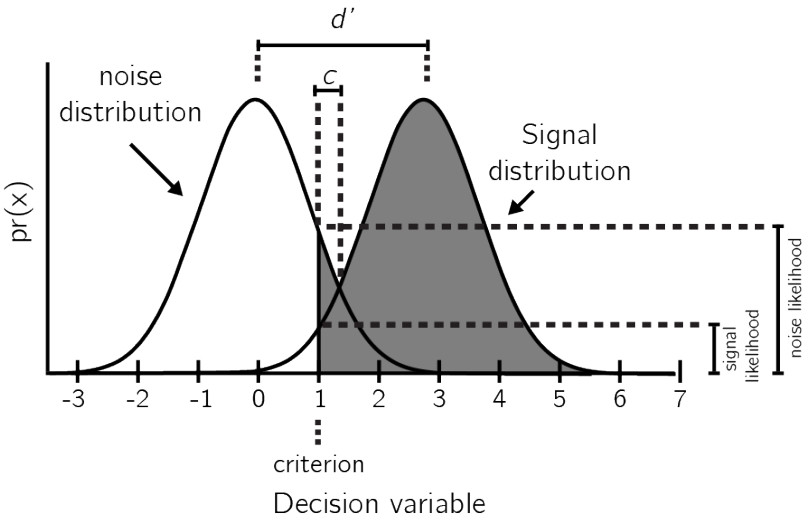


Figure A.1: Distribution of the decision variable across noise and signal trials, showing d', c, and the likelihoods. Figure based on Stanislaw & Todorov, 1999.

Given the noisiness of the incoming input, some signal trials will result in a 'no' response and some noise trials will result in a 'yes' response. This makes a total of

four groups of trials that can be ordered in a two by two table:

Table A.1: SDT response classification.

response	signal	noise
'yes'	hit	false alarm
'no'	miss	correct rejection

Two conditional probabilities are sufficient to provide a full description of the behaviour profile of a participant, namely $p(yes|Signal)$ (the 'hit rate'), and $p(yes|Noise)$ (the 'false alarm rate'). SDT makes it possible to translate these two probabilities to properties of the signal and noise distributions and their positioning with respect to the decision criterion. The parameter d' represents the distance between the two distributions in standard deviations. Under the assumption of equal variance of the two distributions d' can be approximated as $\hat{d}' = Z(h) - Z(f)$, with Z representing the inverse cumulative normal distribution. The parameter λ stands for the position of the criterion relative to the mean of the noise distribution, and can be approximated as $\hat{\lambda} = -Z(f)$.

A.1 ROC and zROC curves

The false alarm and hit rates are often insufficient to provide a full description of a system. For example, they are not sufficient to determine the ratio between the variance terms of the two distributions, and therefore to decide if the equal variance assumption holds. To obtain a fuller picture, false alarm and hit rates can be recorded under different settings of the decision criterion. One way to experimentally shift the criterion is by manipulation of the task incentive structure. For example, in order to encourage participants to make more 'no' responses, rewards for correct rejections can be set higher than rewards for hits. Alternatively, confidence ratings can be collected for every decision. The criterion can then be theoretically placed between every two possible confidence ratings, to generate a full set of false positive and hit rates.

A "*Receiver Operating Characteristic*" (ROC) curve is the plot of false alarm and hit rates for all possible settings of a decision criterion value. It can be approximated by plotting the false alarm and hit rates for the criterion values available by the experimental manipulation (see figure A.2). For a system that performs at chance, false positive and hit rates should be equal for every criterion, giving rise to an ROC that follows the identity line. The area under the ROC curve ("AUROC") can be interpreted as the proportion of times the system will identify the stimulus in a 2AFC task where noise and signal are presented simultaneously (Stanislaw & Todorov, 1999).

Often it is informative to plot the inverse of the cumulative distribution for $p(f)$ and $p(h)$, resulting in what is known as a "zROC curve" (see figure A.3). The zROC curve is linear when the noise and signal distributions are approximately normal. The slope of the zROC curve equals the ratio between the standard deviations of

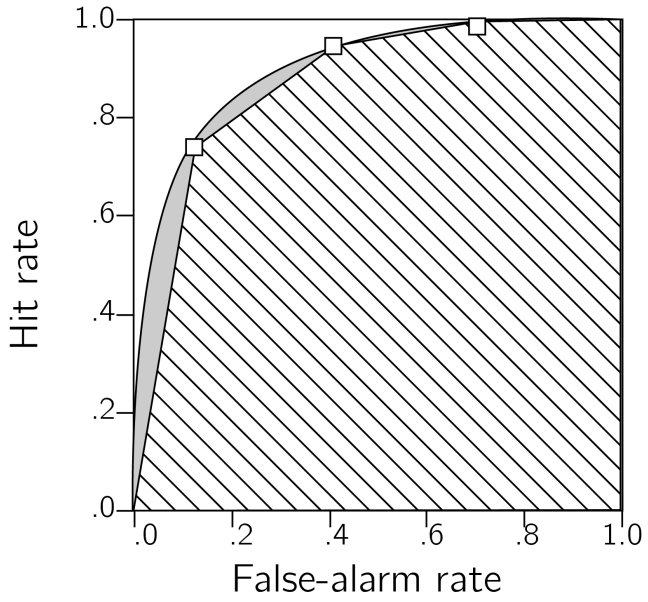


Figure A.2: Receiver Operating Characteristic (ROC) curve. Three points on the ROC curve are shown (open squares). The area under the curve, as estimated by linear extrapolation, is indicating by hatching; the true area includes the gray regions. Figure based on Stanislaw & Todorov, 1999.

the noise and signal distributions (Stanislaw & Todorov, 1999). Hence, the standard equal-variance SDT model predicts a linear zROC curve with a slope of 1.

A.2 Unequal-variance (uv) SDT

Unequal variance (uv) SDT can be applied to settings in which one distribution is assumed to be wider. For example, in perceptual detection tasks it is plausible that the signal distribution will be wider, as every sample comprises two sources of variance: a baseline noise component that is shared with the noise distribution, and the stimulus noise that represents fluctuations in the evidence strength available in the physical stimulus. A similar pattern is typically observed in recognition memory tasks.

This simple change to the model has profound effects on the decision making process. Under the assumption of equal-variance, the "log likelihood-ratio" (LLR; $log(\frac{p(x|signal)}{p(x|noise)})$) increases monotonically as a function of the decision variable, so that an optimal solution to the inference problem can rely on one decision criterion: samples to the right of the criterion are labeled as 'signal,' and samples to its left are labeled as 'noise' (Wickens, 2002, p. 30). The introduction of unequal variance to the SDT model makes inference more complex. Both extreme positive and extreme negative values are more likely to be drawn from the signal distribution when it is wider than the noise distribution, making a single-criterion decision rule sub-optimal. More

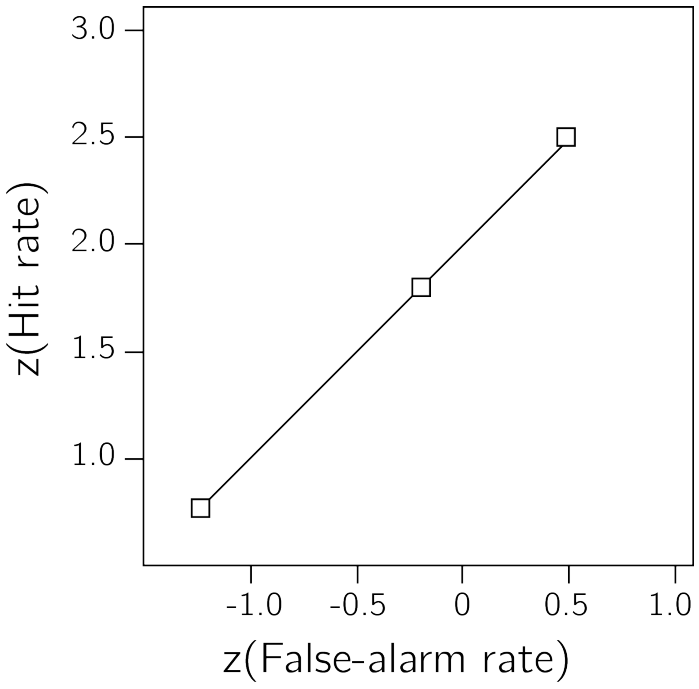


Figure A.3: zROC curve

specifically, in an unequal-variance setting, the LLR is proportional to the square of the decision variable. This means that it can be arbitrarily high for extremely positive or negative decision variables, but has a strict lower bound around the peak of the noise distribution.

A.3 SDT Measures for Metacognition

the ability to reliably track one's objective performance in a perceptual or a memory task is commonly taken as a measure of one's metacognitive ability (e.g., Fleming & Dolan, 2012). This ability can be quantified by asking participants for confidence judgments ("type-2 task") following their primary decision ("type-1 task"). The match or mismatch between objective performance and confidence can then be used as a proxy for their "metacognitive sensitivity."

The way this measure is extracted depends on the assumed underlying process. One potential process is a second-order SDT model, where a second variable is sampled following the type-1 decision, and this variable is then compared with an internal criterion that separates 'confident' responses from 'unconfident' responses (or a set of criteria, in the case of more than two possible confidence ratings). This variable is assumed to have higher values on average on trials in which the type-1 response was correct, similar to how the decision variable is higher on average on trials in which a signal is presented in a visual detection task (see figure A.4). Assuming that the two distributions of this confidence variable are normal, and assuming equal-variance, metacognitive sensitivity can then be quantified as the d' of the process

that aims to separate between correct and incorrect responses . Alternatively, a type-2 ROC curve can be generated by plotting $p(confidence > x|incorrect)$ against $p(confidence > x|correct)$ for different values of x, and the area under this curve can be extracted as a measure of metacognitive sensitivity. Under these assumptions, these SDT measures have the desired properties of relative invariance of d' and AuROC to the positioning of the criterion and to performance level in the type-1 task (Kunimoto, Miller, & Pashler, 2001).

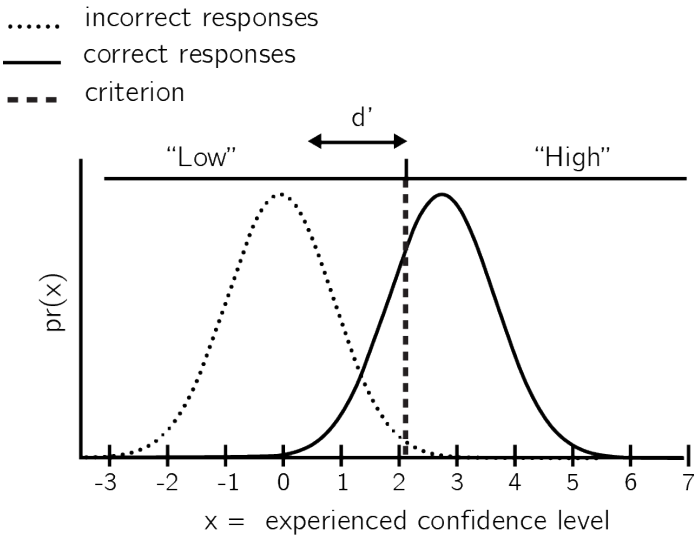


Figure A.4: (ref:ref:app1Kunimotocaption)

However, as discussed by Maniscalco & Lau (2012), this approach is unwarranted if the assumed underlying process uses the decision variable itself, or some transformation of it, in the generation of the confidence rating. In such a first-order model, the distance between the signal and noise distributions d' will be positively correlated with the estimated distance between the hypothetical 'correct' and 'incorrect' internal distributions. To correct for this, the authors propose to extract a measure of metacognitive sensitivity ($meta - d'$) that is fitted to the conditional distribution of confidence given stimulus and response, and compare it with d' (for example, by taking the ratio between these the two ($M_{ratio} = meta - d'/d'$). For an interactive primer on this approach, see matanmazor.shinyapps.io/sdtprimer.

Appendix B

Supp. materials for ch. 1

B.1 Effect of RT-based trial exclusion

Our pre-registered exclusion criterion for particularly slow (>1000 ms) and fast (<250 ms) trials resulted in the exclusion of a non-negligible number of trials per participant (more than two out of 12 trials on average). To test the robustness of our findings to

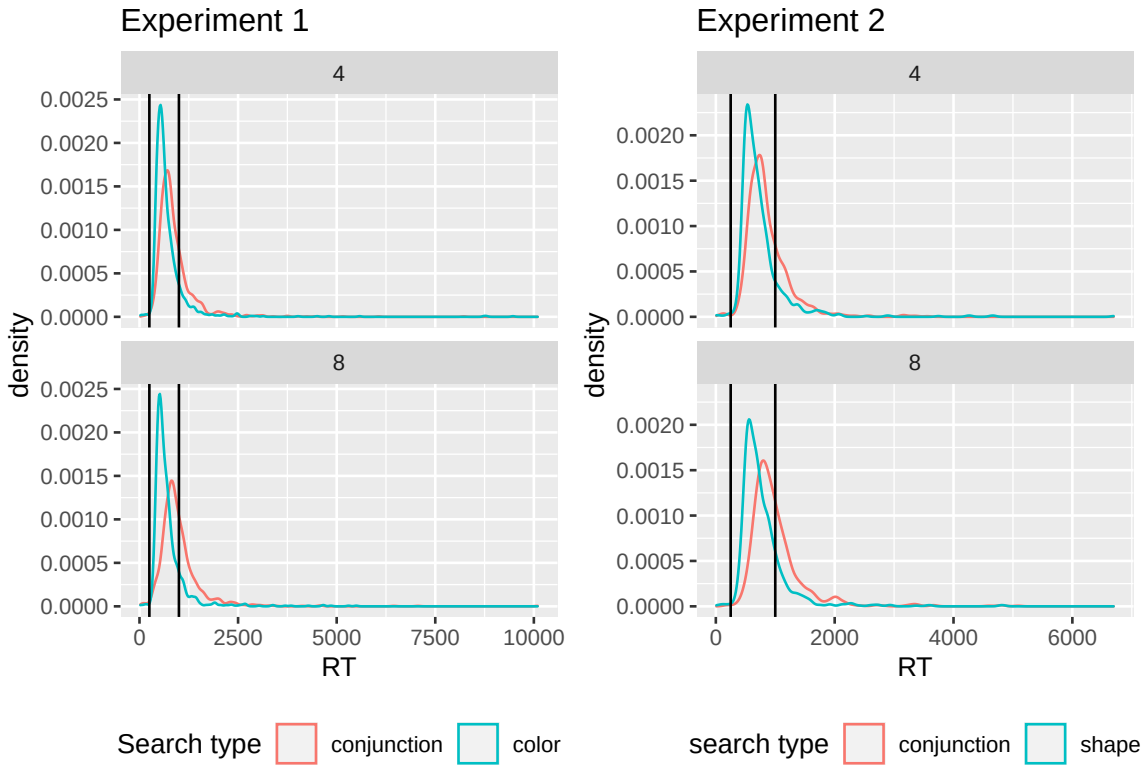


Figure B.1: RT histograms in the first block (first four trials) of Exp. 1 and 2 as a function of search type and set size. Our pre-registered analysis included only trials between the two vertical lines, corresponding to 250 and 1000 ms.

other RT-based exclusion criteria, we report here all pre-registered analyses, this time without excluding trials based on response time. Relaxing the RT-based exclusion criterion did not affect the results of most of our four pre-registered analyses, with the following exceptions: slopes for conjunctions slopes were now generally higher, and in Experiment 2, block 3, shape slope was not significantly different from the slope for conjunction search. Importantly, even when including these trials, shape slope was significantly different from conjunction slope in block 1. Furthermore, like in the original report, here also we find no learning effect between blocks 1 and 3.

B.1.1 Experiment 1

Hypothesis 1 (positive control): Search times in block 2 (target-present) followed the expected pattern, with a steep slope for conjunction search (M = 17.27, 95% CI [12.38, 22.15]) and a shallow slope for color search (M = 2.90, 95% CI [−0.59, 6.39]). The slope for color search was significantly lower than 10 ms/item and thus met our criterion for being considered 'pop-out' ($t(1,024) = -3.99$, $p < .001$). Furthermore, the difference between the slopes was significant ($t(891) = 4.25$, $p < .001$).

Hypothesis 2: Similar to the second block, the slope for the conjunction search was steep (M = 35.99, 95% CI [27.56, 44.43]). A clear 'pop-out' effect for color search was also evident ($M = -1.03$, 95% CI [$-\infty$, 5.18], $t(1,063) = -2.92$, $p = .002$). Furthermore, the average search slope for color search in this first block was significantly different from that of the conjunction search ($t(874) = 6.36$, $p < .001$), indicating that a color-absence pop-out is already in place prior to direct task experience.

Hypothesis 3: Like in the first block, in the third block color search complied with our criterion for 'pop-out' (M = 1.91, 95% CI [$-\infty$, 5.06], $t(1,053) = -4.24$, $p < .001$), and was significantly different from the conjunction search slope ($t(964) = 7.92$, $p < .001$).

Hypothesis 4: We find no evidence for a learning effect ($t(996) = -0.86$, $p = .389$). Furthermore, a Bayesian t-test with a scaled Cauchy prior for effect sizes (r=0.707) provided strong evidence in favour of the absence of a learning effect ($BF_{01} = 19.35$).

Hypothesis 5: The change in slope between blocks 1 and 3 was similar for color and conjunction search ($M = -9.31$, 95% CI [−21.70, 3.09], $t(745) = -1.47$, $p = .141$).

B.1.2 Experiment 2

Hypothesis 1 (positive control): Search times in block 2 (target-present) followed the expected pattern, with a steep slope for conjunction search (M = 21.87, 95% CI [15.55, 28.19]) and a shallow slope for shape search (M = 1.99, 95% CI [−3.81, 7.79]). The slope for shape search was significantly lower than 10 ms/item and thus met our criterion for being considered 'pop-out' ($t(792) = -2.71$, $p = .003$). Furthermore, the difference between the slopes was significant ($t(680) = 5.39$, $p < .001$).

Hypothesis 2: Also in the first block, the slope for conjunction search was steep (M = 34.71, 95% CI [27.71, 41.71]). The slope for shape search was numerically lower than 10 ms/item, but not significantly so (M = 9.68, 95% CI [$-\infty$, 15.17], $t(790) = -0.10$, $p = .462$). Still, the average search slope for shape search in this first

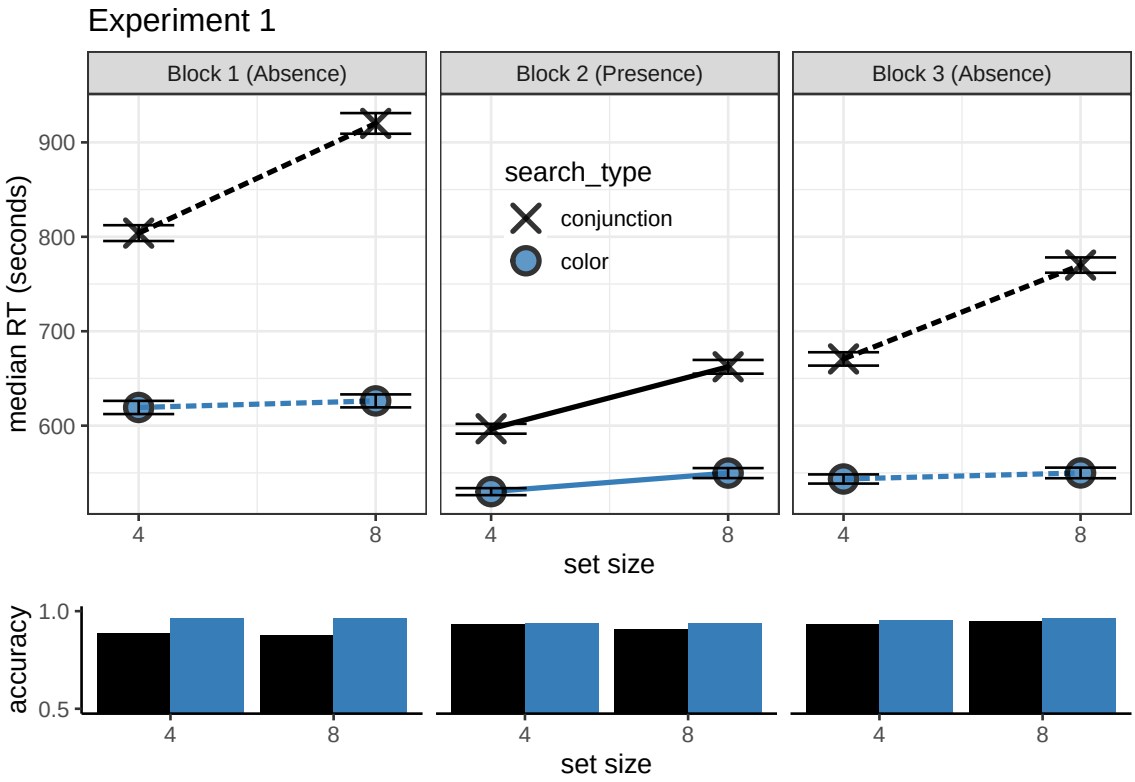


Figure B.2: Results from Experiment 1 without RT-based trial exclusion

block was significantly different from that of the conjunction search ($t(701) = 5.02$, $p < .001$).

Hypothesis 3: In the third block the slope for shape search was *higher* than 10 ms/item ($M = 17.97$, 95% CI [4.99, 30.96]), and not significantly different from the the slope for conjunction search ($t(751) = 0.81$, $p = .419$).

Hypothesis 4: To quantify a potential learning effect for shape search between blocks 1 and 3, we directly contrasted the search slope for shape search in these two 'target-absent' blocks. We find no evidence for a learning effect ($t(751) = -1.03$, $p = .303$). Furthermore, a Bayesian t-test with a scaled Cauchy prior for effect sizes (r=0.707) provided strong evidence against a learning effect ($BF_{01} = 14.37$).

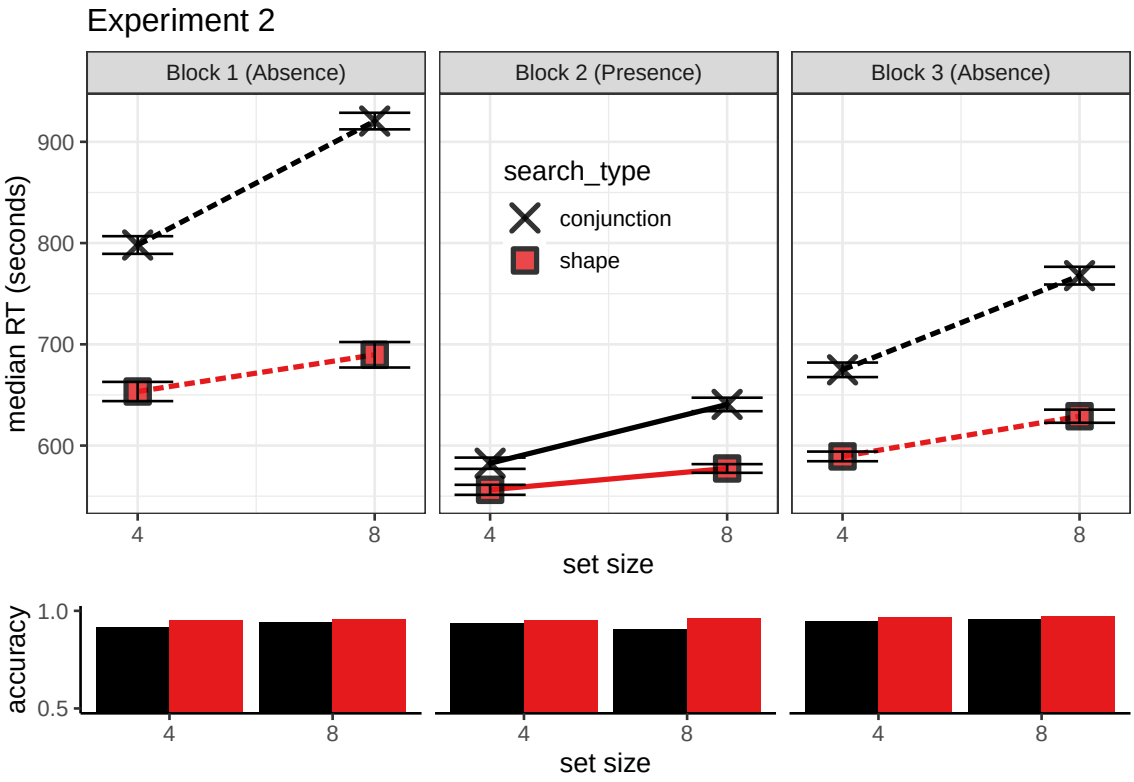


Figure B.3: Results from Experiment 1 without RT-based trial exclusion

Appendix C

Supp. materials for ch. 2

C.1 Bonus structure

We assume that participants represent the distribution of response times conditional on a specific search array as a right-skewed, positive distribution. Here, we assume that internal distributions of response times abide by the rule that

$$log(RT) \sim N(\mu, \sigma) \tag{C.1}$$

where σ is fixed per participant, and μ varies as a function of search difficulty.

The participants produces an estimate x. The expected bonus given for a trial is now:

$$E[bonus|x] = Pr_{n\sim N(\mu,\sigma)}[\log(x) > n] \cdot e^{-\log(x)/2}. \tag{C.2}$$

We can write $log(x) = \mu + \alpha \cdot \sigma$ for some number α. This number represents the position of the estimate relative to the distribution of response times, with lower values corresponding to more risky estimates, and higher values to more conservative ones. Then the expected bonus is:

$$\begin{aligned} E[bonus|\alpha] &= Pr_{n\sim N(0,1)}[\alpha > n] \cdot e^{-(\mu+\alpha\cdot\sigma)/2} \\ &= Pr_{n\sim N(0,1)}[\alpha > n] \cdot e^{-(\alpha\cdot\sigma)/2} \cdot e^{-\mu/2}. \end{aligned} \tag{C.3}$$

μ only appears in the third term in the product, which functions as a constant multiplier which scales the expected bonus equally for all choices of α. It then follows that the function relating the choice of α to the expected bonus preserves its shape for all possible values of μ:

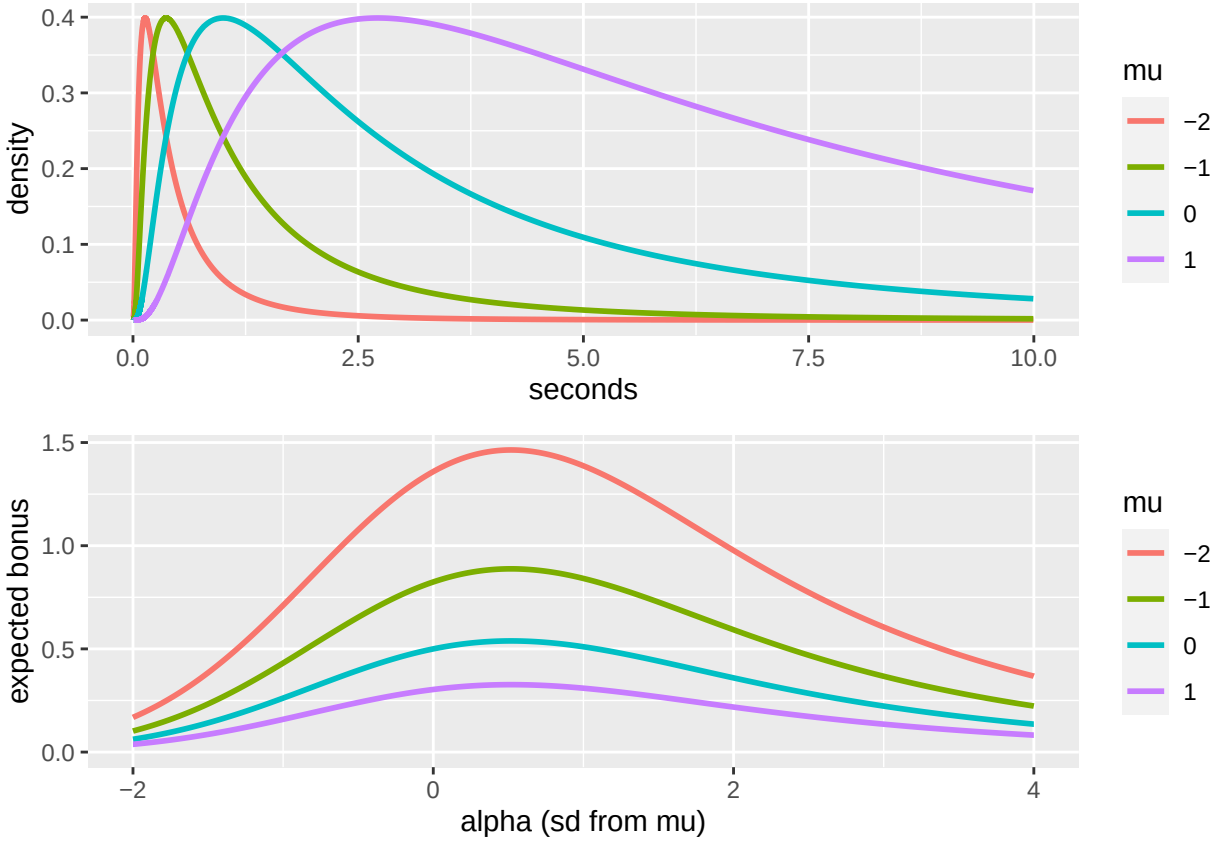


Figure C.1: Upper panel: response time distributions are modeled as exponents of values drawn from a normal distribution with different values of mu. Lower panel: the estimate value that maximizes the expected bonus is fixed with respect to the the mean of the log(RT) distributions, regardless of what the mean is. The expected bonus is higher for lower values of mu, but to maximize their bonus participants should always choose an estimate that is positioned in the 70 quantile of the RT distribution.

Appendix D

Supp. materials for ch. 3

D.1 Additional analyses: Exp. 1

D.1.1 Response time, confidence, and metacognitive sensitivity differences

In detection, participants were generally slower to deliver 'no' responses compared to 'yes' responses (median difference: 85.37 ms, $t(9) = -3.46$, $p = .007$ for a t-test on the log-transformed response times; see Fig. 3.3, third row). No significant difference in response times was observed for the discrimination task (median difference: 6.16 ms, $t(9) = -0.43$, $p = .676$).

Confidence in detection was generally higher than in discrimination ($M_d = 0.06$, 95% CI [0.01, 0.12], $t(9) = 2.49$, $p = .035$; see Fig. 3.3, fourth row). Within detection, confidence in 'yes' responses was generally higher than confidence in 'no' responses ($M = 0.08$, 95% CI [0.03, 0.13], $t(9) = 3.49$, $p = .007$). No difference in average confidence levels was found between the two discrimination responses ($M = 0.02$, 95% CI [-0.03, 0.06], $t(9) = 0.91$, $p = .384$).

Following Meuwese, Loon, Lamme, & Fahrenfort (2014), we extracted response-conditional type-2 ROC (rc-ROC) curves for the two tasks. Unlike traditional type-I ROC curves that provide a summary of subjects' ability to distinguish between two external world states, type 2 ROC curves represent their ability to track the accuracy of their own responses. The area under the response-conditional ROC curve (auROC2) is a measure of metacognitive sensitivity, with higher values corresponding to more accurate metacognitive monitoring.

Mean response-conditional ROC curves for the two responses in the discrimination task closely matched ($M = 0.00$, 95% CI [-0.05, 0.05], $t(9) = 0.13$, $p = .900$), indicating that on average, participants had similar metacognitive insight into the accuracy of the two discrimination responses. In contrast, auROC2 estimates for 'yes' responses were significantly higher than for 'no' responses, indicating a metacognitive asymmetry between the two detection responses (group difference in auROC2: $M = 0.11$, 95% CI [0.03, 0.18], $t(9) = 3.28$, $p = .010$).

D.1.2 zROC curves

An asymmetry in metacognitive sensitivity for 'yes' and 'no' responses is predicted by unequal-variance Signal Detection Theory (*uvSDT*). Specifically, if the signal distribution is wider than the noise distribution, the overlap between the distributions will be more pronounced for misses and correct rejections than for hits and false alarms, making metacognitive judgments for 'no' responses objectively more difficult. Unequal-variance SDT predicts that plotting the type-1 ROC curve in z-space (taking the inverse cumulative distribution of the confidence rating histogram) will result in a straight line with a slope equal to $\frac{\sigma_{noise}}{\sigma_{signal}}$. Because the variance of the signal distribution is higher than that of the noise distribution, zROC slopes are typically shallow, with slopes below 1.

We used linear regression to estimate the slope of the zROC curve. To control for underestimation of the slope due to regression to the mean (Wickens, 2002, p. 56), we fitted two regression models for the task data of each participant: one predicting $Z(h)$ based on $Z(f)$ (slope s_1) and one predicting $Z(f)$ based on $Z(h)$ (slope s_2). We then used $\frac{log(s_1)-log(s2)}{2}$ as a bias-free measure of the zROC slope. In equal-variance SDT, this value is predicted to be 0, corresponding to a slope of 1.

Indeed, slopes were generally shallow for detection zROC curves (as predicted by an unequal-variance SDT model; $M = -0.15$, 95% CI $[-0.27, -0.04]$, $t(9) = -2.95$, $p = .016$), and not significantly different from 1 for discrimination zROC curves (as predicted by equal-variance SDT; $M = 0.00$, 95% CI $[-0.09, 0.10]$, $t(9) = 0.07$, $p = .946$).

These results support a difference in the variance-structure of the representation of signal and noise, such that the representation of signal is more varied across trials. However, it is still possible that some of the metacognitive asymmetry in detection (the difference in auROC between 'yes' and 'no' responses) reflects additional higher-order processes that cannot be captured by a first-order signal-detection model. If this was the case, zROC curves for detection should not only be more shallow, but also less linear than for discrimination, reflecting poorer fit of the signal-detection model to detection. In order to test if this was the case, we compared the subject-wise R^2 values for the detection and discrimination zROC regression lines. R^2 values reflect the goodness of fit of a linear model to the data. These values were similar for the two tasks ($M_d = -0.01$, 95% CI $[-0.03, 0.01]$, $t(9) = -0.91$, $p = .385$), suggesting that a first-order SDT model accounted equally well for the two tasks.

D.1.3 Confidence response-time alignment

Following our pre-registered analysis plan, we extracted a Spearman correlation coefficient between confidence and response times separately for the two tasks and four responses. We find a negative correlation in all four cases (discrimination responses: -0.40 and -0.39, detection 'yes': -0.41, detection 'no': -0.33). As hypothesized, this negative correlation was significantly attenuated in detection 'no' responses compared to detection 'yes' responses (tested with a one-tailed t-test: $t(9) = -1.97$, $p = .040$). The difference in correlation strength between detection 'no' responses and

discrimination responses was only marginally significant ($t(9) = -1.68$, $p = .063$).

D.1.4 Global metacognitive estimates

At the end of each 100-trial block, participants estimated their block-wise accuracy. Mean estimated accuracy was 0.71 for discrimination and 0.69. These figures are close to true correct response rates: 0.74 in discrimination and 0.72 in detection.

A difference of 0.03 between mean accuracy estimates for discrimination and detection was not significant at the group level ($t(9) = 1.71$, $p = .121$).

D.2 Additional analyses: Exp. 2

D.2.1 Response time, confidence, and metacognitive sensitivity differences

Participants were slower to deliver 'no' responses compared to 'yes' responses (median difference: 77.12 ms, $t(101) = -6.84$, $p < .001$ for a t-test on the log-transformed response times; see Fig. 3.7, third row). No significant difference in response times was observed for the discrimination task (median difference: 10.90 ms, $t(101) = -1.40$, $p = .165$).

Confidence in detection was generally lower than in discrimination, consistent with lower accuracy in this task ($M_d = -0.09$, 95% CI $[-0.11, -0.07]$, $t(101) = -8.41$, $p < .001$; see Fig. 3.7, fourth row). Within detection, confidence in 'yes' responses was generally higher than confidence in 'no' responses ($M = 0.10$, 95% CI [0.07, 0.12], $t(101) = 8.15$, $p < .001$). No difference in average confidence levels was observed between the two discrimination responses ($M = 0.00$, 95% CI $[-0.02, 0.02]$, $t(101) = -0.03$, $p = .974$).

In contrast to the results of Exp. 1, auROC2 for 'yes' and 'no' responses were not significantly different (group difference in area under the response-conditional curve, AUROC2: $M = 0.02$, 95% CI $[-0.02, 0.06]$, $t(58) = 1.13$, $p = .264$; see Fig. 3.7, first and second rows). auROC2s were not significantly different also when controlling for type-1 response and confidence biases ($M = 0.01$, 95% CI $[-0.03, 0.05]$, $t(58) = 0.59$, $p = .560$).

D.2.2 zROC curves

Unlike in Experiment 1, detection zROC slopes were not significantly different from 1 ($M = -0.04$, 95% CI $[-0.09, 0.01]$, $t(100) = -1.52$, $p = .131$), whereas discrimination zROC slopes were significantly shallower than 1 ($M = -0.14$, 95% CI $[-0.25, -0.02]$, $t(93) = -2.29$, $p = .024$). This unexpected result indicates equal variance for the signal and noise distributions, but higher variance for targets presented on the right than on the left. Furthermore, first-order SDT fitted the data significantly better for the detection task than for the discrimination (difference in R^2 for the two tasks: $M = 0.15$, 95% CI [0.12, 0.18], $t(93) = 8.85$, $p < .001$).

D.3 Additional analyses: Exp. 3

D.3.1 Response time, confidence, and metacognitive sensitivity differences

Participants were also slower to deliver 'no' responses compared to 'yes' responses (median difference: 71.81 ms, $t(97) = -6.66$, $p < .001$ for a t-test on the log-transformed response times; see Fig. 3.11, third row). No significant difference in response times was observed for the discrimination task (median difference: 19.28 ms, $t(96) = -0.28$, $p = .781$).

Confidence in detection was generally lower than in discrimination, consistent with lower accuracy in this task ($M_d = -0.04$, 95% CI $[-0.06, -0.02]$, $t(97) = -3.77$, $p < .001$; see Fig. 3.11, fourth row). Within detection, confidence in 'yes' responses was generally higher than confidence in 'no' responses ($M = 0.09$, 95% CI $[0.07, 0.11]$, $t(97) = 8.00$, $p < .001$). No difference in average confidence levels was observed between the two discrimination responses ($M = -0.01$, 95% CI $[-0.03, 0.02]$, $t(97) = -0.65$, $p = .519$).

D.3.2 Reverse correlation analysis of standard trials only

In the following, we repeat the reverse correlation analysis for Exp 3. on the subset of trials where luminance was not increased by 2/255.

Discrimination decisions Discrimination decisions were sensitive to fluctuations in luminance during the first 300 milliseconds of the trial ($t(97) = 8.47$, $p < .001$). We found no evidence for a positive evidence bias in discrimination decisions, even when grouping evidence based on the location of the true signal rather than subjects' decisions ($t(93) = -0.23$, $p = .819$).

Discrimination confidence Luminance within the first 300 milliseconds had a significant effect on confidence ratings ($t(97) = 6.38$, $p < .001$; see Fig. D.1, right panels). A positive evidence bias in discrimination confidence was not significant in this sample ($t(97) = 1.42$, $p = .157$).

D.4 Pseudo-discrimination analysis

In our pre-registration document (https://osf.io/8u7dk/), we specified our plan for *pseudo-discrimination analysis*, where we analyze detection 'signal' trials as if they were discrimination trials:

> In this analysis, we will assume that in the majority of 'different' trials, when participants responded 'yes' they correctly identified the brighter set. For example, a detection trial in which the brighter set was presented on the right and in which the participant responded 'yes' will be treated as a

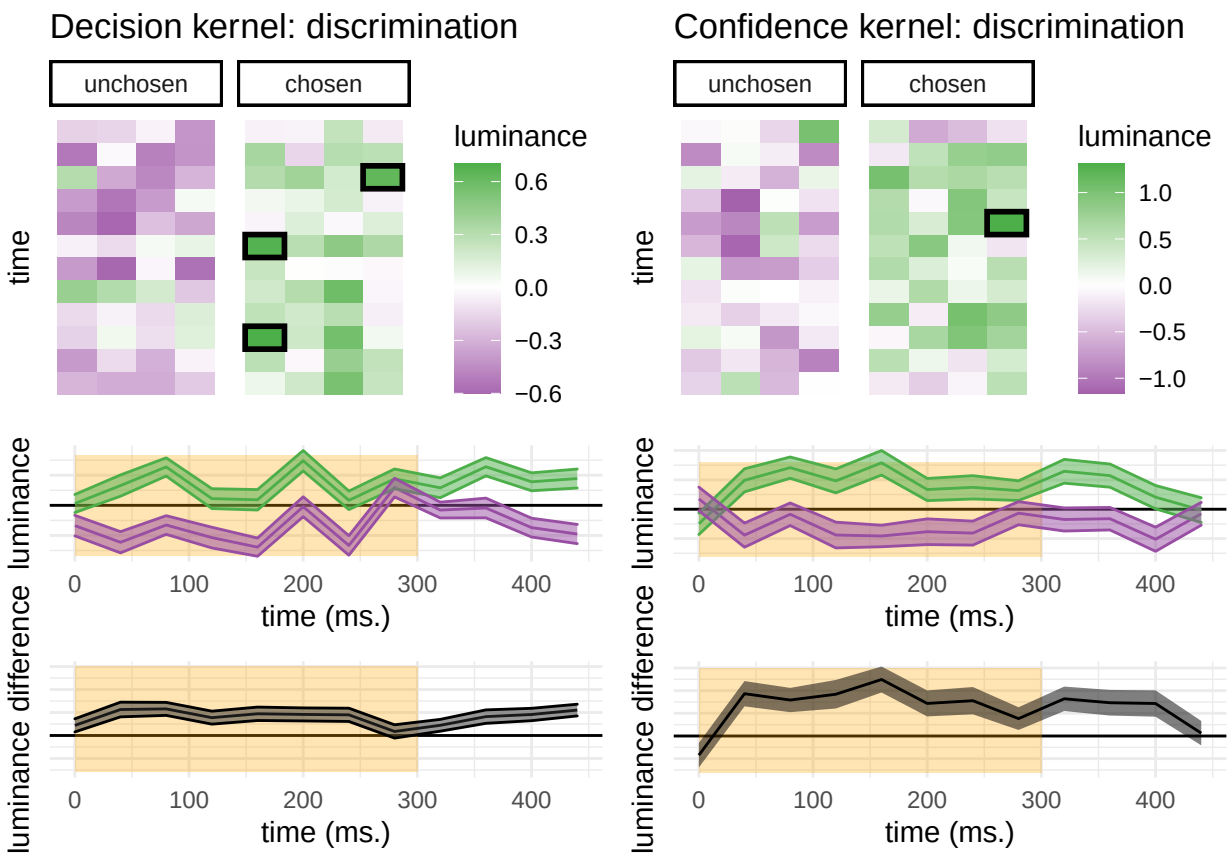


Figure D.1: Decision and confidence discrimination kernels, Experiment 3, standard trials only.

discrimination trial in which the participant responded 'right.' Conversely, a trial in which the brighter set was presented on the right and in which the participant responded 'no' will be treated as a discrimination trial in which the participant responded 'left.' These hypothetical responses will then be submitted to the same reverse correlation analysis described in the previous section confidence kernels.

We subsequently realized that a much simpler approach is to contrast 'yes' and 'no' responses for the true and opposite direction of motion (or flickering stimuli) in signal trials. This alternative approach does not entail treating 'no' responses as the successful detection of a wrong signal. The results of this analysis mostly agreed with the pre-registered pseudo-discrimination analysis. For completeness, we include the pre-registered pseudo-discrimination analysis for both experiments here.

D.4.1 Exp. 1

Pseudo-discrimination decision kernels were highly similar to discrimination decision kernels. Here also, motion energy during the first 300 milliseconds of the stimulus had a significant effect on decision ($t(9) = 4.18$, $p = .002$) and on decision confidence

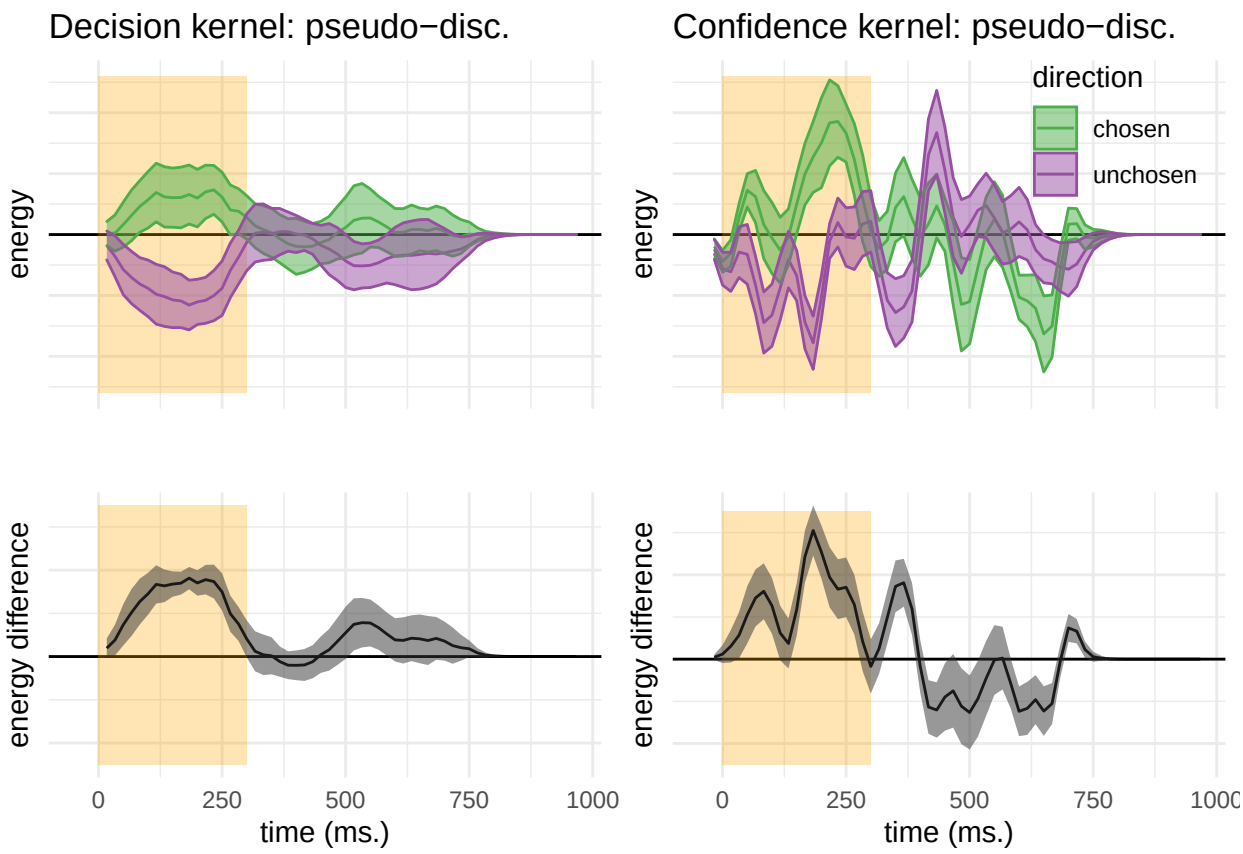


Figure D.2: Decision and confidence pseudo-discrimination kernels, Experiment 1. Upper left: motion energy in the "chosen" (green) and "unchosen" (purple) direction as a function of time. Bottom left: a subtraction between energy in the "chosen" and "unchosen" directions. Upper right: confidence effects for motion energy in the "chosen" (green) and "unchosen" (purple) directions. Lower right: a subtraction between confidence effects in the "chosen" and "unchosen" directions. Shaded areas represent the the mean +- one standard error. The first 300 milliseconds of the trial are marked in yellow.

($t(9) = 3.26$, $p = .010$). However, unlike discrimination, where motion energy in the chosen direction influenced decision confidence more than motion energy in the unchosen direction, no such bias was observed for detection responses ($t(9) = 0.20$, $p = .849$).

While motion energy during the first 300 milliseconds of the trial significantly affected confidence in 'yes' responses ($t(9) = 5.52$, $p < .001$), it had no significant effect on confidence in 'no' responses ($t(9) = -0.09$, $p = .932$). However, given that the pseudo-discrimination analysis was performed on signal trials only, confidence kernels for 'no' responses were based on fewer trials than confidence kernels for 'yes' responses, such that the absence of a significant effect in 'no' responses may reflect insufficient statistical power to detect one.

D.4.2 Exp. 2

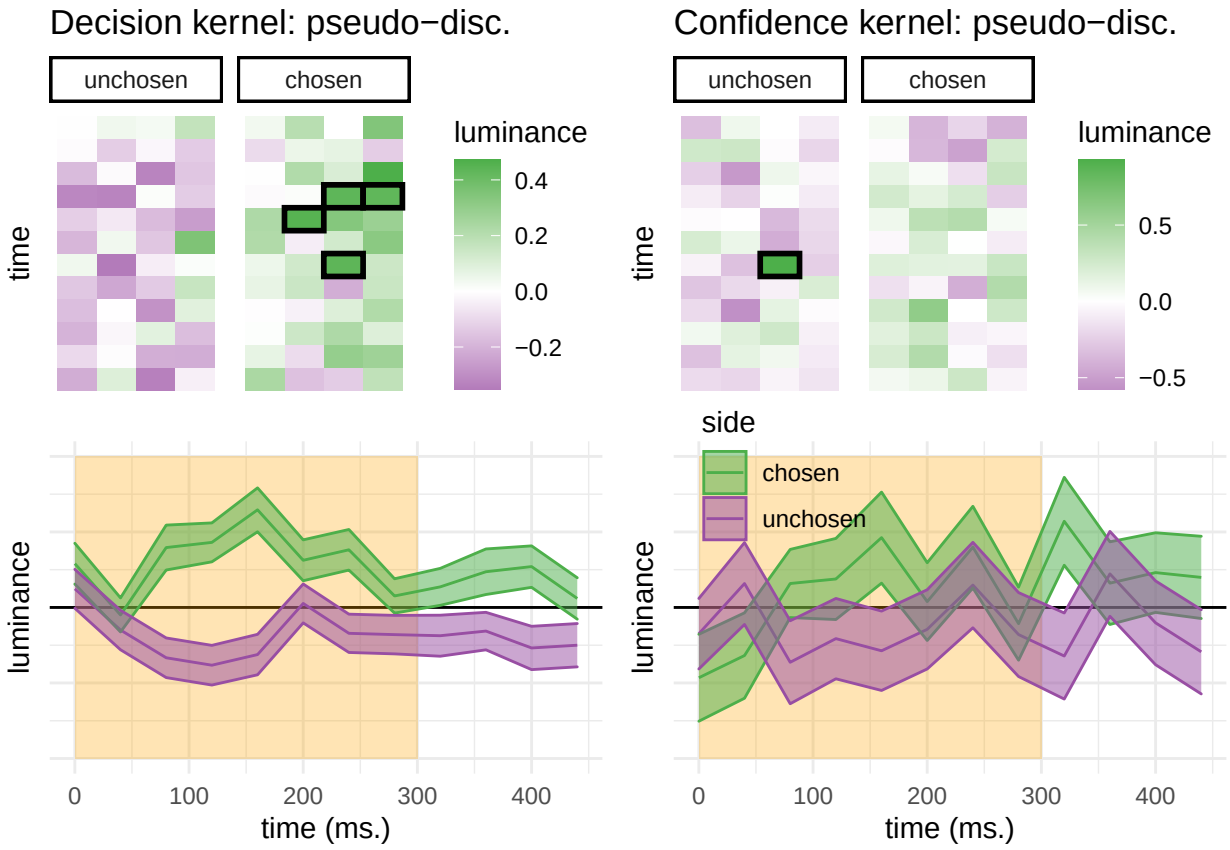


Figure D.3: Decision and confidence pseudo-discrimination kernels, Experiment 2. Upper left: luminance in the "chosen" (green) and "unchosen" (purple) stimulus as a function of time and spatial position. Bottom left: decision kernel averaged across the four spatial positions. Upper right: confidence effects for motion energy in the "chosen" (green) and "unchosen" (purple) stimuli. Bottom right: confidence effects averaged across the four spatial positions. Shaded areas represent the the mean +- one standard error. The first 300 milliseconds of the trial are marked in yellow. Black frames denote significance at the 0.05 level controlling for family-wise error rate for 48 comparisons.

Similar to decision kernels in Exp. 2, random fluctuations in luminance during the first 300 milliseconds of the stimulus had a significant effect on decision ($t(101) = 6.68$, $p < .001$). However, in Exp. 2 this anslysis revealed no effect of luminance on decision confidence ($t(99) = 1.36$, $p = .178$), and no positive evidence bias in confidence judgments ($t(99) = -0.66$, $p = .512$).

D.5 Stimulus-dependent noise model

D.5.1 Discrimination

Generative model

Stimuli were represented as pairs of numbers, corresponding to the two sensory channels (e.g., right and left motion). One sensory channel transmitted pure noise, and one channel had additional signal in it. The signal channel was chosen randomly for each trial with equal probability.

$$x_t^c \sim \begin{cases} \mathcal{N}(0,1), & \text{if signal.} \\ \mathcal{N}(1,1), & \text{if noise.} \end{cases} \tag{D.1}$$

On top of the presented noise, we added perceptual noise to the stimulus. Importantly, this additional noise affected the decisions and confidence ratings of the simulated agent, but did not affect trial-wise estimates of stimulus energy for the reverse correlation analysis. The noise was channel specific, and its magnitude dependent on the magnitude of the underlying signal:

$$\epsilon_t^c \sim \mathcal{N}(0, 2^{x_t^c}) \tag{D.2}$$

$$x_t'^c = x_t^c + \epsilon_t^c \tag{D.3}$$

Inference

The log likelihood ratio is computed to decide whether it is more likely that the signal was in channel 1 or 2.

$$LLR = log(p([x_t'^1, x_t'^2]|stim = [x^s, x^n]) - log(p([x_t'^1, x_t'^2]|stim = [x^n, x^s]) \tag{D.4}$$

$$decision_t = \begin{cases} 1, & \text{if } LLR > 1. \\ 2, & \text{else.} \end{cases} \tag{D.5}$$

$$confidence_t = |LLR| \tag{D.6}$$

```
class Model:
    def __init__(self, mu, sigma, noise_factor):

        self.df = pd.DataFrame()
        self.mu = mu
        self.sigma = sigma
        self.noise_factor = noise_factor

        # if noise factor > 0, approximate density function
```

```
        # with grid.
        if noise_factor > 0:

          X = np.arange(-100,100,0.1)
          Xboundaries = np.arange(-100,100.1,0.1)-0.05
          marginal_signal=[0]*len(X)
          marginal_noise=[0]*len(X)

          for x in X:
            conditional = stats.norm(x,self.noise_factor**x).pdf(X);
            prior_noise = stats.norm(self.mu[0],self.sigma[0]).pdf(x)
            prior_signal = stats.norm(self.mu[1],self.sigma[1]).pdf(x)
            marginal_noise = [p+conditional[i]*prior_noise
                              for i,p in enumerate(marginal_noise)];
            marginal_signal = [p+conditional[i]*prior_signal
                               for i,p in enumerate(marginal_signal)];


          self.signal_dist = stats.rv_histogram(
              (np.array(marginal_signal),Xboundaries));

          self.noise_dist = stats.rv_histogram(
              (np.array(marginal_noise),Xboundaries));

        # else, use normal distributions.
        else:
          self.signal_dist = stats.norm(self.mu[1],self.sigma[1]);
          self.noise_dist = stats.norm(self.mu[0],self.sigma[0])

    def runModel(self, num_trials):

        # first, decide which is the true direction in each trial
        # (p=0.5)
        self.df['direction'] = ['r' if flip else 'l'
               for flip in np.random.binomial(1,0.5,num_trials)]

        self.getMotionEnergy()

        self.extractLLR()

        self.makeDecision()

        self.rateConfidence()
```

```
        self.df['correct'] = self.df.apply(lambda row:
               row.direction==row.decision, axis=1)

        #energy in chosen direction
        self.df['E_c'] = self.df.apply(lambda row:
                                       row.E_r if row.decision=='r'
                                       else row.E_l, axis=1)

        #energy in unchosen direction
        self.df['E_u'] = self.df.apply(lambda row:
                                       row.E_l if row.decision=='r'
                                       else row.E_r, axis=1)

    def runModelForSpecifiedValues(self,
          specified_values,
          repetitions=1):

        # no direction here
        self.df['direction'] =
          ['x']*len(specified_values)**2*repetitions

        self.df['E_r'] =
          specified_values*len(specified_values)*repetitions;

        self.df['E_l'] = list(np.repeat(
            specified_values,len(specified_values)))*repetitions;

        # how it appears to subjects
        if self.noise_factor>0:
          self.df['E_ra'] = self.df.apply(lambda row: row.E_r +
                 np.random.normal(0, self.noise_factor**row.E_r),
                 axis=1);

          self.df['E_la'] = self.df.apply(lambda row: row.E_l +
                 np.random.normal(0, self.noise_factor**row.E_l),
                 axis=1);
        else:
          self.df['E_ra']=self.df['E_r'];
          self.df['E_la']=self.df['E_l'];

        self.extractLLR()

        self.makeDecision()
```

```
        self.rateConfidence()

        self.df['correct'] = self.df.apply(lambda row:
               row.direction==row.decision, axis=1)

        #energy in chosen direction
        self.df['E_c'] = self.df.apply(lambda row:
                                       row.E_r if row.decision=='r'
                                       else row.E_l, axis=1)

        #energy in unchosen direction
        self.df['E_u'] = self.df.apply(lambda row:
                                       row.E_l if row.decision=='r'
                                       else row.E_r, axis=1)

    def getMotionEnergy(self):
    # sample the motion energy for left and right as a function of
    # the true direction
        self.df['E_r'] = self.df.apply(lambda row:
               np.random.normal(self.mu[1],self.sigma[1])
               if row.direction=='r'
               else np.random.normal(self.mu[0],self.sigma[0]),
               axis=1)

        self.df['E_l'] = self.df.apply(lambda row:
               np.random.normal(self.mu[1],self.sigma[1])
               if row.direction=='l'
               else np.random.normal(self.mu[0],self.sigma[0]),
               axis=1)

        # how it appears to subjects
        if self.noise_factor>0:
          self.df['E_ra'] = self.df.apply(lambda row: row.E_r +
                 np.random.normal(0, self.noise_factor**row.E_r),
                 axis=1);
          self.df['E_la'] =  self.df.apply(lambda row: row.E_l +
                 np.random.normal(0, self.noise_factor**row.E_l),
                 axis=1)
        else:
          self.df['E_ra']=self.df['E_r'];
          self.df['E_la']=self.df['E_l'];

    def extractLLR(self):
```

```
        # extract the Log Likelihood Ratio (LLR)
        #log(p(Er|r))-log(p(Er|l)) + log(p(El|r))-log(p(El|l))
        self.df['LLR'] = self.df.apply(lambda row:
               np.log(self.signal_dist.pdf(row.E_ra))-
               np.log(self.noise_dist.pdf(row.E_ra)) +
               np.log(self.noise_dist.pdf(row.E_la))-
               np.log(self.signal_dist.pdf(row.E_la)), axis=1)
    def makeDecision(self):

        # we assume that our participant chooses the direction associated
        # with higher likelihood
        self.df['decision'] =  self.df.apply(lambda row:
                                             'r' if row.LLR>0 else 'l',
                                             axis=1)

    def rateConfidence(self):

        # and rates their confidence in proportion to the absolute LLR
        self.df['confidence'] = abs(self.df['LLR'])
```

D.5.2 Detection

Generative model

Similar to detection, except that on half of the trials both channels transmitted noise only.

Inference

The log likelihood ratio is computed to decide whether it is more likely that the signal was present or absent.

$$p(x|signal) = 0.5 \times p([x_t'^1, x_t'^2]|stim = [x^s, x^n]) + 0.5 \times p([x_t'^1, x_t'^2]|stim = [x^n, x^s]) \quad \text{(D.7)}$$

$$p(x|noise) = p([x_t'^1, x_t'^2]|stim = [x^n, x^n]) \quad \text{(D.8)}$$

$$LLR = log(p(x|signal)) - log(p(x|noise)) \quad \text{(D.9)}$$

$$decision_t = \begin{cases} 1, & \text{if } LLR > 1. \\ 2, & \text{else.} \end{cases} \quad \text{(D.10)}$$

$$confidence_t = |LLR| \quad \text{(D.11)}$$

```
class DetectionModel(Model):

    def runModel(self, num_trials):
```

```
        # first, decide which is the true direction in each trial
        #(p=0.5)
        self.df['direction'] = ['r' if flip else 'l'
               for flip in np.random.binomial(1,0.5,num_trials)]

        # decide whether motion is present or absent.
        self.df['motion'] = ['p' if flip else 'a'
               for flip in np.random.binomial(1,0.5,num_trials)]

        self.getMotionEnergy()

        self.extractLLR()

        self.makeDecision()

        self.rateConfidence()

        self.df['correct'] = self.df.apply(lambda row:
               row.motion==row.decision,
               axis=1)

         #energy in true direction
        self.df['E_t'] = self.df.apply(lambda row:
               row.E_r if row.direction=='r'
               else row.E_l,
               axis=1)

        #energy in opposite direction
        self.df['E_o'] = self.df.apply(lambda row:
               row.E_l if row.direction=='r'
               else row.E_r,
               axis=1)

    def runModelForSpecifiedValues(self, specified_values, repetitions=1):

        # no direction/motion here
        self.df['direction'] =
          ['x']*len(specified_values)**2*repetitions

        self.df['motion'] =
          ['x']*len(specified_values)**2*repetitions

        self.df['E_r'] =
          specified_values*len(specified_values)*repetitions;
```

```
        self.df['E_l'] = list(np.repeat(
            specified_values,len(specified_values)))*repetitions;

        # how it appears to subjects
        if self.noise_factor>0:
          self.df['E_ra'] = self.df.apply(lambda row: row.E_r +
                        np.random.normal(0, self.noise_factor**row.E_r),
                        axis=1);
          self.df['E_la'] = self.df.apply(lambda row: row.E_l +
                        np.random.normal(0, self.noise_factor**row.E_l),
                        axis=1)
        else:
          self.df['E_ra']=self.df['E_r'];
          self.df['E_la']=self.df['E_l'];

        self.extractLLR()

        self.makeDecision()

        self.rateConfidence()

        self.df['correct'] =
          self.df.apply(lambda row:
            row.motion==row.decision, axis=1)

    def getMotionEnergy(self):
    # sample the motion energy for left and right as a function of
    # the true direction
        self.df['E_r'] = self.df.apply(lambda row:
               np.random.normal(self.mu[1],self.sigma[1])
               if row.direction=='r' and row.motion=='p'
               else np.random.normal(self.mu[0],self.sigma[0]),
               axis=1)

        self.df['E_l'] = self.df.apply(lambda row:
               np.random.normal(self.mu[1],self.sigma[1])
               if row.direction=='l' and row.motion=='p'
               else np.random.normal(self.mu[0],self.sigma[0]),
               axis=1)

        # how it appears to subjects
        if self.noise_factor>0:
          self.df['E_ra'] = self.df.apply(lambda row: row.E_r +
                 np.random.normal(0, self.noise_factor**row.E_r),
```

```
                axis=1);
          self.df['E_la'] = self.df.apply(lambda row: row.E_l +
                np.random.normal(0, self.noise_factor**row.E_l),
                axis=1)
        else:
          self.df['E_ra']=self.df['E_r'];
          self.df['E_la']=self.df['E_l'];

    def extractLLR(self):

        self.df['LLR'] = self.df.apply(lambda row:
               np.log(0.5*self.signal_dist.pdf(row.E_ra)*
               self.noise_dist.pdf(row.E_la) +
               0.5*self.signal_dist.pdf(row.E_la)*
               self.noise_dist.pdf(row.E_ra)) -
               np.log(self.noise_dist.pdf(row.E_la) *
               self.noise_dist.pdf(row.E_ra)),
               axis=1)

    def makeDecision(self):

        # we assume that our participant just chooses the option
        # associated with higher likelihood
        self.df['decision'] =  self.df.apply(lambda row:
               'p' if row.LLR>0
                else 'a',
                axis=1)

    def rateConfidence(self):

        # and rates their confidence in proportion to the absolute
        # LLR
        self.df['confidence'] = abs(self.df['LLR'])
```

D.5.3 Effects of evidence on decision and confidence: Exp. 2 and 3

To compare participants' empirical behaviour to our model simulations, we plotted optimal behaviour, participants' responses, and confidence in correct responses, as a function of perceptual evidence in a two-dimensional representational space. First, for each trial we extracted mean luminance (minus background luminance) in the first 300 milliseconds in the right and left stimuli. These numbers were rounded to the closest integer. For each tuple of such integers, we extracted the posterior probability for stimulus category (Fig. D.4, top row), participants' empirical discrimination and

detection decisions (middle row), and participants' subjective confidence in correct responses (bottom row).

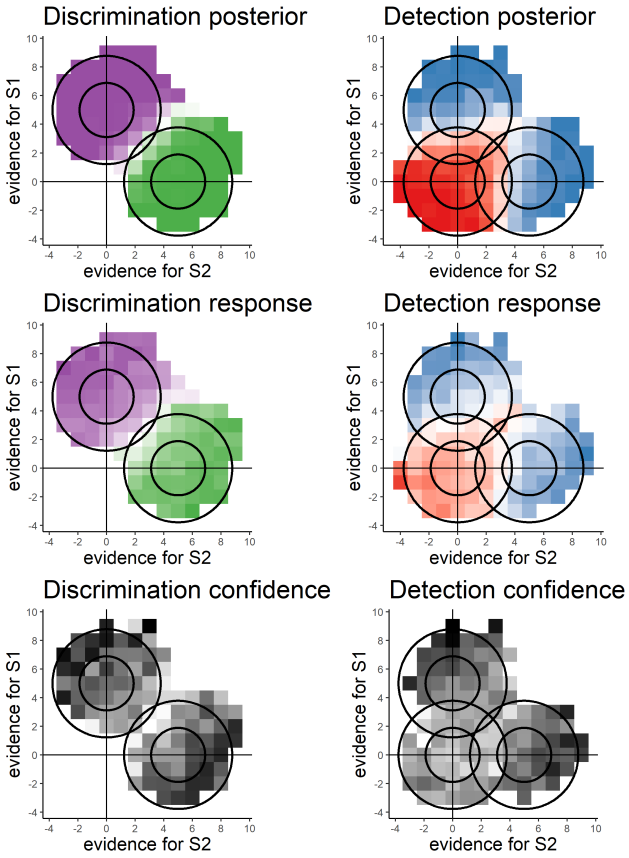


Figure D.4: Top row: posterior probability of stimulus category given perceptual evidence for discrimination (left) and detection (right). Middle row: decision probability as a function of perceptual evidence. Bottom row: mean confidence in correct responses as a function of perceptual evidence.

Appendix E

Supp. materials for ch. 4

E.1 Confidence button presses

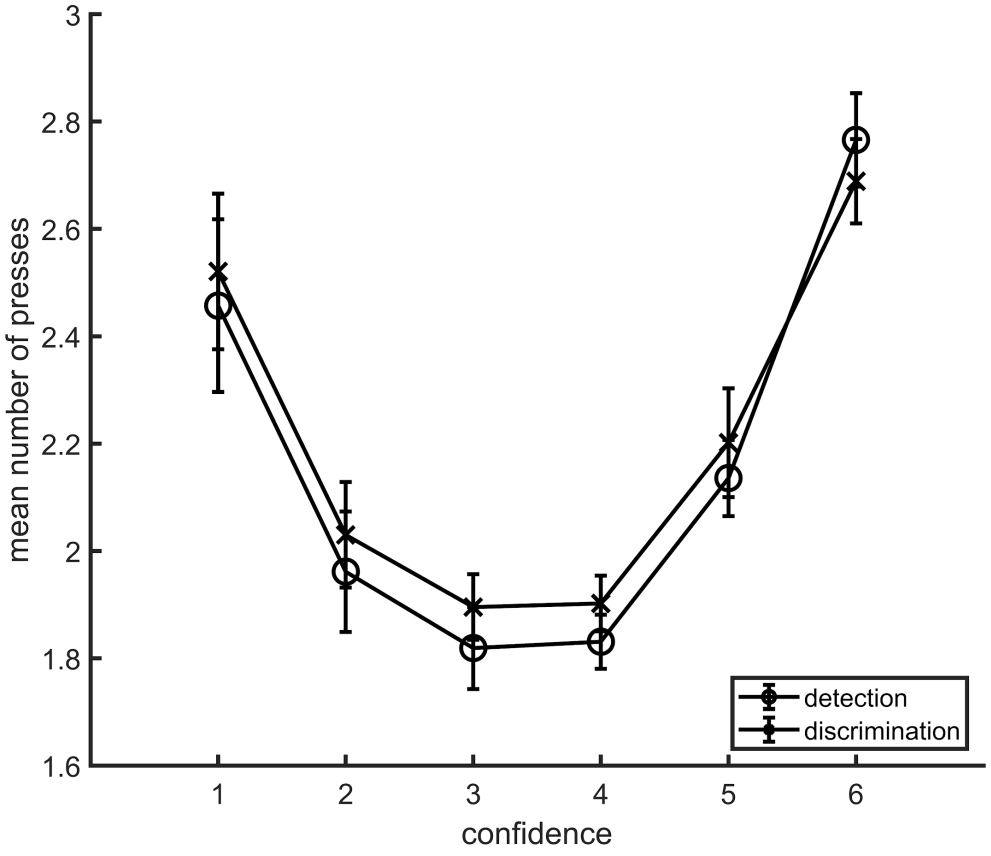


Figure E.1: Average number of button presses for each confidence level, as a function of task. More button presses were needed on average to reach the extreme confidence ratings, hence the quadratic shape. No difference between the two tasks was observed in the mean number of button presses for any of the confidence levels. Error bars represent the standard error of the mean.

E.2 zROC curves

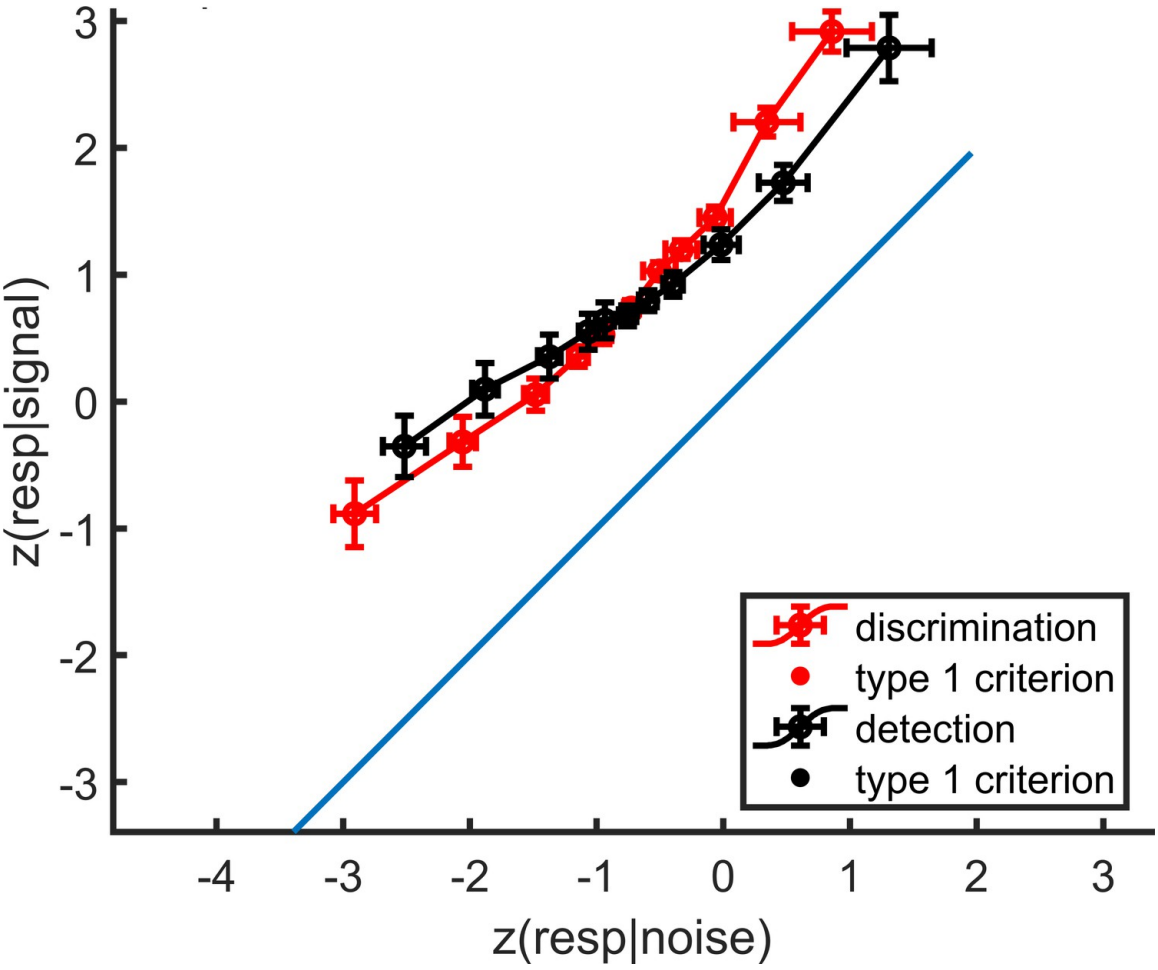


Figure E.2: mean zROC curves for the discrimination and detection tasks. As expected in a uv-SDT setting, the discrimination curve is approximately linear with a slope of 1, and the detection curve is approximately linear with a shallower slope. Error bars represent the standard error of the mean.

E.3 Global confidence design matrix

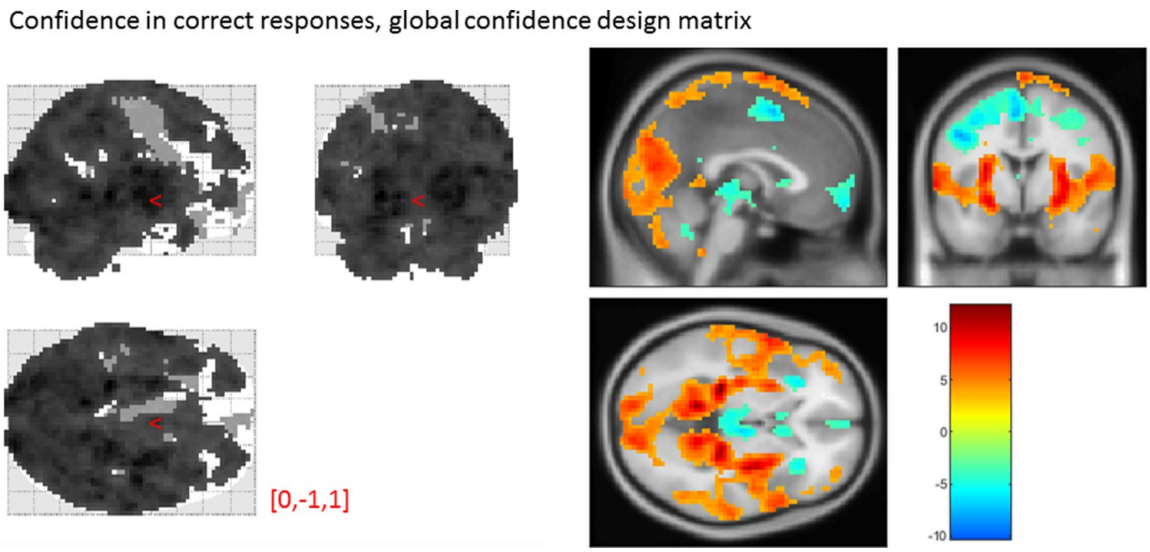


Figure E.3: Effect of confidence in correct responses, from the global-confidence design matrix. Uncorrected, thresholded at p<0.001. Left: glass brain visualization of the whole brain contrast. Right: yellow-red represent a positive correlation with subjective confidence ratings, and green-blue represent a negative correlation.

From our pre-specified ROIs, only the vmPFC and BA46 ROIs showed a significant linear effect of confidence in correct responses, in the opposite direction to what we expected based on previous studies. This is likely to be due to the differences in confidence profiles between the detection and discrimination tasks:

Average beta	T value	P value	Standard deviation	
vmPFC	-0.35	-3.06	4 × 10-3	0.67
pMFC	-0.31	-2.48	0.02	0.74
precuneus	0.25	2.30	0.03	0.64
ventral striatum	-0.056	-1.51	0.14	0.22
FPl	0.16	1.52	0.14	0.64
FPm	-0.12	-1.46	0.16	0.48
BA 46	0.37	3.77	6 × 10-4	0.57

E.4 Effect of confidence in our pre-specified ROIs

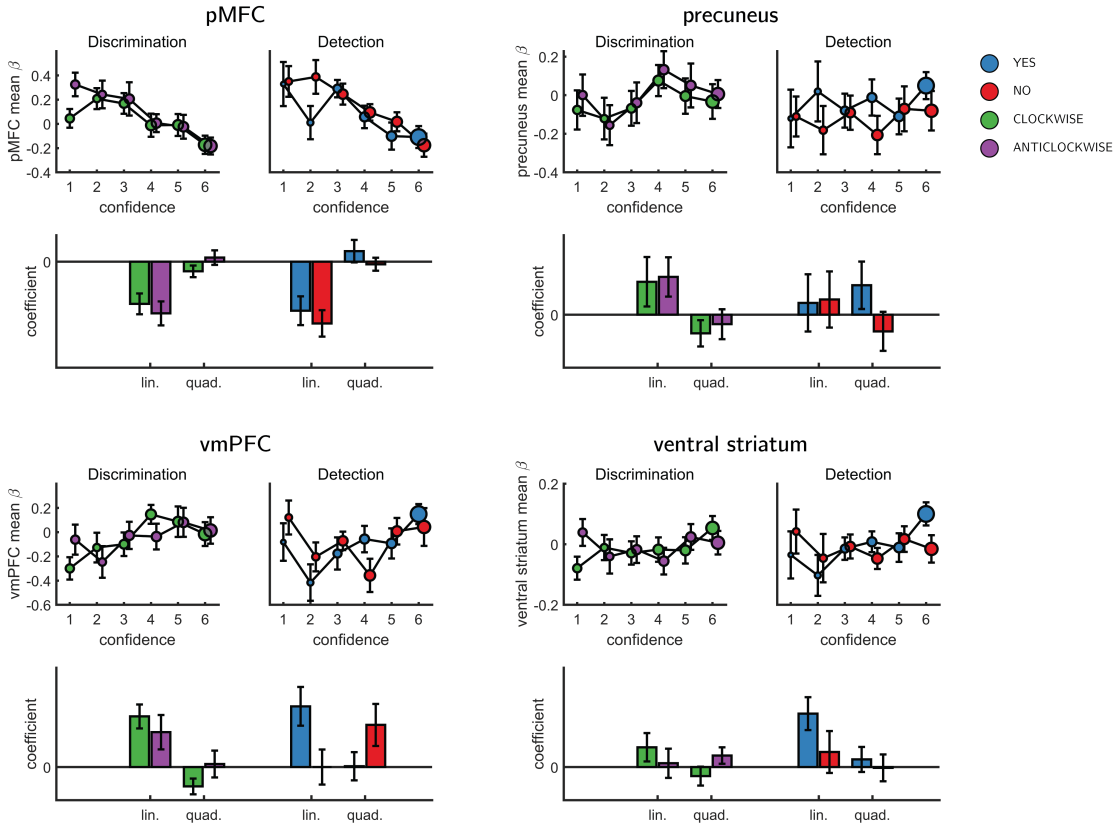


Figure E.4: Effect of confidence in all 4 ROIs, as a function of task and response, as extracted from the categorical design matrix. No significant interaction between the linear or quadratic effects and task or response was observed in any of the ROIs.

E.5 SDT variance ratio correlation with the quadratic confidence effect

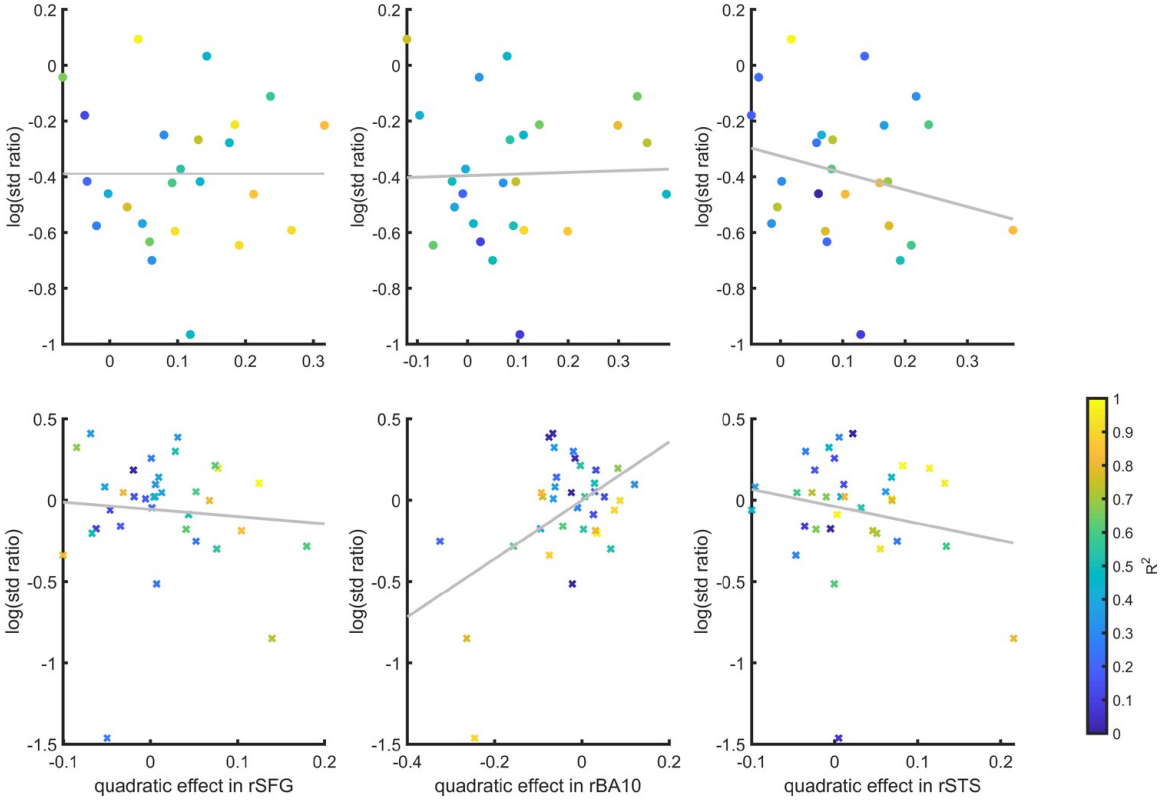


Figure E.5: Inter-subject correlation between the quadratic effect in the right hemisphere clusters and the ratio between the detection (top panel) and discrimination (lower panel) distribution variances, as estimated from the zROC curve slopes in the two tasks. Marker color indicates the goodness of fit of the second-order polynomial model to the BOLD data. All Spearman correlation coefficients are <0.25.

E.6 Correlation of metacognitive efficiency with linear and quadratic confidence effects

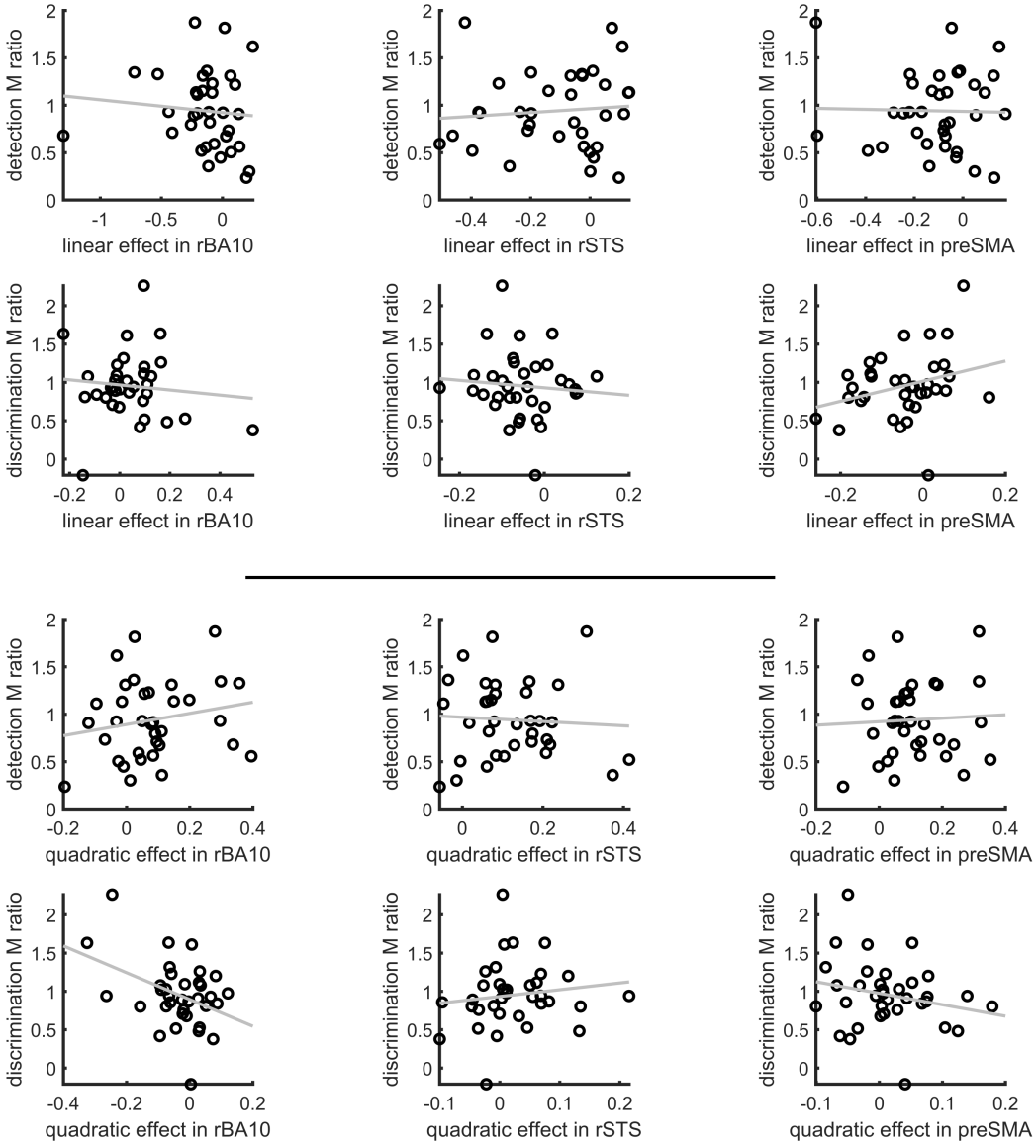


Figure E.6: Inter-subject correlation between the linear (upper panel) and quadratic (lower panel) effects in the right hemisphere clusters and metacognitive efficiency scores (measured as M ratio = meta-d'/d', Maniscalco and Lau, 2012).

E.7 Confidence-decision cross classification

In order to dissociate between brain regions that encode stimulus visibility and brain regions that encode decision confidence, we performed a multivariate cross-classification analysis. We trained a linear classifier on detection decisions ('yes' and 'no'), and tested it on discrimination confidence (high and low), and vice versa. Shared information content between detection responses and confidence in discrimination is expected in brain regions that encode stimulus visibility, rather than accuracy estimation. In detection, yes responses are associated with higher stimulus visibility compared to no responses (regardless of decision confidence), and in discrimination high confidence trials are associated with higher visibility than low confidence trials (regardless of subjective confidence).

Presented cross classification scores are the mean of cross classification accuracies in both directions. Detection-response and discrimination-confidence cross-classification was significantly above chance in in the pMFC ($t(29) = 2.76, p < 0.05$, corrected for family-wise error across the four ROIs), and in the BA46 anatomical subregion of the frontopolar ROI ($t(29) = 2.64, p < 0.05$, corrected).

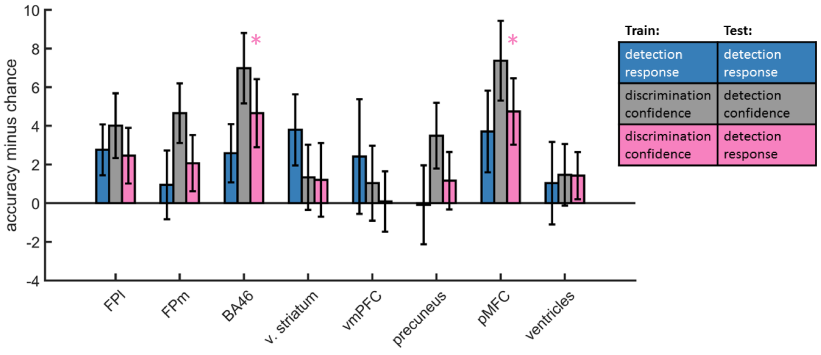


Figure E.7: Accuracy minus chance for classification of response in detection (yes vs. no; blue), and from a cross-classification between tasks: confidence in detection and confidence in discrimination (gray), and confidence in discrimination and decision in detection (pink).

E.8 Static Signal Detection Theory

E.8.1 Discrimination

Generative model

According to SDT, a decision variable x is sampled from one of two distributions on each experimental trial.

$$\mu_t = \begin{cases} 0.5, & \text{if cw.} \\ -0.5, & \text{if acw.} \end{cases} \tag{E.1}$$

$$x_t \sim \mathcal{N}(\mu_t, 1) \tag{E.2}$$

Inference

x is compared against a criterion to generate a decision about which of the two distributions was most likely, given the sample. For a discrimination task with symmetric distributions around 0, the optimal placement for a criterion is at 0.

$$decision_t = \begin{cases} \text{cw}, & \text{if } x_t > 0. \\ \text{acw}, & \text{else.} \end{cases} \tag{E.3}$$

In standard discrimination tasks, a common assumption is that the two distributions are Gaussian with equal variance. This assumption has a convenient computational consequence: the log-likelihood ratio (LLR), a quantity that reflects the degree to which the sample is more likely under one distribution or another, is linear with respect to x. Confidence is then assumed to be proportional to the distance of x_t from the decision criterion.

In what follows $\phi(x, \mu, \sigma)$ is the likelihood of observing x when sampling from a normal distribution with mean μ and standard deviation σ.

$$LLR = log(\phi(x_t, 0.5, 1)) - log(\phi(x_t, -0.5, 1)) \tag{E.4}$$

$$confidence_t \propto |x_t| \tag{E.5}$$

E.8.2 Detection

Generative model

A common assumption is that in detection the signal distribution is wider than the noise distribution [unequal-variance SDT; Wickens (2002), 48].

$$\mu_t = \begin{cases} 1.3, & \text{if P.} \\ 0, & \text{if A.} \end{cases} \tag{E.6}$$

$$\sigma_t = \begin{cases} 2, & \text{if P.} \\ 1, & \text{if A.} \end{cases} \tag{E.7}$$

$$x_t \sim \mathcal{N}(\mu_t, \sigma_t) \tag{E.8}$$

Inference

Here $med(x)$ represents the median sensory sample x. This criterion was chosen to ensure that detection responses are balanced.

$$decision = \begin{cases} \text{P}, & \text{if } x_t > med(x). \\ \text{A}, & \text{else.} \end{cases} \tag{E.9}$$

Importantly, in uv-SDT, LLR is quadratic in x.

$$LLR = log(\phi(x, 1.3, 2)) - log(\phi(x, 0, 1)) \tag{E.10}$$

$$confidence \propto |x_t - med(x)| \tag{E.11}$$

E.9 Dynamic Criterion

In SDT, task performance depends on the degree of overlap between the underlying distributions (d') and on the positioning of the decision criterion (c). Participants may optimize criterion placement based on their changing beliefs about the underlying distributions (Ko & Lau, 2012). To model this dynamic process of criterion setting we simulated a model where beliefs about the underlying distributions are the Maximum Likelihood Estimates of the mean and standard deviation, based on the last 5 samples that were (correctly or not) categorized.

E.9.1 Discrimination

Generative model

As in the Static Signal Detection model.

Inference

Means and standard deviations of the two distributions are estimated based on the last 5 samples in each category. To model prior beliefs about these parameters, each participant starts the task with 5 imaginary samples from the veridical distributions. Means and standard deviations are then extracted from these imaginary samples. In what follows, $\vec{cw}$ and $\vec{acw}$ are vectors with entries corresponding to the last 5 samples that were (correcly or not) labelled as 'cw' and 'acw,' respectively. $\bar{x}_{cw}$ and $\bar{x}_{acw}$

correspond to the sample means of these vectors. σ_{cw} and σ_{acw} correspond to their standard deviations.

$$LLR = log(\phi(x, \bar{x}_{cw}, \sigma_{cw})) - log(\phi(x, \bar{x}_{acw}, \sigma_{acw})) \tag{E.12}$$

Decisions and confidence are extracted from the LLR as in the Static Signal Detection model.

E.9.2 Detection

Generative model

As in the Static Signal Detection model.

Inference

As in discrimination. In what follows, $\vec{a}$ and $\vec{p}$ are vectors with entries corresponding to the last 5 samples that were (correcly or not) labelled as 'signal absent' and 'signal present,' respectively. $\bar{x}_a$ and $\bar{x}_p$ correspond to the sample means of these vectors. σ_a and σ_p correspond to their standard deviations.

$$LLR = log(\phi(x, \bar{x}_p, \sigma_p)) - log(\phi(x, \bar{x}_a, \sigma_a)) \tag{E.13}$$

In detection, $LLR = 0$ at two points (see figure @ref{fig:models). The decision criterion c_t is chosen to coincide with the rightmost point, which is positioned between the Signal and Noise distribution means.

$$decision = \begin{cases} \text{p}, & \text{if } x_t > c_t. \\ \text{a}, & \text{else.} \end{cases} \tag{E.14}$$

$$confidence \propto |LLR| \tag{E.15}$$

E.10 Attention Monitoring

Similar to the Dynamic Criterion model, in the Attention Monitoring model participants adjusts a decision criterion based on changing beliefs about the underlying distributions. However, unlike the Dynamic Criterion model, here beliefs change not as a function of recent perceptual samples, but as a function of access to an internal variable that represents the expected sensory precision (attention).

E.10.1 Discrimination

Generative model

In our schematic formulation of this model, participants have a true attentional state, which for simplicity we treat as either being on (1) or off (0). When attending,

participatns enjoy higher sensitivity than when they don't.

$$p(attended_t) = 0.5 \tag{E.16}$$

The attentional state determines the means of sensory distributions.

$$\mu_t = \begin{cases} 0.5, & \text{if cw and } \neg attended_t. \\ -0.5, & \text{if acw and } \neg attended_t. \\ 2, & \text{if cw and } attended_t. \\ -2, & \text{if acw and } attended_t. \end{cases} \tag{E.17}$$

$$x_t \sim \mathcal{N}(\mu_t, 1) \tag{E.18}$$

However, they don't have direct access to their attentional state, but only to a noisy approximation of the probability that they were attending.

$$onTask_t \sim \begin{cases} Beta(2,1), & \text{if } attended_t. \\ Beta(1,2), & \text{if } \neg attended_t. \end{cases} \tag{E.19}$$

Inference

Inference

Participants are then assumed to use their knowledge about the $onTask$ variable when making a decision and confidence estimate.

$$\begin{aligned} p(x_t|\text{cw}) &= p(attended_t|onTask_t)\phi(x_t, 2, 1) + p(\neg attended_t|onTask_t)\phi(x_t, 0.5, 1) \\ &= onTask_t\phi(x_t, 2, 1) + (1 - onTask_t)\phi(x_t, 0.5, 1) \end{aligned} \tag{E.20}$$

$$\begin{aligned} p(x_t|\text{acw}) &= p(attended_T|onTask_t)\phi(x_t, -2, 1) + p(\neg attended_t|onTask_t)\phi(x_t, -0.5, 1) \\ &= onTask_t\phi(x_t, -2, 1) + (1 - onTask_t)\phi(x_t, -0.5, 1) \end{aligned} \tag{E.21}$$

$$LLR = log(p(x_t|\text{w})) - log(p(x_t|\text{acw}) \tag{E.22}$$

$$decision_t = \begin{cases} \text{cw}, & \text{if } LLR > 0. \\ \text{acw}, & \text{else.} \end{cases} \tag{E.23}$$

$$confidence_t \propto |LLR| \tag{E.24}$$

E.10.2 Detection

Generative model

In detection, attentional states only affect the signal distribution, as noise is always centred at 0.

$$\mu_t = \begin{cases} 0, & \text{if a and } \neg attended_t. \\ 0.5, & \text{if p and } \neg attended_t. \\ 0, & \text{if a and } attended_t. \\ 2, & \text{if p and } attended_t. \end{cases} \tag{E.25}$$

$$x_t \sim \mathcal{N}(\mu_t, 1) \tag{E.26}$$

Inference

$$\begin{aligned} p(x_t|\text{p}) &= p(attended_t|onTask_t)\phi(x_t, 2, 1) + p(\neg attended_t|onTask_t)\phi(x_t, 0.5, 1) \\ &= onTask_t\phi(x_t, 2, 1) + (1 - onTask_t)\phi(x_t, 0.5, 1) \end{aligned} \tag{E.27}$$

The likelihood of observing x_t if no stimulus was presented is independent of the attention state.

$$\begin{aligned} p(x_t|\text{a}) &= p(attended_t|onTask_t)\phi(x_t, 0, 1)) + p(\neg attended_t|onTask_t)\phi(x_t, 0, 1) \\ &= \phi(x_t, 0, 1) \end{aligned} \tag{E.28}$$

$$LLR = log(p(x_t|\text{p})) - log(p(x_t|\text{a}) \tag{E.29}$$

$$decision_t = \begin{cases} \text{p}, & \text{if } LLR > 0. \\ \text{a}, & \text{else.} \end{cases} \tag{E.30}$$

Nevertheless, confidence in judgments about stimulus absence is dependent on beliefs about the attentional state. This is mediated by the effect of attention on the likelihood of observing x_t if a stimulus were present. This is the counterfactual part.

$$confidence_t \propto |LLR| \tag{E.31}$$

Appendix F

Supp. materials for ch. 5

F.1 Robustness Region

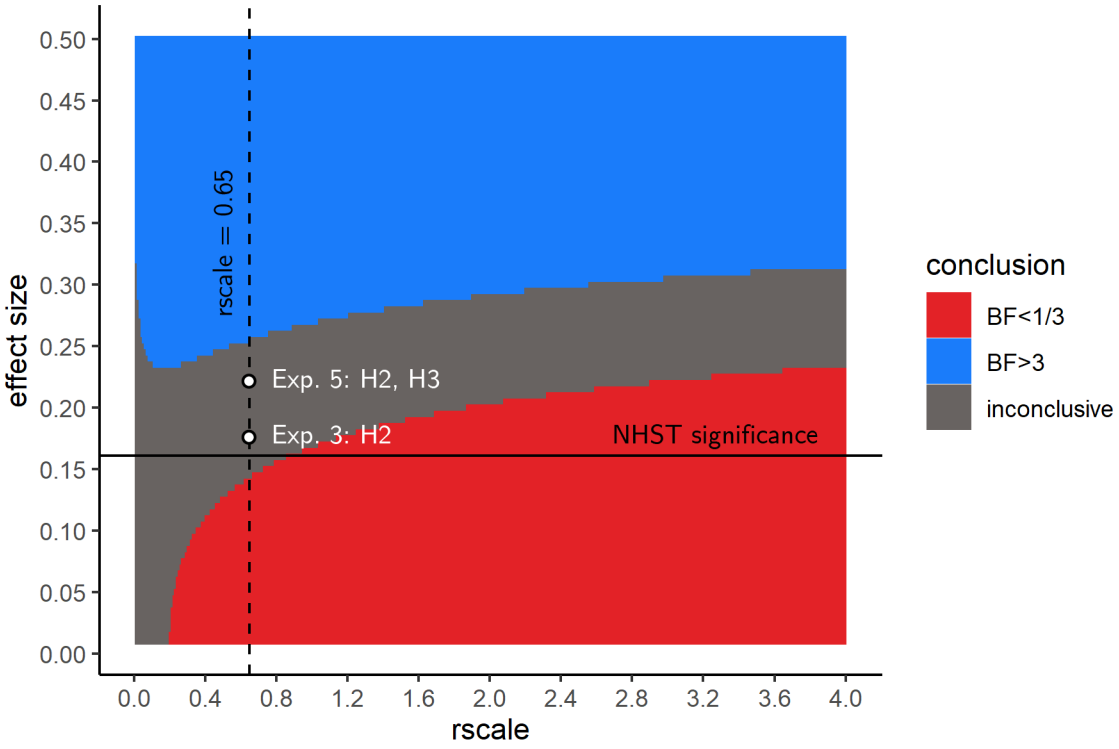


Figure F.1: A Robustness Region plot, visualizing Bayes Factors for hypothetical effect sizes and scale factors on the prior, for our sample size of 106 participants. Points above the horizontal line are significant in a one-tailed t-test. The dashed line indicates our choice of a scale factor on the prior.

Appendix G

Reproducibility receipt

```
## datetime
Sys.time()
```

```
[1] "2021-12-10 15:05:01 GMT"
```

```
## session info
sessionInfo()
```

```
R version 4.0.5 (2021-03-31)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19042)

Matrix products: default

locale:
[1] LC_COLLATE=English_United Kingdom.1252
[2] LC_CTYPE=English_United Kingdom.1252
[3] LC_MONETARY=English_United Kingdom.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.1252

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets
[7] methods   base

other attached packages:
 [1] scales_1.1.1           gsubfn_0.7
 [3] proto_1.0.0            pracma_2.3.3
 [5] shiny_1.6.0            bookdown_0.24
 [7] RColorBrewer_1.1-2     zoo_1.8-9
 [9] egg_0.4.5              gridExtra_2.3
```

```
[11] modeest_2.4.0          png_0.1-7
[13] ppcor_1.1              MASS_7.3-53.1
[15] lmerTest_3.1-3         lme4_1.1-27.1
[17] magrittr_2.0.1         knitr_1.34
[19] jsonlite_1.7.2         BayesFactor_0.9.12-4.2
[21] Matrix_1.3-2           coda_0.19-4
[23] brms_2.16.1            Rcpp_1.0.7
[25] pwr_1.3-0              lsr_0.5
[27] MESS_0.5.7             cowplot_1.1.1
[29] forcats_0.5.1          stringr_1.4.0
[31] dplyr_1.0.7            purrr_0.3.4
[33] readr_2.0.1            tidyr_1.1.3
[35] tibble_3.1.4           ggplot2_3.3.5
[37] tidyverse_1.3.1        reticulate_1.20
[39] papaja_0.1.0.9997      broom_0.7.9
[41] thesisdown_0.2.0.9000  remotes_2.4.0

loaded via a namespace (and not attached):
  [1] utf8_1.2.2           ggstance_0.3.5       tidyselect_1.1.1
  [4] htmlwidgets_1.5.4    munsell_0.5.0        codetools_0.2-18
  [7] ragg_1.2.0           DT_0.19              miniUI_0.1.1.1
 [10] withr_2.4.2          Brobdingnag_1.2-6    colorspace_2.0-2
 [13] rstudioapi_0.13      stats4_4.0.5         bayesplot_1.8.1
 [16] labeling_0.4.2       rstan_2.21.2         polyclip_1.10-0
 [19] bit64_4.0.5          farver_2.1.0         bridgesampling_1.1-2
 [22] fBasics_3042.89.1    rprojroot_2.0.2      vctrs_0.3.8
 [25] generics_0.1.0       afex_1.0-1           xfun_0.26
 [28] geepack_1.3-2        R6_2.5.1             markdown_1.1
 [31] clue_0.3-60          gamm4_0.2-6          projpred_2.0.2
 [34] assertthat_0.2.1     vroom_1.5.4          promises_1.2.0.1
 [37] gtable_0.3.0         processx_3.5.2       spatial_7.3-13
 [40] timeDate_3043.102    rlang_0.4.11         MatrixModels_0.5-0
 [43] systemfonts_1.0.3    splines_4.0.5        mosaicCore_0.9.0
 [46] checkmate_2.0.0      inline_0.3.19        yaml_2.2.1
 [49] reshape2_1.4.4       abind_1.4-5          modelr_0.1.8
 [52] threejs_0.3.3        crosstalk_1.1.1      backports_1.2.1
 [55] httpuv_1.6.3         rsconnect_0.8.24     tcltk_4.0.5
 [58] tensorA_0.36.2       tools_4.0.5          ellipsis_0.3.2
 [61] posterior_1.1.0      geeM_0.10.1          ggformula_0.10.1
 [64] stabledist_0.7-1     ggridges_0.5.3       plyr_1.8.6
 [67] base64enc_0.1-3      ps_1.6.0             prettyunits_1.1.1
 [70] rpart_4.1-15         pbapply_1.4-3        statip_0.2.3
 [73] cluster_2.1.1        haven_2.4.3          fs_1.5.0
 [76] here_1.0.1           timeSeries_3062.100  data.table_1.14.0
 [79] openxlsx_4.2.4       colourpicker_1.1.0   reprex_2.0.1
```

```
 [82] mvtnorm_1.1-2         matrixStats_0.60.1   hms_1.1.0
 [85] shinyjs_2.0.0         mime_0.11            evaluate_0.14
 [88] xtable_1.8-4          shinystan_2.5.0      rio_0.5.27
 [91] readxl_1.3.1          rstantools_2.1.1     compiler_4.0.5
 [94] V8_3.4.2              crayon_1.4.1         minqa_1.2.4
 [97] StanHeaders_2.21.0-7 htmltools_0.5.2      mgcv_1.8-34
[100] later_1.3.0           tzdb_0.1.2           RcppParallel_5.1.4
[103] lubridate_1.7.10      DBI_1.1.1            tweenr_1.0.2
[106] rmutil_1.1.5          dbplyr_2.1.1         rappdirs_0.3.3
[109] boot_1.3-27           car_3.0-11           cli_3.0.1
[112] parallel_4.0.5        igraph_1.2.6         pkgconfig_2.0.3
[115] numDeriv_2016.8-1.1   foreign_0.8-81       xml2_1.3.2
[118] dygraphs_1.1.1.6      rvest_1.0.1          distributional_0.2.2
[121] callr_3.7.0           digest_0.6.27        rmarkdown_2.10
[124] cellranger_1.1.0      curl_4.3.2           gtools_3.9.2
[127] nloptr_1.2.2.2        lifecycle_1.0.0      nlme_3.1-152
[130] carData_3.0-4         fansi_0.5.0          labelled_2.8.0
[133] pillar_1.6.2          lattice_0.20-41      loo_2.4.1
[136] fastmap_1.1.0         httr_1.4.2           pkgbuild_1.2.0
[139] glue_1.4.2            xts_0.12.1           zip_2.2.0
[142] shinythemes_1.2.0     bit_4.0.4            ggforce_0.3.3
[145] stringi_1.7.4         stable_1.1.4         textshaping_0.3.6
```

References

10 Adams, O. J., & Gaspelin, N. (2020). Assessing introspective awareness of attention capture. *Attention, Perception, & Psychophysics*, 1–13.

Adams, O. J., & Gaspelin, N. (2021). Introspective awareness of oculomotor attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, *2*(2), 284–299.

Albonico, A., Furubacke, A., Barton, J. J., & Oruc, I. (2018). Perceptual efficiency and the inversion effect for faces, words and houses. *Vision Research*, *153*, 91–97.

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *Neuroimage*, *13*(5), 903–919.

Angel, E. (2000). *Interactive computer graphics : A top-down approach with OpenGL*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with QuickTime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, *26*(3), 839–851.

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological*

Bulletin, *74*(2), 81.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, *76*, 412–427.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*(2), 297–305.

Blackwell, H. R. (1952). Studies of psychophysical methods for measuring visual thresholds. *JOSA*, *42*(9), 606–616.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, *1*(7), 635–640.

Bonawitz, E., Ullman, T. D., Bridgers, S., Gopnik, A., & Tenenbaum, J. B. (2019). Sticking to the evidence? A behavioral and computational case study of micro-theory change in the domain of magnetism. *Cognitive Science*, *43*(8), e12765.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*(5), 733–743.

Borchers, H. W. (2019). *Pracma: Practical numerical math functions*. Retrieved from https://CRAN.R-project.org/package=pracma

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel'touch that eyes see. *Nature*, *391*(6669), 756–756.

Brown, J., Lewis, V., & Monk, A. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology*, *29*(3), 461–473.

Burgess, P. W., Gilbert, S. J., & Dumontheil, I. (2007). Function and localization within rostral prefrontal cortex (area 10). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 887–899.

Calder-Travis, J., Charles, L., Bogacz, R., & Yeung, N. (2020). Bayesian confidence in optimal decisions.

Cassini, M. H., Kacelnik, A., & Segura, E. T. (1990). The tale of the screaming hairy armadillo, the guinea pig and the marginal value theorem. *Animal Behaviour*, *39*(6), 1030–1050.

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through

adaptive experimentation. *Psychonomic Bulletin & Review*, *18*(1), 204–210.

Champely, S. (2020). *Pwr: Basic functions for power analysis*. Retrieved from https://CRAN.R-project.org/package=pwr

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Christensen, M. S., Ramsøy, T. Z., Lund, T. E., Madsen, K. H., & Rowe, J. B. (2006). An fMRI study of the neural correlates of graded visual perception. *Neuroimage*, *31*(4), 1711–1725.

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.

Clark, A. (2013). The many faces of precision (replies to commentaries on "whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Psychology*, *4*, 270.

Coldren, J. T., & Haaf, R. A. (2000). Asymmetries in infants' attention to the presence or absence of features. *The Journal of Genetic Psychology*, *161*(4), 420–434.

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory*, *18*(6), 595–609.

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *Quarterly Journal of Experimental Psychology*, *68*(8), 1489–1501.

Cowie, D., Makin, T. R., & Bremner, A. J. (2013). Children's responses to the rubber-hand illusion reveal dissociable pathways in body representation. *Psychological Science*, *24*(5), 762–769.

Cowie, R. J. (1977). Optimal foraging in great tits (parus major). *Nature*, *268*(5616), 137–139.

D'Zmura, M. (1991). Color in visual search. *Vision Research*, *31*(6), 951–966.

Darwin, C., & Darwin, F. (1958). *Autobiography and selected letters* (Vol. 479). Courier Corporation.

De Cornulier, B. (1988). Knowing whether, knowing who, and epistemic closure. *Questions and Questioning*, 182–192.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. *Journal of Neuroscience*, *37*(25), 6066–6074.

Denecker, M., Marek, V. W., & Truszczynski, M. (2011). Reiter's default logic is a logic

of autoepistemic reasoning and a good one, too. *arXiv Preprint arXiv:1108.3278*.

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, *115*(43), 11090–11095.

Dienes, Z. (2019). How do i know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377.

Domenech, P., & Koechlin, E. (2015). Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, *1*, 101–106.

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486.

Dosher, B. A., Han, S., & Lu, Z.-L. (2004). Parallel processing in visual search asymmetry. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(1), 3.

Dugué, L., Merriam, E. P., Heeger, D. J., & Carrasco, M. (2018). Specific visual subregions of TPJ mediate reorienting of spatial attention. *Cerebral Cortex*, *28*(7), 2375–2390.

Ehinger, K. A., & Wolfe, J. M. (2016). When is it time to move to the next map? Optimal foraging in guided visual search. *Attention, Perception, & Psychophysics*, *78*(7), 2135–2151.

Ekstrøm, C. T. (2019). *MESS: Miscellaneous esoteric statistical scripts*. Retrieved from https://CRAN.R-project.org/package=MESS

Farennikova, A. (2013). Seeing absence. *Philosophical Studies*, *166*(3), 429–454.

Farennikova, A. (2015). Perception of absence and penetration from expectation. *Review of Philosophy and Psychology*, *6*(4), 621–640.

Fechner, G. T., & Adler, H. E. (1860). Elements of psychophysics [elemente der psychophysik]. *Leipzig, Germany: Breitkopf and Ha Rtel*.

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*(18), 6117–6125.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Fleming, S. M., Van Der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, *21*(4), 617.

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

Forrester, J. W. (1971). Counterintuitive behavior of social systems. *Theory and Decision*, *2*(2), 109–140.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213–2223.

Frith, U. (1974). Acurious effect with reversed letters explained by a theory of schema. *Perception & Psychophysics*, *16*(1), 113–116.

Gandolfo, M., & Downing, P. E. (2020). Asymmetric visual representation of sex from human body shape. *Cognition*, *205*, 104436.

Geng, J. J., & Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: Contextual updating? *Neuroscience & Biobehavioral Reviews*, *37*(10), 2608–2620.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. *Oxford Handbook of Causal Reasoning*, 515–548.

Gherman, S., & Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *Elife*, *7*, e38293.

Ghetti, S., & Alexander, K. W. (2004). "If it happened, i would remember it": Strategic use of event memorability in the rejection of false autobiographical events. *Child Development*, *75*(2), 542–561.

Ghetti, S., Castelli, P., & Lyons, K. E. (2010). Knowing about not remembering: Developmental dissociations in lack-of-memory monitoring. *Developmental Science*, *13*(4), 611–621.

Ghetti, S., Lyons, K. E., Lazzarin, F., & Cornoldi, C. (2008). The development of metamemory monitoring during retrieval: The case of memory strength and memory absence. *Journal of Experimental Child Psychology*, *99*(3), 157–181.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,

16(1), 5.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 21.

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*(3), 431–455.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory.

Gow, L. (2021). A new theory of absence experience. *European Journal of Philosophy*, *29*(1), 168–181.

Graziano, M. S. (2013). *Consciousness and the social brain*. Oxford University Press.

Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*, 500.

Greene, R. L., & Thapar, A. (1994). Mirror effect in frequency discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 946.

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *Elife*, *5*, e13388.

Guttentag, R., & Carroll, D. (1998). Memorability judgments for high-and low-frequency words. *Memory & Cognition*, *26*(5), 951–958.

Haarsma, J., Fletcher, P. C., Ziauddeen, H., Spencer, T. J., Diederen, K. M., & Murray, G. K. (2018). Precision weighting of cortical unsigned prediction errors is mediated by dopamine and benefits learning. *bioRxiv*, 288936.

He, Y., Chen, P., & Li, Y. (2020). New efficient and practicable adaptive designs for calibrating items online. *Applied Psychological Measurement*, *44*(1), 3–16.

Hearst, E. (1991). Psychology and nothing. *American Scientist*, *79*(5), 432–443.

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The decoding toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88.

Helmholtz, H. von. (1948). Concerning the perceptions in general, 1867. In *Readings in the history of psychology* (pp. 214–230). East Norwalk, CT, US: Appleton-Century-Crofts. http://doi.org/10.1037/11304-027

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186.

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2019). Lab. Js: A free, open, online study builder.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 57.

Hoffman, J. E. (1979). A two-stage model of visual search. *Perception & Psychophysics*, *25*(4), 319–327.

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*, 1155–1167.

Hsu, A. S., Martin, J. B., Sanborn, A. N., & Griffiths, T. L. (2019). Identifying category representations for complex stimuli using discrete markov chain monte carlo with people. *Behavior Research Methods*, *51*(4), 1706–1716.

Hulleman, J., & Olivers, C. N. (2017). The impending demise of the item in visual search. *Behavioral and Brain Sciences*, *40*.

Igelström, K. M., Webb, T. W., & Graziano, M. S. (2015). Neural processes in the human temporoparietal cortex separated by localized independent component analysis. *Journal of Neuroscience*, *35*(25), 9432–9445.

Igelström, K. M., Webb, T. W., Kelly, Y. T., & Graziano, M. S. (2016). Topographical organization of attentional, social, and memory processes in the human temporoparietal cortex. *Eneuro*, *3*(2).

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2013). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482.

Jannati, A., & Di Lollo, V. (2012). Relative blindsight arises from a criterion confound in metacontrast masking: Implications for theories of consciousness. *Consciousness and Cognition*, *21*(1), 307–314.

Kahneman, D. (1968). Method, findings, and theory in studies of visual masking. *Psychological Bulletin*, *70*(6p1), 404.

Kammers, M. P., Vignemont, F. de, Verhagen, L., & Dijkerman, H. C. (2009). The rubber hand illusion in action. *Neuropsychologia*, *47*(1), 204–211.

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.

Kay, K. N., & Yeatman, J. D. (2017). Bottom-up and top-down computations in word-and face-selective cortex. *Elife*, *6*, e22341.

Kellij, S., Fahrenfort, J., Lau, H., Peters, M. A., & Odegaard, B. (2018). The foundations of introspective access: How the relative precision of target encoding influences metacognitive performance.

Kellij, S., Fahrenfort, J., Lau, H., Peters, M. A., & Odegaard, B. (2021). An investigation of how relative precision of target encoding influences metacognitive performance. *Attention, Perception, & Psychophysics*, *83*(1), 512–524.

King, J.-R., & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130204.

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 B: Biological Sciences*, *367*(1594), 1401–1411.

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics*, *77*(4), 1295–1306.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behaviour*, *22*, 953–IN3.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. http://doi.org/10.18637/jss.v082.i13

Labes, D., Schütz, H., Lang, B., & Labes, M. D. (2020). Package 'PowerTOST.' *Power*, *2*, 49.

Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Lakens, D. (2016, January). *Power analysis for default Bayesian t-tests*. Retrieved from http://daniellakens.blogspot.com/2016/01/power-analysis-for-default-bayesian-t.html

Lange, K., Kuhn, S., & Filevich, E. (2015). Just another tool for online studies (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6), e0130834.

Lange, K., Kühn, S., & Filevich, E. (2015). " just another tool for online stud-

ies"(JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6).

Langlois, T. A., Jacoby, N., Suchow, J. W., & Griffiths, T. L. (2021). Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences*, *118*(13).

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159.

Leckey, S., Selmeczy, D., Kazemi, A., Johnson, E. G., Hembacher, E., & Ghetti, S. (2020). Response latencies and eye gaze provide insight on how toddlers gather evidence under uncertainty. *Nature Human Behaviour*, *4*(9), 928–936.

Lee, S. M., & McCarthy, G. (2016). Functional heterogeneity and convergence in the right temporoparietal junction. *Cerebral Cortex*, *26*(3), 1108–1116.

Levin, D. T., & Angelone, B. L. (2001). Visual search for a socially defined feature: What causes the search asymmetry favoring cross-race faces? *Perception & Psychophysics*, *63*(3), 423–435.

Levin, D. T., & Angelone, B. L. (2008). The visual metacognition questionnaire: A measure of intuitions about vision. *The American Journal of Psychology*, 451–472.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477.

Limanowski, J., & Friston, K. (2018). 'Seeing the dark': Grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in Psychology*, *9*, 643.

Locke, J. (1836). An essay concerning human understanding (bk. iv. *Chap. XXVII.*

Malinowski, P., & Hübner, R. (2001). The effect of familiarity on visual-search performance: Evidence for learned basic features. *Perception & Psychophysics*, *63*(3), 458–463.

Maniscalco, B., & Lau, H. (2010). Comparing signal detection models of perceptual decision confidence. *Journal of Vision*, *10*(7), 213–213.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*(3), 923–937.

Marois, R., Yi, D.-J., & Chun, M. M. (2004). The neural fate of consciously perceived and missed events in the attentional blink. *Neuron*, *41*(3), 465–472.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

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

Mazor, M., & Fleming, S. (2021). Zero-shot search termination reveals a dissociation between implicit and explicit metacognitive knowledge.

Mazor, M., & Fleming, S. M. (2020). Distinguishing absence of awareness from awareness of absence. *Philosophy and the Mind Sciences*, *1*(II).

Mazor, M., Friston, K. J., & Fleming, S. M. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. *Elife*, *9*, e53900.

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.

Mazor, M., Moran, R., & Fleming, S. (2021). Stage 2 registered report: Metacognitive asymmetries in visual perception.

McCarthy, L. (2015). p5. js. *URL: Https://P5js. Org*, *3*.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., De Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*(5), 1897–1906.

Merkle, E. C., & Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*(3), 391.

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*(4), 353–393.

Meuwese, J. D., Loon, A. M. van, Lamme, V. A., & 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.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92.

Miller, P. H., & Bigi, L. (1977). Children's understanding of how stimulus dimensions affect performance. *Child Development*, 1712–1715.

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review*, *127*(5), 655.

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, *38*(14), 3534–3546.

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.

Moran, R., Zehetleitner, M., Liesefeld, H. R., Müller, H. J., & Usher, M. (2016). Serial vs. Parallel models of attention in visual search: Accounting for benchmark RT-distributions. *Psychonomic Bulletin & Review*, *23*(5), 1300–1315.

Moran, R., Zehetleitner, M., Müller, H. J., & Usher, M. (2013). Competitive guided search: Meeting the challenge of benchmark RT distributions. *Journal of Vision*, *13*(8), 24–24.

Morey, Richard D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Morey, Richard D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor.' *URLh Http://Cran/r-Projectorg/Web/Packages/BayesFactor/BayesFactor Pdf i (Accessed 1006 15)*.

Navarro, D. (2015). *Learning statistics with r: A tutorial for psychology students and other beginners. (Version 0.5)*. Adelaide, Australia: University of Adelaide. Retrieved from http://ua.edu.au/ccs/teaching/lsr

Neubert, F.-X., Mars, R. B., Thomas, A. G., Sallet, J., & Rushworth, M. F. (2014). Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron*, *81*(3), 700–713.

Newman, J. P., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 630.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Oaksford, M. (2002). Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking & Reasoning*, *8*(2), 135–151.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Oaksford, M., & Hahn, U. (2004). A bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *58*(November 2015), 75–85. http://doi.org/10.1037/h0085798

Odegaard, B., Chang, M. Y., Lau, H., & Cheung, S.-H. (2018). Inflation versus filling-in: Why we feel we see more than we actually do in peripheral vision. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170345.

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between JSON data and r objects. *arXiv:1403.2805 [Stat.CO]*. Retrieved from https://arxiv.org/abs/1403.2805

Palmer, C. E., Auksztulewicz, R., Ondobaka, S., & Kilner, J. M. (2019). Sensorimotor beta power reflects the precision-weighting afforded to sensory prediction errors. *NeuroImage*.

Papeo, L., & Vega, M. de. (2020). The neurobiology of lexical and sentential negation. In *The oxford handbook of negation*.

Parr, T., Benrimoh, D. A., Vincent, P., & Friston, K. J. (2018). Precision and false perceptual inference. *Frontiers in Integrative Neuroscience*, *12*.

Parr, T., & Friston, K. J. (2019). Attention or salience? *Current Opinion in Psychology*, *29*, 1–5.

Peters, M. A., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., … others. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 1–8.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., De Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*(12), 1513–1515.

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.

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*(1-2), 81–132.

Robinson, D., & Hayes, A. (2020). *Broom: Convert statistical analysis objects into tidy tibbles*. Retrieved from https://CRAN.R-project.org/package=broom

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1), 1–11.

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry*, *84*(6), 443–451.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Rust, N. C., & Mehrpour, V. (2020). Understanding image memorability. *Trends in Cognitive Sciences*.

Rutishauser, U., Aflalo, T., Rosario, E. R., Pouratian, N., & Andersen, R. A. (2018). Single-neuron representation of memory strength and recognition confidence in left human posterior parietal cortex. *Neuron*, *97*(1), 209–220.

Saiki, J. (2008). Stimulus-driven mechanisms underlying visual search asymmetry revealed by classification image analyses. *Journal of Vision*, *8*(4), 30–30.

Sainsbury, R. (1971). The "feature positive effect" and simultaneous discrimination learning. *Journal of Experimental Child Psychology*, *11*(3), 347–356.

Samaha, J., & Denison, R. (2020). The positive evidence bias in perceptual confidence is not post-decisional. *bioRxiv*.

Sanborn, A., & Griffiths, T. L. (2008). Markov chain monte carlo with people. In *Advances in neural information processing systems* (pp. 1265–1272).

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–1399.

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241–1260.

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *Elife*, *9*, e60705.

Shen, J., & Reingold, E. M. (2001). Visual search asymmetry: The influence of stimulus familiarity and low-level features. *Perception & Psychophysics*, *63*(3), 464–475.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Shulman, G. L., Astafiev, S. V., McAvoy, M. P., d'Avossa, G., & Corbetta, M. (2007). Right TPJ deactivation during visual search: Functional significance and support for a filter hypothesis. *Cerebral Cortex*, *17*(11), 2625–2633.

Siegel, M. H., Magid, R. W., Pelz, M., Tenenbaum, J. B., & Schulz, L. E. (2021). Children's exploratory play tracks the discriminability of hypotheses. *Nature Communications*, *12*(1), 1–9.

Simons, J. S., Davis, S. W., Gilbert, S. J., Frith, C. D., & Burgess, P. W. (2006). Discriminating imagined from perceived information engages brain areas implicated in schizophrenia. *Neuroimage*, *32*(2), 696–703.

Sladky, R., Friston, K. J., Tröstl, J., Cunnington, R., Moser, E., & Windischberger, C. (2011). Slice-timing effects and their correction in functional MRI. *Neuroimage*, *58*(2), 588–594.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

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.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition*, *40*(8), 1189–1199.

Stein, T., & Peelen, M. V. (2021). Dissociating conscious and unconscious influences on visual detection effects. *Nature Human Behaviour*, 1–13.

Strack, F., Förster, J., & Werth, L. (2005). "Know thyself!" The role of idiosyncratic self-knowledge in recognition memory. *Journal of Memory and Language*, *52*(4), 628–638.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1379.

Takeda, Y., & Yagi, A. (2000). Inhibitory tagging in visual search can be found if search stimuli remain visible. *Perception & Psychophysics*, *62*(5), 927–934.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Treisman, A. (1986). Features and objects in visual processing. *Scientific American*,

255(5), 114B–125.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*(1), 15.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 459.

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, *114*(3), 285.

Tsakiris, M., & Haggard, P. (2005). The rubber hand illusion revisited: Visuotactile integration and self-attribution. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(1), 80.

Turner, M. S., Simons, J. S., Gilbert, S. J., Frith, C. D., & Burgess, P. W. (2008). Distinct roles for lateral and medial rostral prefrontal cortex in source monitoring of perceived and imagined events. *Neuropsychologia*, *46*(5), 1442–1453.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Ushey, K., Allaire, J., & Tang, Y. (2020). *Reticulate: Interface to 'python'*. Retrieved from https://github.com/rstudio/reticulate

Vallesi, A. (2014). Monitoring mechanisms in visual search: An fMRI study. *Brain Research*, *1579*, 65–73.

Vincent, B. T. (2011). Search asymmetries: Parallel processing of uncertain sensory information. *Vision Research*, *51*(15), 1741–1750.

Von Grünau, M., & Dubé, S. (1994). Visual search asymmetry for viewing direction. *Perception & Psychophysics*, *56*(2), 211–220.

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

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception & Psychophysics*, *56*(5), 495–500.

Watanabe, A., Grodzinski, U., & Clayton, N. S. (2014). Western scrub-jays allocate longer observation time to more valuable information. *Animal Cognition*, *17*(4), 859–867.

Webb, T., Miyoshi, K., So, T. Y., & Lau, H. (2021). A task-optimized neural network model of decision confidence. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of*

Psychology, *69*, 105–129.

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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2020). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wilke, C. O. (2019). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=cowplot

Wilterson, A. I., Kemper, C. M., Kim, N., Webb, T. W., Reblando, A. M., & Graziano, M. S. (2020). Attention control and the attention schema theory of consciousness. *Progress in Neurobiology*, *195*, 101844.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 681.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152.

Wokke, M. E., Cleeremans, A., & Ridderinkhof, K. R. (2017). Sure i'm sure: Prefrontal oscillations support metacognitive monitoring of decision making. *Journal of Neuroscience*, *37*(4), 781–789.

Wolfe, J., & Horowitz, T. S. (2008). Visual search. *Scholarpedia*, *3*(7), 3325.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, *9*(1), 33–39.

Wolfe, J. M. (2001). Asymmetries in visual search: An introduction. *Perception & Psychophysics*, *63*(3), 381–389.

Wolfe, J. M. (2012). When do i quit? The search termination problem in visual search. *The Influence of Attention, Learning, and Motivation on Visual Search*, 183–208.

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 1–33.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419.

Wolfe, J. M., & Gray, W. (2007). Guided search 4.0. *Integrated Models of Cognitive Systems*, 99–119.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 1–8.

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. *Vision Research*, *50*(14), 1304–1311.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Xue, G., Chen, C., Jin, Z., & Dong, Q. (2006). Language experience shapes fusiform activation when processing a logographic artificial language: An fMRI training study. *Neuroimage*, *31*(3), 1315–1326.

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., ... Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neuroscience Research*, *68*(3), 199–206.

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition*, *5*(4), 418–441.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*(24), 2256–2262.

Zhang, Y. R., & Onyper, S. (2020). Visual search asymmetry depends on target-distractor feature similarity: Is the asymmetry simply a result of distractor rejection speed? *Attention, Perception, & Psychophysics*, *82*(1), 80–97.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*, 79.