Self-Modeling in Inference about Absence

A Thesis
Presented to
The Division of 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

April 2021

Approved for the Division
(Brain Sciences)

Stephen M. Fleming	Karl J. Friston

Acknowledgements

Rani, Roni, Halely, Dina, Karl, Josh, Alisa, Lesley, B7, SIPS & Neurohackademy (Ariel and Tal), Steve, Noam, scihub, Roy, Chudi, Dan, Peter Zeidman, Rani, Dominic and the Hari Krishna volunteers, my participants, imaging team, Kamlyn, Maddi and Shiv, the developers of JATOS (Elisa and Kristian), labjs (Felix), thesisdown (Chester Ismay) and the numerous other pacakges that I've used, stackoverflow community,

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

Introduction **1**
0.1 Inference about absence 2
0.2 Probabilistic reasoning, criterion setting, and self knowledge 3
Symmetrical definition: 3
Dissymmetrical definition: 3
0.2.1 Second-order cognition 4
0.2.2 Computational models of detection 6
The High-Threshold model 7
Signal Detection Theory 8
0.3 Detection: "I would have noticed it" 9
0.4 Visual search: "I would have found it" 12
0.5 Memory: "I would have remembered it" 16
0.6 The development of a self-model 18
0.7 This thesis 19

Chapter 1: Zero-shot search termination reveals a dissociation between implicit and explicit metacognitive knowledge **21**
1.1 Introduction 21
1.2 Experiment 1 22
1.2.1 Participants 23
1.2.2 Procedure 23
1.2.3 Data analysis 25
1.2.4 Results 27
1.3 Experiment 2 29
1.3.1 Participants 30
1.3.2 Procedure 30
1.3.3 Results 30
1.4 Discussion 34
1.4.1 Is implicit metacognitive knowledge metacognitive? 35
1.4.2 Inference about absence as a tool for studying implicit self knowledge 36
1.4.3 Conclusion 36

Chapter 2: Prospective search time estimates for unseen displays reveal a rich intuitive theory of visual search **37**

2.1 Introduction . . . 37
2.2 Experiments 1 and 2: shape, orientation, and color . . . 38
2.2.1 Participants . . . 39
2.2.2 Procedure . . . 39
2.2.3 Results . . . 41
2.3 Experiments 3 and 4: complex, unfamiliar stimuli . . . 43
2.3.1 Participants . . . 44
2.3.2 Procedure . . . 44
2.3.3 Results . . . 45

Chapter 3: Distinct neural contributions to metacognition for detecting (but not discriminating) visual stimuli . . . 51
3.1 Introduction . . . 51
3.2 Methods and Materials . . . 53
3.2.1 Participants . . . 53
3.2.2 Design and procedure . . . 54
3.2.3 Scanning parameters . . . 56
3.2.4 Analysis . . . 56
3.2.5 Exclusion criteria . . . 57
3.2.6 Response conditional type-II ROC curves . . . 57
3.2.7 Imaging analysis . . . 58
3.2.8 Statistical inference . . . 62
3.3 Results . . . 62
3.4 Behavioural results . . . 62
3.4.1 Imaging results . . . 64
3.4.2 Computational models . . . 69
3.5 Discussion . . . 72

Chapter 4: Paradoxical evidence weightings in confidence judgments for detection and discrimination . . . 77
4.1 Introduction . . . 77
4.2 Experiment 1 . . . 78
4.2.1 Methods . . . 78
4.2.2 Analysis . . . 80
4.2.3 Results . . . 81
4.3 Experiment 2 . . . 89
4.3.1 Methods . . . 90
4.3.2 Results . . . 92
4.3.3 Detection signal trials . . . 98
4.4 Discussion . . . 99
4.4.1 Model 1: a rational agent + symmetric evidence structure . . . 99
4.4.2 Model 2: a rational agent + symmetric evidence structure . . . 101
4.4.3 Model 3: confidence decision cross . . . 103
4.4.4 Evidence for absence . . . 104

General Discussion . . . **107**
4.5 Summary of results . . . 107
4.6 What I didn't find . . . 107
4.6.1 Chapter 1: no correlation with explicit metacognition . . . 107
4.6.2 Chapter 2: no effect of confidence in signal presence . . . 107
4.6.3 Chapter 3: small differences in brain activity between inference about absence and presence . . . 107
4.6.4 . . . 107
4.7 Future directions . . . 107
4.7.1 Failures of a self-model . . . 108
4.8 Conclusion . . . 108

Appendix A: Signal Detection Theory . . . **109**
A.1 ROC and zROC curves . . . 110
A.2 Unequal-variance (uv) SDT . . . 112
A.3 SDT Measures for Metacognition . . . 113

Appendix B: Supp. materials for ch. 2 . . . **115**
B.1 Pseudo-discrimination analysis . . . 115
B.1.1 Exp. 1 . . . 116
B.1.2 Exp. 2 . . . 117
B.2 Unequal-variance model . . . 118
B.2.1 Discrimination . . . 118
B.2.2 Detection . . . 120

Appendix C: Supp. materials for ch. 3 . . . **123**
C.1 Confidence button presses . . . 124
C.2 zROC curves . . . 125
C.3 Global confidence design matrix . . . 126
C.4 Effect of confidence in our pre-specified ROIs . . . 127
C.5 SDT variance ratio correlation with the quadratic confidence effect . . . 128
C.6 Correlation of metacognitive efficiency with linear and quadratic confidence effects . . . 129
C.7 Confidence-decision cross classification . . . 130
C.8 Static Signal Detection Theory . . . 130
C.8.1 Discrimination . . . 130
C.8.2 Detection . . . 131
C.9 Dynamic Criterion . . . 132
C.9.1 Discrimination . . . 132
C.9.2 Detection . . . 133
C.10 Attention Monitoring . . . 133
C.10.1 Discrimination . . . 133
C.10.2 Detection . . . 134

References . . . **137**

List of Tables

3.1 List of regressors in the main design matrix (DM-1). 59

A.1 SDT response classification. 110

List of Figures

1 Guavas 2

2 A symmetric implementation of a predator-detector. 4

3 An asymmetric implementation of a predator-detector. 5

4 An asymmetric implementation of a predator-detector with a pessimistic prior. 6

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. 'No' responses are controlled by the parameter *g* - a 'guessing parameter' that determines the probability of responding 'yes' in case no stimulus was detected. 7

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 no in 'target-absent' trials, the sensory sample falls far away from the decision criterion. 9

7 Task design for Meuwese et al (2014). 10

8 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. 12

9 Models of search termination 14

10 Upper panel: A target that is marked by a unique colour imemdiately captures attention (left). This fact is available to particiapnts' 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 canincally 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). 16

1.1 Experimental design for Exp. 1 24

1.2 Visualization of Hypotheses. Top left: typical search time results 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 of target-absent search also holds in the first trials in an experiment. Different models make different predictions about target-absent serach times in the first block of the experiment. Top right: one possible pattern 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 metacognitive 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 metacognitive 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. 26

1.3 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. 28

1.4 Upper panel: median search time by distractor set size for the two search tasks across the three blocks. 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. 32

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. 40
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 $y = x$ and the dotted line indicates $y = 2x$. 43
2.3 Stimuli used for Experiments 3 and 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 clockface. 45
2.4 Estimated search times plotted against true search times in Experiment 2. The dashed line indicates $y = x$ and the dotted line indicates $y = 2x$. 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). 47
2.5 Median estimated search times plotted against true search times in Experiment 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. 48

3.1 Experimental design, imaging experiment 55
3.2 Behavioural results, imaging experiment 63
3.3 Univariate parametric effect of confidence 65
3.4 Effect of confidence in the frontopolar cortex 67
3.5 Quadratic effect of confidence 69
3.6 Computational models, imaging experiment 70

4.1 Experimental design for Exp. 1 80
4.2 Response time and Confidence histograms for Experiment 1 82
4.3 Response conditional ROC curves for Experiment 1 83

4.4 Reverse correlation of discrimination trials, Exp. 1 85
4.5 Reverse correlation of detection trials, Exp. 1 87
4.6 Reverse correlation of detection signal trials, Exp. 1 89
4.7 Experimental design for Exp. 2 91
4.8 Response time and confidence distributions, Exp. 2 93
4.9 Response conditional ROC curves for Experiment 2. 94
4.10 Decision kernels in discrimination, Exp. 2 95
4.11 Decision kernels in detection, Exp. 2 97
4.12 Decision kernels in detection signal trials, Exp. 2 98
4.13 Simulation results: Model 1 101
4.14 Simulation results: Model 2 103
4.15 Simulation results: Model 2 104

A.1 Signal Detection Theory 109
A.2 Receiver Operative Characteristic (ROC) curve 111
A.3 zROC curve 112
A.4 A second order SDT model 114

B.1 Pseudo-discrimination kernels for detection signal trials. 116
B.2 Pseudo-discrimination kernels for detection signal trials. 117

C.1 Button presses, imaging experiment 124
C.2 zROC curves, imaging experiment 125
C.3 Parametric effect of confidence in correct responses 126
C.4 Effect of confidence in Regions of Interest 127
C.5 Inter-subject correlation between the quadratic effect in the right hemisphere clusters and the ratio between the detection and discrimination distribution variances 128
C.6 Inter-subject correlation between the linear and quadratic effects in the right hemisphere clusters and metacognitive efficiency scores 129
C.7 Cross-classification analysis 130

Abstract

Representing the absence of things is qualitatively different from representing their presence. Specifically, to represent something as absent one must know that they would have known if it was present. This form of counterfactual reasoning critically relies on having a mental self-model which specifies 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 its relation with default-reasoning and predictive-coding. First, the timing of decisions about the absence of an item has been shown to be sensitive to search time and accuracy in previous trials, but it remains unknown how decisions about the absence of an item are made in the very first trials of the experiment, before previous trials are available. In a set of behavioural experiments I provide evidence for that implicit metacognitive knowledge about spatial attention supports inference about the absence of items already in these first trials, and that this implicit knowledge is dissociable from explicit metacognitive knowledge about search difficulty. Second, subjective confidence in perceptual decisions is mostly sensitive to perceptual evidence supporting the decision, but decisions about stimulus absence are unique in that they are based on the absence of evidence, rendering positive evidence unavailable. Using reverse-correlation I identify positive stimulus features that contribute to decision confidence in decisions about absence, and discuss these findings in the context of sensory noise estimation. Third, neuroimaging studies of metacognitive monitoring have identified a network of frontal and parietal regions that are sensitive to decision confidence. Using functional MRI, I find that these regions 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 a series of six behavioural experiments I show that different levels of the cognitive hierarchy are sensitive to different notions of absence. I conclude with a discussion of specific ways in which inference about absence can be used by cognitive scientists for probing implicit metacognitive beliefs and studying the mental self-model.

Dedication

You can have a dedication here if you wish.

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.

Figure 1: Guavas.

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 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 work that all share 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.

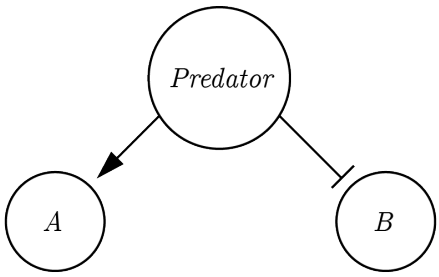
Dissymmetrical definition:

'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

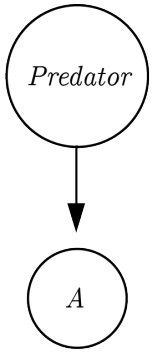
The symmetric definition entails a *first-order process*, as no knowledge about the system itself is used in the process of inferring about the world state. This definition applies to scenarios in which it is possible to have direct knowledge against the veracity of a proposition. For example, a hypothetical 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.



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

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

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 definition entails a *second-order process*.



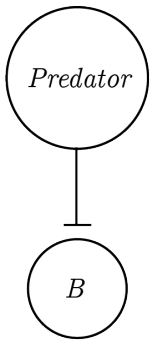
	Predator present	Predator absent
Sensor A	1	0

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

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.

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.



	Predator present	Predator absent
Sensor B	0	1

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

This architecture is perfectly equivalent to the previous one for systems 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 he 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.

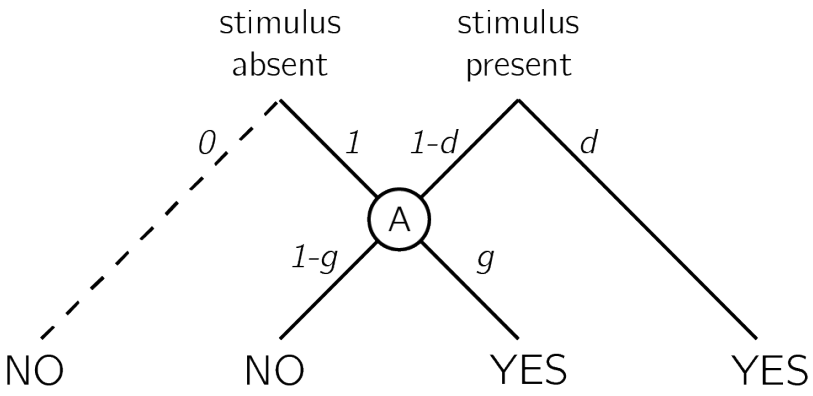


Figure 5: In discrete high-threshold models the presence of a signal can sometimes lead directly to a 'yes' response, but the absence of a signal is never sufficient to lead to a 'no' response. 'No' responses are controlled by the parameter *g* - a 'guessing parameter' that determines the probability 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 $p(\neg e|\neg T) = \frac{1-0.5}{1-0.5*0.5} = \frac{2}{3}$. 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.

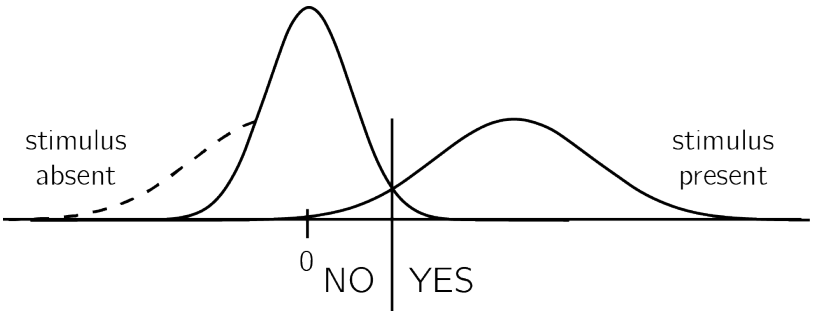


Figure 6: In unequal-variance SDT models, decisions are made based on the relative position of the sensory sample to a decision criterion. The presense/absence asymmetry manifests in the fact that only in some 'target-present' trials, but no in 'target-absent' trials, the sensory sample falls far away from the decision criterion.

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; M. Mazor et al., 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 et al. (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 (see figure 7). 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.

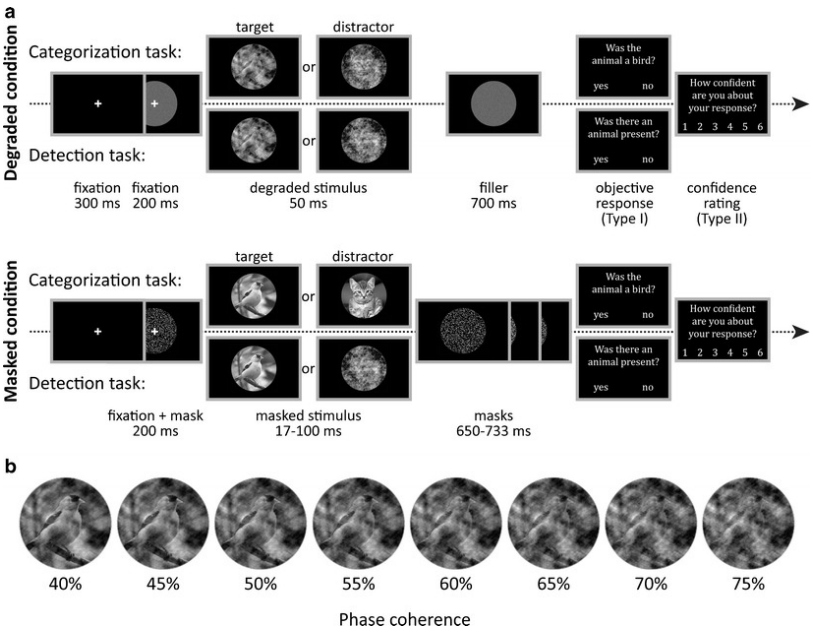


Figure 7: Task design for Meuwese et al (2014). Subjects performed both the detection task and the categorization task in 12 interleaved blocks of 60 trials. Stimulus visibility was manipulated between subjects, by either pattern masking or degrading (phase scrambling). During the detection task, the degraded or masked stimulus contained either an animal (cat, bird, or fish; target) or a fully phase-scrambled image (distractor). On every trial, subjects were asked "Was there an animal present?" For the categorization task, a target category was randomly selected for each block (i.e., "bird"), and the stimuli consisted of a degraded/masked cat, bird, or fish. Subjects were asked whether the animal was a member of the target category (i.e., "Was the animal a bird?"). Subjects rated their confidence in the correctness of their response on a scale from 1 (not at all confident) to 6 (very confident). By linking confidence ratings with objective performance, metacognitive ability (MA) was calculated. b An image that is phase scrambled to different coherence levels: from left to right, 0.4 to 0.75 phase coherence, which was the range of phase coherence levels and step sizes used in degraded condition of the experiment.

These and similar observations of a metacognitive disadvantage for inference about absence (Kanai et al., 2010; Kellij, Fahrenfort, Lau, Peters, & Odegaard, 2018; Mazor et al., 2020; Meuwese et al., 2014), as well as a similar pattern in response times (decisions about absence tend to be slower than decisions about presence; Mazor et

al., 2020) fit well with the high-threshold and unequal-variance SDT models described above. Only in the presence of a target stimulus 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, this metacognitive blindspot for 'stimulus absence' judgments diminishes or reverses when targets are masked from awareness by means of an attentional manipulation (Kanai et al., 2010; Kellij et al., 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 lower for attended versus unattended stimuli (see Fig. 8). Indeed, detection criterion is typically found to be lower for unattended stimuli (Odegaard et al., 2018; Rahnev et al., 2011; Solovey et al., 2015).

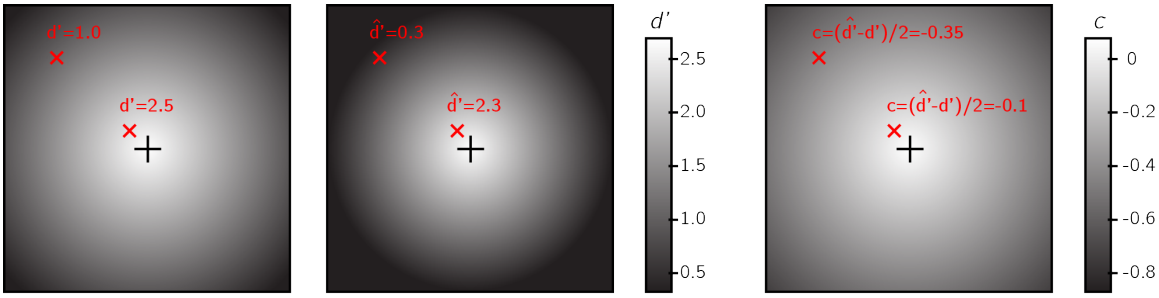


Figure 8: Left panel: Sensitivity to near-threshold stimuli is lower in the visual periphery. For example, d' equals 1.0 in top left of the screen, but is much higher near the center. Right panel: the perceptual decision criterion is lower (more 'yes' responses) in the visual periphery. Middle panel: 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 the visual array. 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 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 target-absent 'conjunction' searches terminate once participants scan all the items in the display (a *self-terminating exhaustive search*; Treisman & Gelade, 1980). However, this model predicts that search-time variability in 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; 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 absent' judgments are the result of exhausting the search only on items that surpass a learned 'activation threshold' (Chun & Wolfe, 1996; Wolfe, 1994). In difficult searches, the activation threshold was 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 could be set to a high value, reflecting a belief that a target would be highly salient (see Fig. 9). Furthermore, some very long searches terminated once subjects concluded that "it rarely takes this long to find a target" (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 can be chosen 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 was 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 the search after rejecting only one item.

In a more recent formulation of the Guided Search model (Wolfe, 2021), the search terminated once a noisy accumulator reached 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*) was 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 reflects explicit or implicit beliefs about the subjective salience of a hypothetical target in the array – a form of self-knowledge.

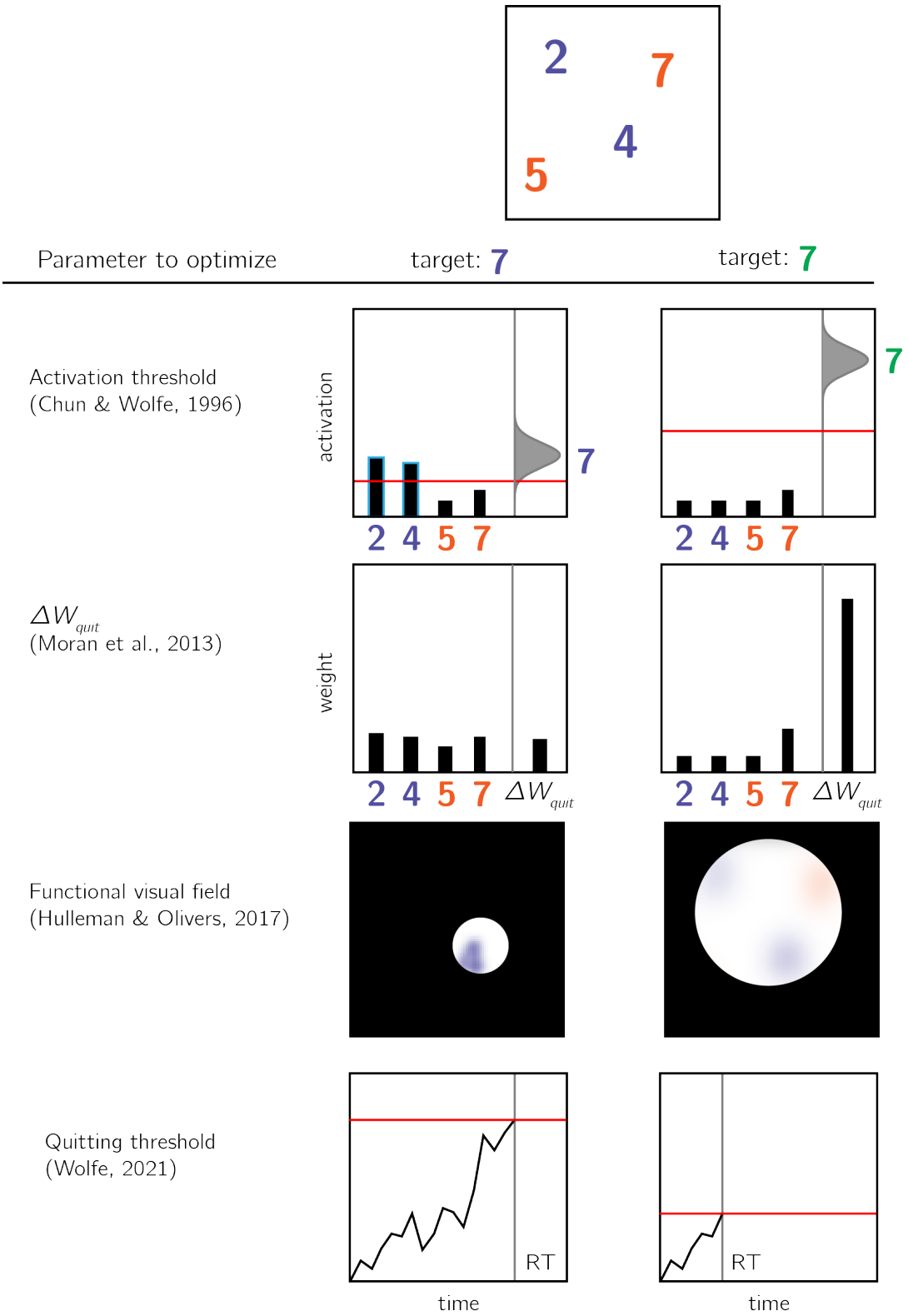


Figure 9: Models of search termination. For the same visual array (uppermost panel) search terminated immediately for one target (a green 7, right column), but takes longer for another target (a purple 7, left column). Different models of visual search explain this difference by postulating search termination mechanisms that are sensitive to the counterfactual difficulty of finding a hypothetical target.

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 (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. 10 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. 10, lower panel).

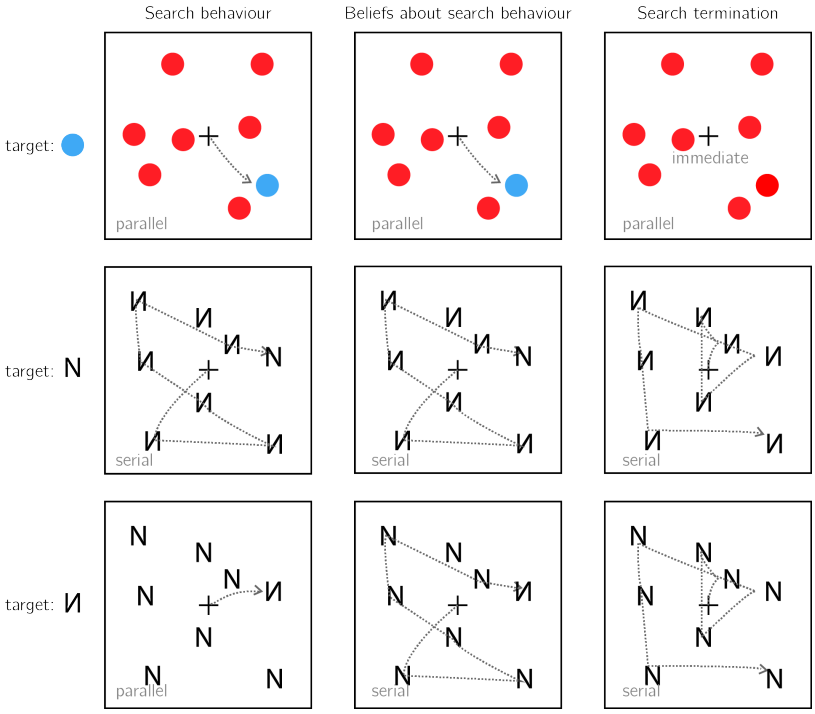


Figure 10: Upper panel: A target that is marked by a unique colour imemdiately captures attention (left). This fact is available to particiapnts' self-model (middle). As a result, participants can immediately terminate a search when no distractor shares the color of the target (right). Middle panel: 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 canincally 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 (see sections 0.2.2 and 0.2.2), 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 give rise to 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 et al., 1977). For example, Brown et al. (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 have remembered if my name was on the list"). The mirror effect has also been demonstrated for the name of one's hometown (Brown et al., 1977), for word frequency (rare words are more likely to be correctly endorsed or rejected with confidence; Brown et al., 1977; Glanzer & Bowles, 1976), word imaginability (Cortese, Khanna, & Hacker, 2010; Cortese, McCarty, & Schock, 2015) and for study time (subjects are more likely to correctly reject items 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 particiapnt 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 over 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 develops through infancy and early childhood. In context of this thesis, the development of this capacity can reflect 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 have noticed 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 et al., 2008), because they lack the self-knowledge to know that if they had, they would have remembered doing so.

0.7 This thesis

This thesis is centred 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. Not unlike near-threshold 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 characterise participants' asearch termination in the first few trials of an experiment.

Chapter 1

Zero-shot search termination reveals a dissociation between implicit and explicit metacognitive knowledge

Matan Mazor, Stephen M. Fleming

In order to infer that a target item is missing from a display, subjects must know that they would have detected it if it was present. This form of counterfactual reasoning critically relies on metacognitive knowledge about spatial attention and visual search behaviour. Previous work on visual search established that this knowledge is constructed and expanded based on task experience. Here we show that some metacognitive knowledge is also available to participants in the first few trials of the task, and that this knowledge can be used to guide decisions about search termination even if it is not available for explicit report.

1.1 Introduction

Searching for the only blue letter in an array of yellow letters is easy, but searching for the only blue X among an array of yellow Xs and blue Ts is much harder (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 (Treisman & Gelade, 1980) and guiding signals (Wolfe, 2021; 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 reasis a fundamental question: what makes some decisions about target absence easier than others?

Metacognitive beliefs about the expected time taken to detect a target can draw on previous experience in the task. Indeed, search time in target-absent trials decreases

following successful target-present trials, and sharply increases following target misses (Chun & Wolfe, 1996). This simple heuristic provided an excellent fit to data from a visual search task with hundreds of trials. However, in everyday life visual searches rarely come in a blocks of hundreds of similar trials, such that relying on previous repetitions of the same search to guide search termination is impossible (Wolfe, 2021). Only the first trials of a visual search experiment, where participants meet the stimuli for the first time, are a good model of this *zero-shot search termination* behaviour. In these trials, search time should rely solely on metacognitive beliefs about search efficiency that are available to subjects prior to engaging with the task. This fact makes search time in the first few trials of a task a critical window into participants' metacognitive knowledge about attention and visual search. Furthermore, participants' ability to learn from positive examples (target-present trials), and their ability to generalize their knowledge across stimulus types and displays, offers an opportunity to study the structure of this simplified metacognitive knowledge, its building blocks, and the inductive biases that guide its acquisition. In this study, we use target-absent trials in visual search to ask what participants know about their spatial attention before engaging with the visual search task, and how this knowledge is built and expanded based on experience.

In two pre-registered experiments here we focus on feature search for colour and shape. Focusing on the first four trials in a visual search 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. Furthermore, by making sure that the first displays do not include the target stimulus, we are able for the first time to ask what knowledge is available to participants about their expected search efficiency prior to engaging with the task.

We dub this approach *zero-shot search termination* in a tribute to the study of 'zero-shot learning' in machine learning: the ability to classify unseen categories of stimuli, based on generalizable knowledge from other categories (Xian, Schiele, & Akata, 2017). Efficient (i.e., fast and accurate) quitting in target-absent trials prior to any target-present trials would indicate that knowledge about the salience of a divergent color or shape is available at some form in the cognitive system, and that this knowledge can flexibly be put to use for counterfactual reasoning in the process of inference about absence. Conversely, inefficient search in these first trials would mean that positive experience is necessary for this knowledge to be acquired, or to be expressed.

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 is dependent on task experience. A detailed pre-registration document for Experiment 1 can be accessed via the following link: https://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 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/experiments/demos/exp1/. 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 are 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 small 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).

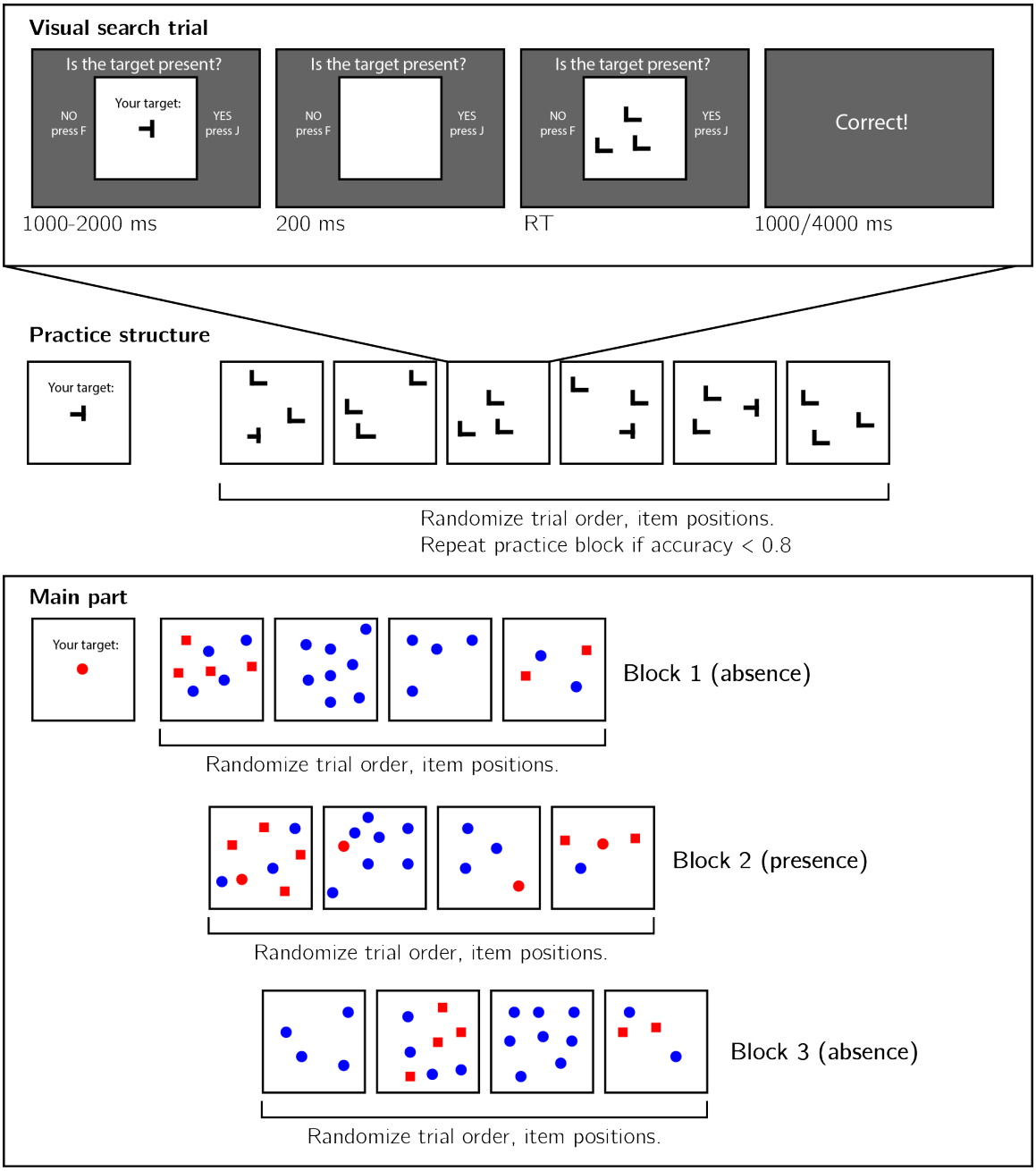


Figure 1.1: Experimental design. Top panel: each visual search trial started with a screen indicating the target stimulus. The search display remained visible until a response is recorded. To motivate accurate responses, the feedback screen remained visible for one second following correct responses 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 per 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 the 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).

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.

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).

1.2.3 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.

Data preprocessing

To control for within-block trial order effects, a separate linear regression model was fitted to the data of each block, predicting search time as a function of trial serial order ($RT \sim \beta_0 + \beta_1 i$, 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 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 trials 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.

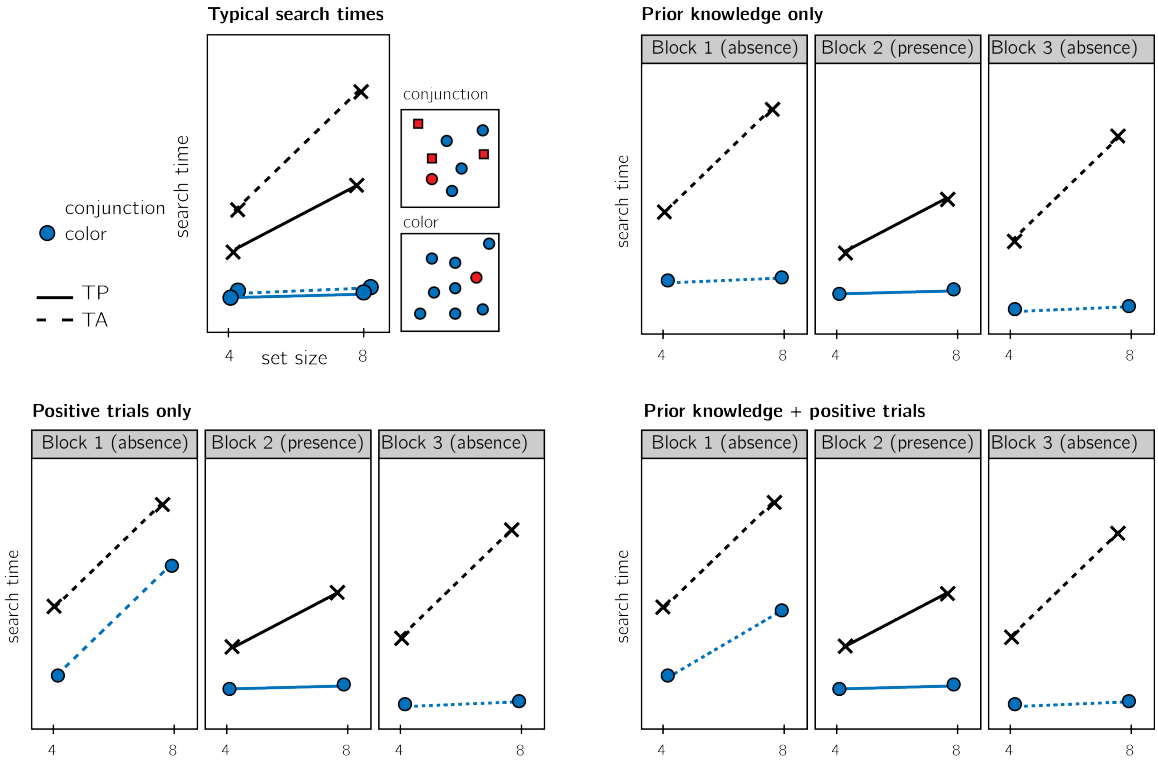


Figure 1.2: Visualization of Hypotheses. Top left: typical search time results in visual search experiments with many trials (where TP = Target Present responses; TA = Target Absent responses). Set size (x axis) affects search time in conjunction search, but much less so in color search. However, it is unclear whether this pattern of target-absent search also holds in the first trials in an experiment. Different models make different predictions about target-absent serach times in the first block of the experiment. Top right: one possible pattern 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 metacognitive 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 metacognitive 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.

Analysis comprised a positive control based 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.

We used R (Version 3.6.0; R Core Team, 2019) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *cowplot* (Version 1.0.0; Wilke, 2019), *dplyr* (Version 1.0.4; Wickham, François, Henry, & Müller, 2020), *ggplot2* (Version 3.3.1; Wickham, 2016), *jsonlite* (Version 1.7.1; Ooms, 2014), *lsr* (Version 0.5; Navarro, 2015), *MESS* (Version 0.5.6; Ekstrøm, 2019), *papaja* (Version 0.1.0.9942; Aust & Barth, 2020), *pwr* (Version 1.3.0; Champely, 2020), *reticulate* (Version 1.16; Ushey, Allaire, & Tang, 2020), and *tidyr* (Version 1.1.0; Wickham & Henry, n.d.) for all our analyses.

1.2.4 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 conjunction search (M = 3.91, 95% CI [2.13, 5.70]; see middle panel in Fig. 1.3). The slope for color search 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, the slope for the conjunction search was steep (M = 18.41, 95% CI [14.95, 21.87]). A clear 'pop-out' effect for color search 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. 1.3), 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. 1.3). 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 $[-10.52, 3.36]$, $t(320) = -1.01$, $p = .311$).

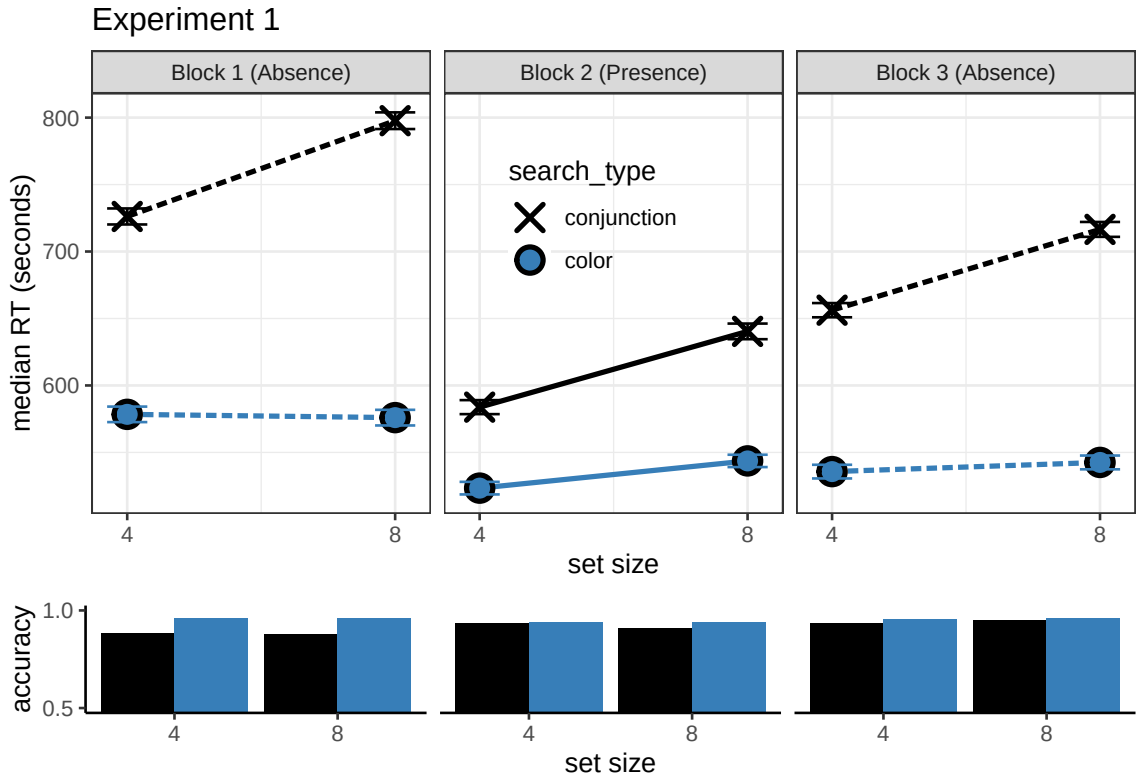


Figure 1.3: Upper panel: median search time by distractor set size for the two search tasks across the three blocks (12 trials per participant). Correct responses only. Lower panel: accuracy as a function of block, set size and search type. Error bars represent the standard error of the median.

Additional analyses

In Experiment 1, we found a clear pop-out effect for color absence in the first trials of the experiment, before participants experienced color pop-out in target-present trials. As per our analysis, this reflects prior metacognitive knowledge about the expected efficiency of color search. In order to terminate the search immediately, participants must have known, implicitly or explicitly, that a red item would have popped out immediately. In the setting of this experiment, this knowledge could not be acquired in previous trials. However, an alternative account is that participants noticed the pop-out of the red distractors in the conjunction trials of block 1, and based their expectation for color pop out on those trials. This account can be directly tested by zeroing in on the subset of participants who performed the two color trials before the two conjunction trials in block 1 (the order of trials within each block was determined pseudorandomly, such that half of the participants had color-search for the first trial, and of those a third had color-search for the second trial as well). This subset of participants showed a clear pop-out effect ($M = -5.07$, 95% CI $[-\infty, 2.25]$, $t(138) = -3.41$, $p < .001$), indicating that the highly efficient search termination in these first trials was not based on prior experience with red distractors.

1.3 Experiment 2

Experiment 1 provided evidence that color-absence pop-out occurs prior to experiencing color pop-out in the context of the same task. We interpret this as indicating that task-naive adults had valid implicit or explicit metacognitive expectations about color pop-out. This metacognitive knowledge may be innate (acquired in the course of evolution, for example driven by the utility of color search for foraging), learned from previous visual experience (for example, first-person experience of attention being immediately drawn to distinct colors), or culturally acquired (for example, through language). Experiment 2 was designed to extend these findings to another stimulus feature that is found to also efficiently guide attention: shape. The time cost of additional distractors in shape search was under 10 ms in our pilot data, rendering it another case of parallel, efficient search. It is possible however that unlike in the case of color, the metacognitive knowledge that gives rise to the pop-out effect for shape-absence is acquired through experience with the task. Unlike the colour space, that 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 requires a richer mental model of attentional processes. Furthrmore, colour is agreed to be a 'guiding attribute of attention', while it is unclear which shape features guide attention (Wolfe & Horowitz, 2017). In this experiment we also include an additional control for prior experience with visual search tasks, and ask whether the implicit metacognitive knowledge about pop-out is available for explicit 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 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 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/experiments/demos/e
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. 1.4). 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 the 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. 1.4), indicating that a processing advantage for the detecting the absence of a shape compared to the absence of shape-color conjunction was already in place before experience with target presence.

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 the slope for conjunction search ($t(565) = 6.02$, $p < .001$; see rightmost panel of Fig. 1.4).

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. 1.4), 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.

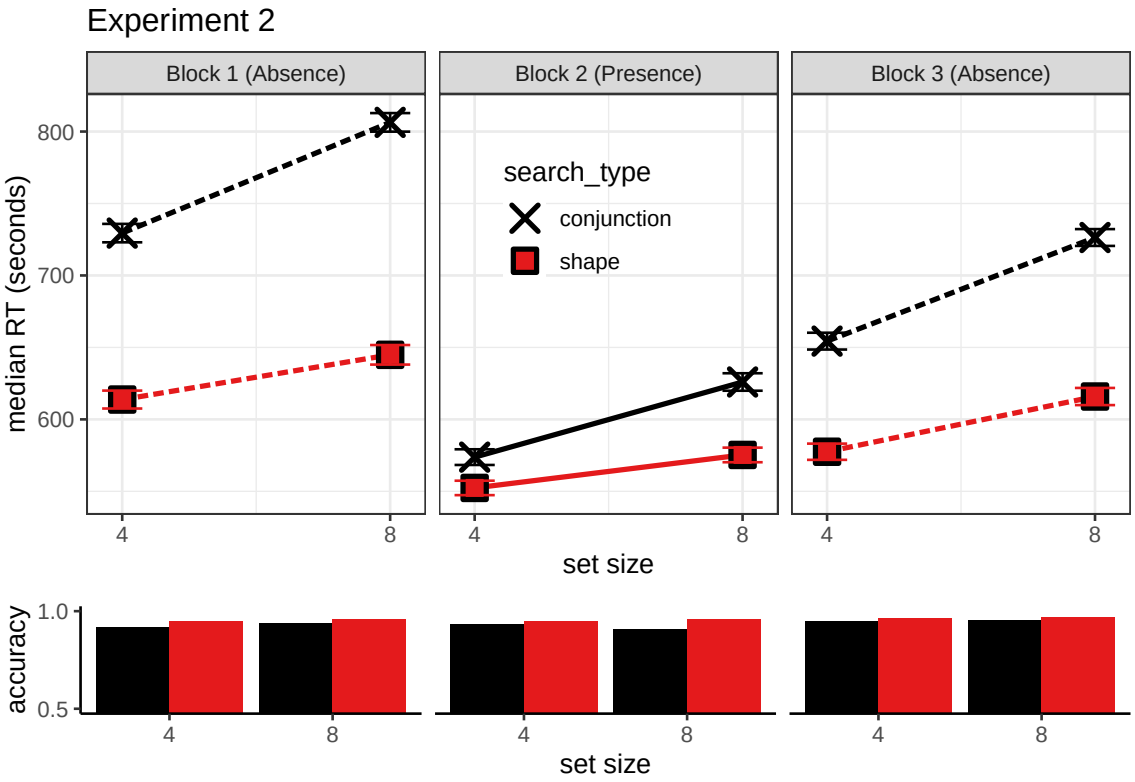


Figure 1.4: Upper panel: median search time by distractor set size for the two search tasks across the three blocks. Correct responses only. Lower panel: accuracy as a function of block, set size and search type. Error bars represent the standard error of the median.

Additional Analyses

Exploratory analysis: 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 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 performing a visual search task with similar stimuli. Participants that 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.

Exploratory analysis: search time estimates

Upon completing the main part of Experiment 2, participants placed the four search arrays (shape and conjunction searches with 4 or 8 distractors) on a perceived difficulty axis. We used these 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 target-absent shape search trials may depend on an internal belief that the target shape would have drawn attention immediately, but this belief may inaccessible to introspection. If introspective access is not a necessary condition 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 rates 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 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], $t(788) = 3.48$, $p = .001$).

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 estimated 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 ($N = 83$). 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, and in support of a functional dissociation between explicit and implicit metacognitive knowledge, search slopes for shape-absence trials were shallower than for conjunction-absence trials ($M_d = 12.45$, 95% CI [5.21, 19.69], $t(82) = 3.42$, $p = .001$).

1.4 Discussion

Deciding that an item is absent requires counterfactual thinking, in the form of 'I would have seen it if it was present'. In some cases, it is immediately clear that an hypothetical target would have been detected (such as when searching for a red item, but seeing only blue items), and in other cases more deliberate searching is needed until this belief can be held with confidence (such as when searching for a conjunction of features, for example colour and shape). Here we sought to determine the origins of this metacognitive knowledge that allows participants to conclude that a target would be found immediately in the first case, but not in the second. Specifically, we asked if this knowledge depends on task experience (such that with time, participants learn that some searches are easier than others), or alternatively, whether it is available already in the first trials of the experiment.

Previous studies of search termination have focused on the calibration of a quitting strategy over long chains of similar trials. For example, in a seminal study by Chun & Wolfe (1996), participants decreased their activation threshold (the necessary activation for an item to be scanned) following misses, but increased the threshold following correct rejections. This calibration mechanism critically depended on two features of the experimental design: a large number of similar trials, and explicit feedback about accuracy. Similarly, in a multi-session perceptual training study by Ellison and Walsh (1998), response times became faster over sessions, and search slopes for conjunction search became shallower. In more recent studies, participants were able to learn statistical regularities in spatial position (Moorselaar & Slagter, 2019) and visual features (Moorselaar, Lampers, Cordesius, & Slagter, 2020) of distractor stimuli in a visual search task, and to use this information for making faster responses. These studies revealed important mechanisms by which task experience can affect visual search behaviour, but they left open the question of what guides search termination in the absence of any task experience. Our zero-shot search termination paradigm revealed 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.

In two experiments, 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 was absent. Participants were sensitive to the counterfactual likelihood of detecting a hypothetical target even in the first trials of the experiment, suggesting that metacognitive knowledge about visual attention (e.g., 'red pops out', or 'a dot would catch my attention') is available to guide zero-shot search termination. In Experiment 2, we find that some of this knowledge is represented explicitly, as expressed in participants' ordering of visual search arrays by difficulty. However, focusing on participants with erroneous metacognitive beliefs about search efficiency, we find that explicit metacognitive knowledge is not a necessary condition for efficient search termination. More broadly, this finding indicates a functional dissociation between explicit and implicit metacognitive knowledge.

1.4.1 Is implicit metacognitive knowledge metacognitive?

In this study we assumed that efficient search termination is impossible without accurate metacognitive knowledge about search difficulty. We base this conjecture on our conceptual analysis of inference about absence: in order to represent something as absent, one must know that they would have detected it had it been present (M. Mazor & Fleming, 2020). Alternative approaches to visual search assume that the absence of a stimulus can sometimes be perceived directly, without alluding to any metacognitive beliefs or counterfactual thinking. For example, ensemble perception allows observers to extract summary statistical information from sets of similar stimuli, without directly perceiving every single stimulus (Whitney & Yamanashi Leib, 2018). According to one alternative explanation of our results, if participants immediately perceive that the search array is all blue, they might not need to rely on any counterfactual thinking or self-knowledge to conclude that no red item was present. Similarly, when searching for a red dot, there is no need to serially scan a search array if it is immediately perceived as comprising only squares.

When contrasting this alternative account with our counterfactual model, it is useful to ask how does the visual system extract ensemble properties from sets of objects. 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 if present. This 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 polygons only'). Even within the ensemble perception framework, inference about the absence of items must be based on some form of meta-level knowledge about the cognitive and perceptual systems. The fact that attention may not be required for ensemble perception (Hochstein, Pavlovskaya, Bonneh, & Soroker, 2015) can inform and constrain our theories of where this meta-level knowledge is represented in the cognitive hierarchy, but it does not, by itself, weigh on the question of whether this is indeed metacognitive knowledge.

We note here that it not a prerequisite that metacognitive knowledge 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 metacognition 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). More relevant to visual search, 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 *Attention Schema* is metacognitive in the sense that it reflects self knowledge about one's own attention. This kind of implicit metacognitive knowledge may be crucial for extracting ensemble statistics from displays, and for representing the absence of objects.

1.4.2 Inference about absence as a tool for studying implicit self knowledge

Participants' early quitting in target-absence feature searches taught us something about implicit self-knowledge. This is not a coincidence, but an example of a general principle: inference about absence critically relies on self-knowledge not only in visual search ('If a target was present, I would have found it') but also in near-threshold detection ('If a stimulus was present, I would have noticed it'), recognition memory ('If this item was in the study list, I would have remembered it'), and problem-solving ('If a solution to this problem was present, I would have come up with it'). This makes inference about absence an important tool for studying implicit self-knowledge in a range of domains without relying on explicit metacognitive reports. For example, in the context of recognition memory, items that are most likely to be remembered are also the ones that are most likely to be correctly rejected as foils when new. This 'mirror effect' (Brown et al., 1977) conceptually resembles the alignment of feature-present and feature-absent search times across items and visual dimensions in visual search: if a target is found easily within a set of distractors S, it would also be easy to conclude that a target is absent if S is presented without the target in it. Just as in the study of visual search, previous studies of the mirror effect adopted a typical many subjects/few trials designs (e.g., Brown et al., 1977; Glanzer & Adams, 1985; Greene & Thapar, 1994). By generalizing the approach we have taken here to implicit metacognitive knowledge of memory, future *Zero-shot negative recognition* experiments could ask whether the self knowledge that gives rise to the mirror effect is also available prior to engaging with the task.

1.4.3 Conclusion

Search termination in the first few trials of an experiment (zero shot search termination) showed the same qualitative response time pattern as that commonly found 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 metacognitive knowledge about search difficulty was often accessible to report, but that this was not a necessary condition for efficient search termination. We interpret our results as indicating a dissociation between implicit and explicit metacognitive knowledge, with the former having a particularly influential role in inference about absence.

Chapter 2

Prospective search time estimates for unseen displays reveal a rich intuitive theory of visual search

Matan Mazor, Max Siegel & Joshua B. Tenenbaum

abstract

2.1 Introduction

The *Intuitive Theories* approach to cognitive science (Gerstenberg & Tenenbaum, 2017) has been successful in accounting for human knowledge and reasoning in the domains of physics (McCloskey, 1983), psychology (Baker, Saxe, & Tenenbaum, 2011) and semantic knowledge (Gelman & Legare, 2011). In recent years, careful experimental and computational work has advanced our understanding of these simplified theories: their ontologies and causal laws, the abstractions that they make, and the consequences of these abstractions for faithfully and efficiently modeling the real world. For example, the computational specification of the intuitive physics model and its deviation from Newtonian physics was informed by empirical measures of biased intuitions about the consequences of object collisions (Sanborn, Mansinghka, & Griffiths, 2013; Smith & Vul, 2013).

Theoretically, there is no reason to believe that Intuitive Theories should be limited in their scope to modeling the external environment and other agents. Indeed, agents may benefit from having an intuitive theory or a simplified model of their own perceptual, cognitive and psychological states. For example, in the context of memory, 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 et al., 1977), and self-modeling has been proposed to play an important role in inference about absence more broadly (Mazor, n.d.). In the context of perception and attention, Graziano & Webb (2015) argued that having a simplified *Attention Schema* (an intuitive theory of attention and its dynamics) is crucial for monitoring and controlling one's attention, similar to how a body-schema supports

motor control.

Still, little experimental work has been devoted to characterizing the computational specifications of this intuitive theory of attention. Is it based on a simulation engine (similar to the game engine proposal; Ullman, Spelke, Battaglia, & Tenenbaum, 2017)? Or instead formatted as a list of propositions (e.g., *'My attention span is shorter when I am tired'*)? How accurate is it? To what extent is it learned from experience and what inductive biases guide its acquisition and tuning based on experience?

Here we take a first step in this direction, using visual search as our model test-case. Participants estimated their prospective search times in visual search tasks and then performed the same searches. Similar to using colliding balls and falling blocks to study intuitive physics, here we chose visual search for being thoroughly studied and amenable to relatively simple modeling. In Experiments 1 and 2, we used simple colorful shapes as our stimuli, and compared participants' intuitive theories to scientific theories of attention that distinguish parallel from serial processing. We found that participants were sensitive to the parallel/serial distinction, 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) to demonstrate the richness and compositional nature of participants' intuitive theories, and their reliance of idiosyncratic knowledge.

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

A good intuitive theory needs to have good predictive value without being overly complex. A reasonable first candidate for what an intuitive theory of visual search may look like is Anne Treisman's *Feature Integration Theory* (FIT). According to FIT, visual search comprises two stages: a *pre-attentive* parallel stage, and a serial *focused attention* stage (Treisman, 1986; 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'. Search targets that are defined by a single feature with respect to its surroundings (*feature search*; for example searching for a red car in a road full of yellow taxis) can be located based on the feature map. Since the extraction of the feature map is pre-attentive, in these cases the search can be completed immediately. However, 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.

In its simplest form, Treisman's FIT predicts that search time should linearly scale with the number of distractors in conjunction search, but not in feature searches. This model provides reasonably accurate search time predictions for simple displays with only three parameters: non-decision time (the y-intercept of the set size X search time curve), the time cost of deploying attention to an item (the slope of the same curve), and a list of the features that can be found without serially scanning the display (for

example, color, orientation, and size).

In Experiments 1 and 2 we used stimuli that lend themselves to a categorical distinction between parallel and serial search: simple geometric shapes of different colors and orientations. We asked whether participants' intuitive theory of visual search can 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. 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).

2.2.1 Participants

For Exp. 1, 100 participants 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 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).

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 respond as fast or faster than their estimation. 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. The report scale ranged from 0.1 to 4 seconds in Exp. 1 and to 2 seconds in Exp. 2.

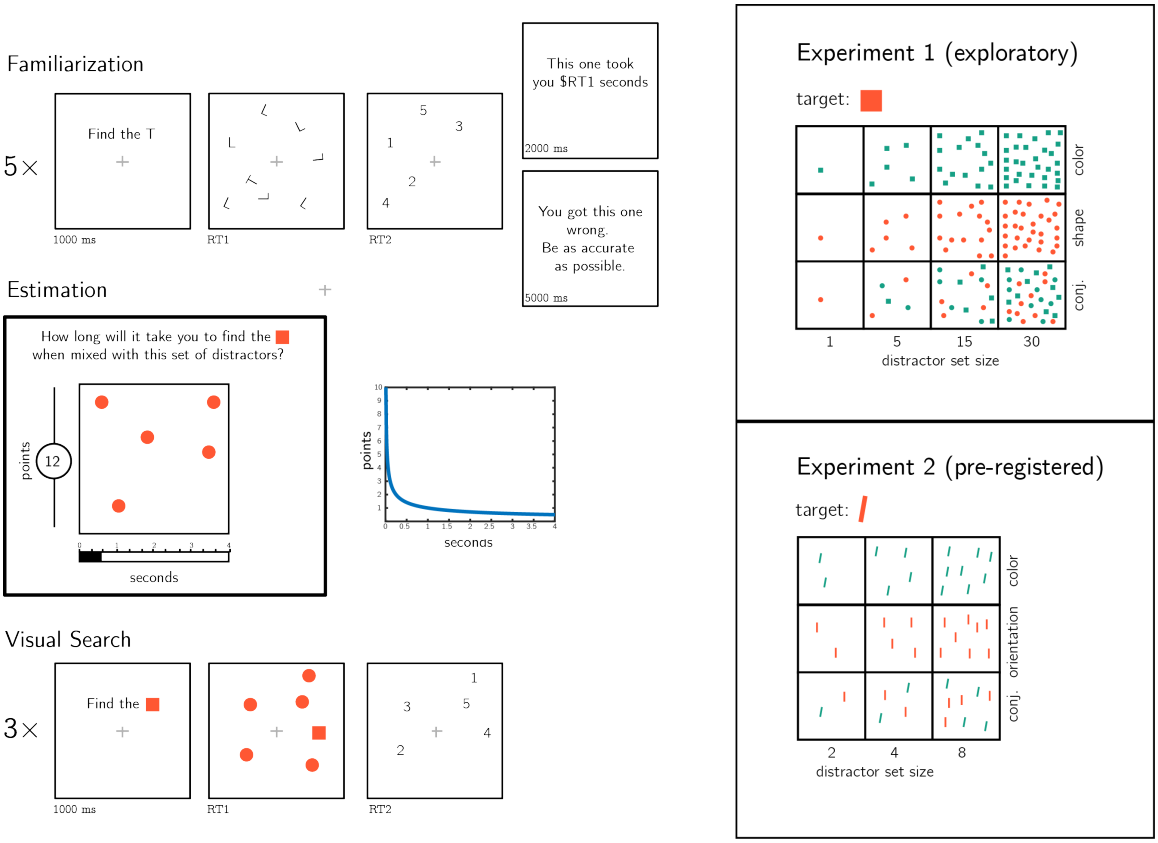


Figure 2.1: Experimental design. Participants first performed five similar visual search trials and received feedback about their speed and accuracy. Then, they were asked to estimate the duration of novel visual search tasks. Bonus points were awarded for accurate estimates, and more points were awarded 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.

After one practice trial (estimating search time for finding one T among 3 randomly positioned Ls), 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 (#FF5733) tilted bar (20° off vertical) among green (#16A085) titled bars (color condition), red vertical bars (orientation condition) and a mix of green tilted and red vertical bars (oriention-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 0.2 seconds 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 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 than 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 (Treisman, 1986; 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 ($t(73) = 1.50$, $p = .139$, $\text{BF}_{01} = 2.70$). 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 **??**, upper panel). This is consistent with the FIT prediction that conjunction search demands serial attention.

Estimation accuracy

We next turned to analyze participants' prospective search time estimates, and their alignment with actual search times. In both tasks, 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). Despite this bias, estimates were correlated with true search times, supporting a metacognitive insight into visual search behaviour (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' intuitive theory of visual search, we analyzed participants' estimates as if they were search times, and extracted search slopes relating estimates to the number of distractors in the display. 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. Still, in both Experiments estimated 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 estimate slopes was not significant (difference between shape and conjuntion slopes: $t(88) = 1.65$, $p = .103$; difference between orientation and conjunction slopes: $t(73) = 1.18$, $p = .244$).

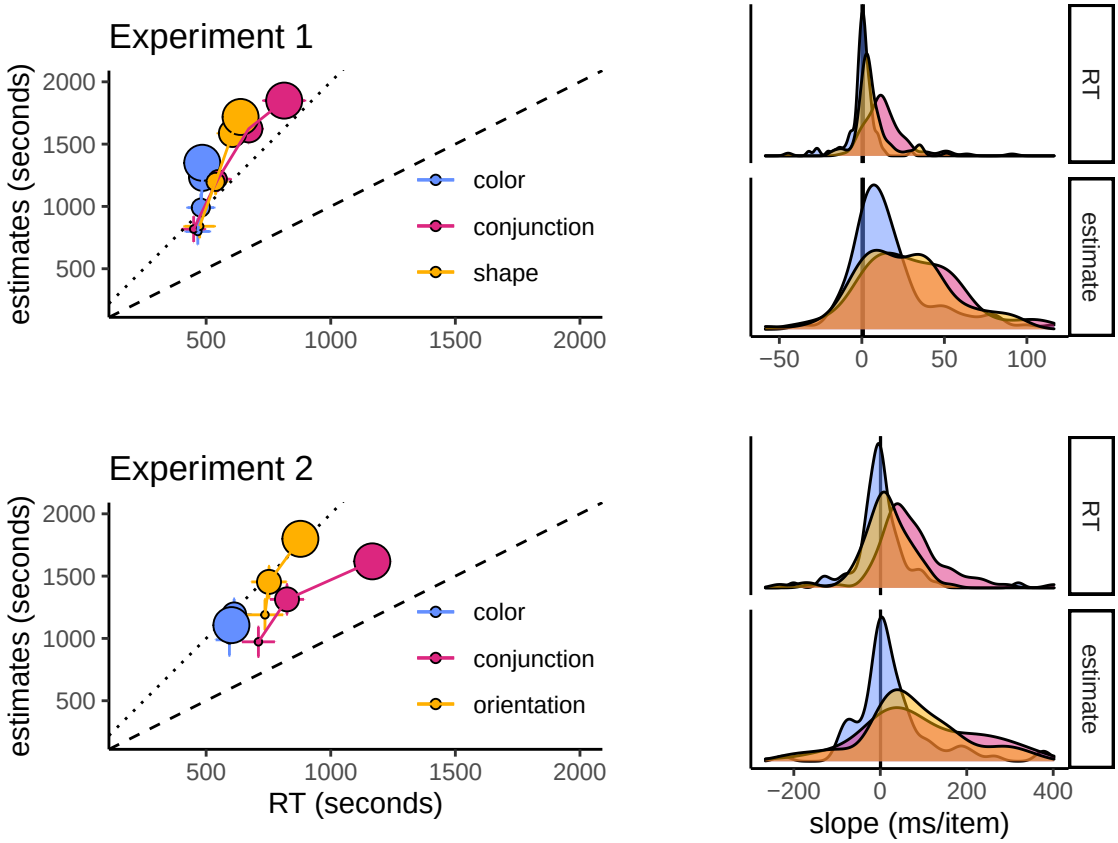


Figure 2.2: Left panels: median estimated search times plotted against true search times for the different search types (coded by color), and set sizes (coded by circle size; from small to large), for Exp. 1 (upper panel) and 2 (lower panel). Error bars represent the standard error of the median. Right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

2.3 Experiments 3 and 4: complex, unfamiliar stimuli

In Experiments 1 and 2 participants' intuitive theory of visual search allowed them 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. Importantly, this knowledge could not have been acquired in the familiarization phase of the experiment, where we used 'T' and 'L's as our stimuli and all displays had the same number of distractors. We also found that participants' intuitive theory of visual search was systematically biased to overestimate the set-size effect, 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 intuitive theory is, by using displays of complex stimuli with which participants are unlikely to have had 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' intuitive theory 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. Experiment 4 followed Experiment 3 and was pre-registered (pre-registration document: osf.io/dprtk).

2.3.1 Participants

For Exp. 3, 100 participants 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 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 (section 2.3.3. 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 et al., 2011), instead of geometrical shapes. 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) from among copies of another letter (the distractor letter). In Exp. 3, target and distractor letters were drawn from the Alphabet of the Magi, and distractors were drawn by different Mechanical Turk workers. In Exp. 4, the target and distractor letters were drawn from different Alphabets, with the target being a Latin letter on half of the trials and a Futurama letter on the other half. In this experiments, distractors were copies of the same letter drawn by the same Mechanical Turk worker. This was important for our visual search asymmetry analysis (section 2.3.3).

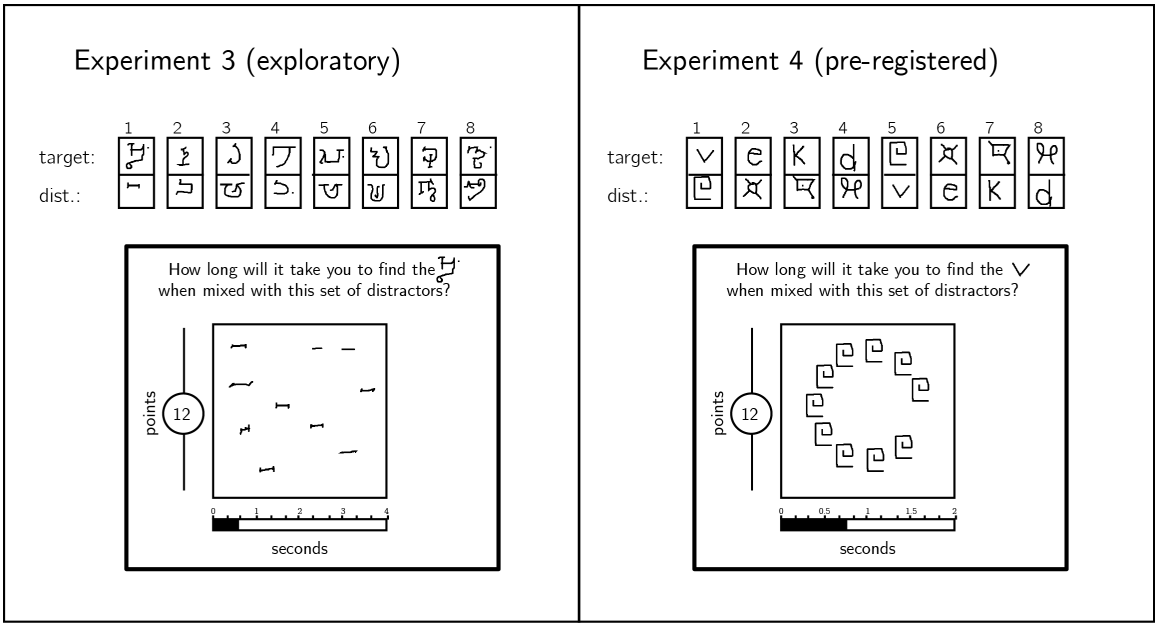


Figure 2.3: Stimuli used for Experiments 3 and 4. In Exp. 3, stimuli were characters from the Alphabet of the Magi, and distractors were drawn by different Mechanical Turk Users. In Exp.4, stimuli were characters from the Latin and Futurama alphabets. Stimulus pairs 1-4 and 5-8 are identical except for 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 clockface.

In the familiarization part, we used as target and distractors two letters from the Alphabet of the Magi in Exp. 3 and two letters from the Futurama alphabet in 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, in Exp. 4 letters appeared on an invisible clockface surrounding the fixation cross. 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 Exp. 3 ($M = 0.89$, 95% CI [0.86, 0.92]) and at ceiling in Exp. 4 ($M = 0.97$, 95% CI [0.96, 0.98]). Error trials and visual search trials that took 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 2.4 and 2.5). 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 more narrow range of search times in Exp. 4 (with median search times ∞ - $-\infty$ ms, per display) than in Exp. 3 (∞ - $-\infty$ 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 be 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 correlation reveal a more intricate theory 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.

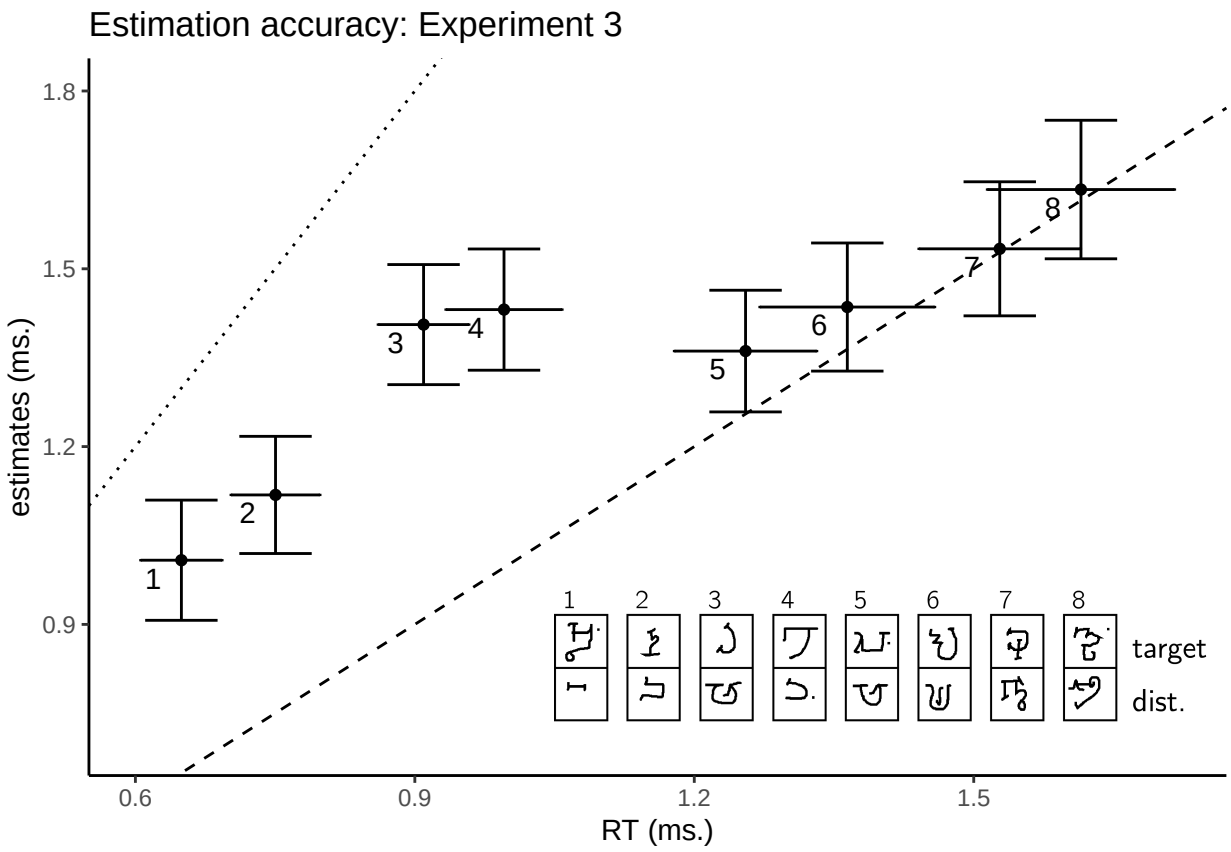


Figure 2.4: Estimated search times plotted against true search times in Experiment 2. The dashed line indicates $y = x$ and the dotted line indicates $y = 2x$. Legend: each search task involved searching for one Omniglot character (top letter) among ten tokens of a second Omniglot character, drawn by 10 different MTurk workers (bottom letter).

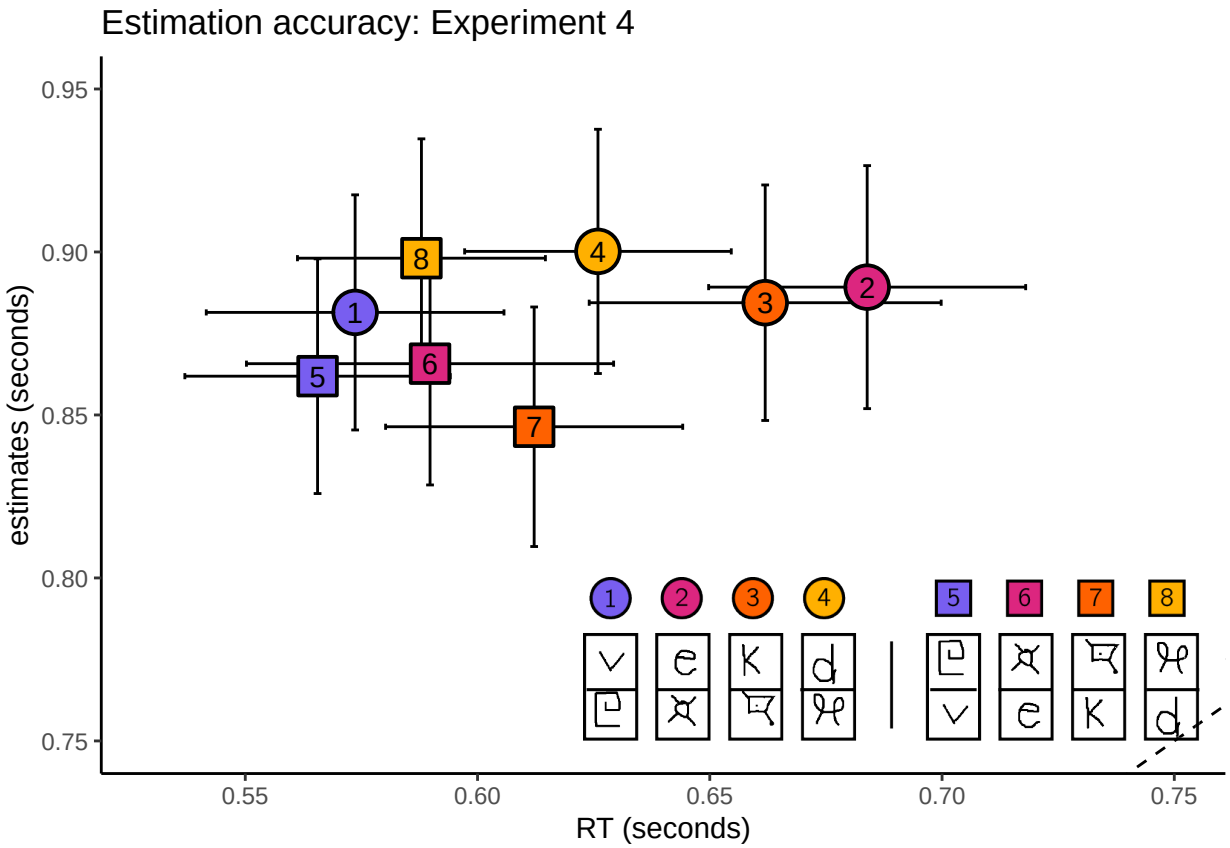


Figure 2.5: Median estimated search times plotted against true search times in Experiment 4. The dashed line indicates $y = x$. Legend: each search task involved searching for one character (top letter) among ten tokens of a different character (bottom letter). In four searches, the target character 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.

Cross-participant correlations

In our choice of stimuli for Experiments 3 and 4 we were motivated to make heuristic-based estimation more difficult, and instead 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 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-seld (mean Spearman correlation $M_r = 0.44$) and seld-other alignment ($M_r = 0.41$) was marginally significant ($p_{perm} = 0.05$). In Experiment 3, we found a significant advantage for self-self alignment compared with self-other alignment (mean Spearman correlations for self-self $M_r =$

0.10 and self-other $M_r = 0.04$, $p_{perm} = 0.01$).

This advantage for self-self alignment is unlikely to reflect motivated slowing in searches that were rated as difficult. Our bonus scheme incentivized accurate search time estimates in the Estimation part, but in the search part points were awarded for speed. We interpret this result as indicating that at least some of participants' intuitive theory of visual search builds on idiosyncratic knowledge about their own attention.

Estimation time

We next looked at the time taken to give 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 see 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 ($t(191) = 0.74$, $p = .458$).

Visual search asymmetry

[PARAGRAPH ABOUT SEARCH ASYMMETRY]

To test if participants were sensitive to this asymmetry in their prospective visual search estimates, we extracted the estimated/true search time correlations after inverting the identity of the target and distractor stimuli in the estimates, but not in the actual search times. If estimates were affected by the assignment of stimuli to target and distractor, this inversion should reduce 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], $t(191) = 2.63$, $p = .009$).

Chapter 3

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

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. This asymmetry suggests that higher-order cognitive and neural processes may be required for confidence in detection, and more specifically, in judgments about absence. 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.

3.1 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; 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 (for reviews, see Fleming & Dolan, 2012; Domenech & Koechlin, 2015; 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 et al., 2010; Meuwese et al., 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.

3.2 Methods and Materials

All design and analysis details were pre-registered before data acquisition and time-locked using pre-RNG randomization (Mazor et al., 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.

3.2.1 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.

3.2.2 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.

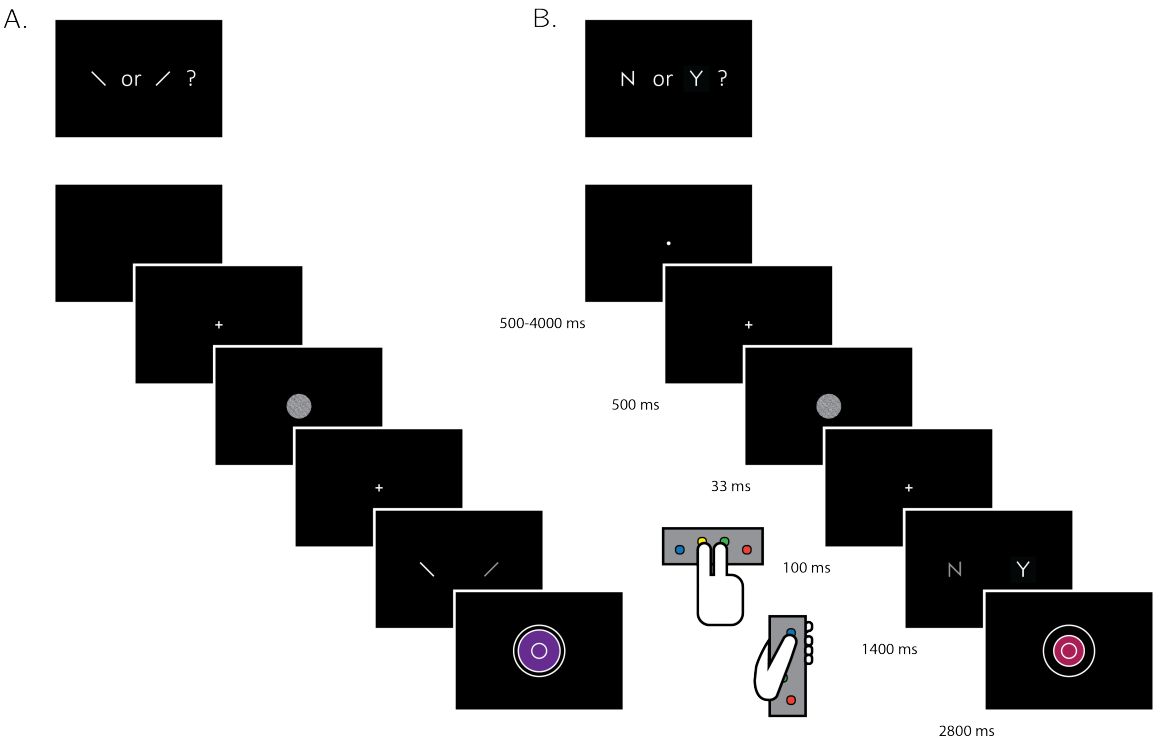


Figure 3.1: Experimental design for discrimination and for detection trials . Perceptual decisions were reported using the right index and middle fingers, and confidence ratings were reported using the left thumb. A) In discrimination blocks, participants indicated the orientation of a visual grating ('clockwise' or 'counterclockwise'). B) In detection blocks, participants indicated whether a grating was embedded in the 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.

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.

3.2.3 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).

3.2.4 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.

- 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 et al., 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.

3.2.5 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.

3.2.6 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.

3.2.7 Imaging analysis

fMRI data preprocessing

Data preprocessing followed the procedure described in Morales et al. (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 et al., 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 et al. (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 (present and present, P_P). 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 3.1)

Table 3.1: 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

		Task	Stimulus	Response
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			
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 et al., 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 (C.7), but do not interpret them further.

3.2.8 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 $p<0.001$.

3.3 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.

3.4 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 0.49ś0.11, 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 0.50ś0.08, 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 3.2). 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 ($t(34) = 2.39, p < 0.05$).

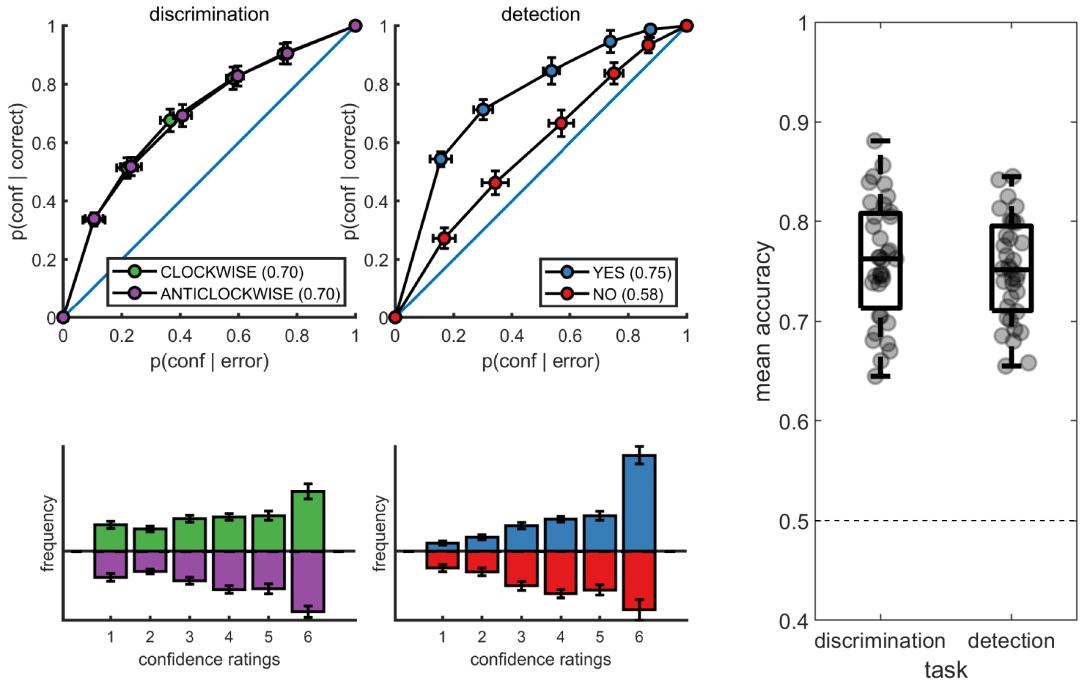


Figure 3.2: Upper panels: response conditional type-2 ROC curves. In parentheses: the mean area under the curve. Lower panels: distribution of confidence ratings for the two tasks and four responses. Right panel: Mean accuracy for both tasks. Error bars represent the standard error of the mean.

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 3.2, as expected (Meuwese et al., 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 et al., 2010; Meuwese et al., 2014).

Response times were faster on average for correct responses (849ś79 milliseconds) compared to incorrect responses (938ś95 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 (850ś90 milliseconds and 896ś103 milliseconds, respectively; $t(34) = 3.16, p < 0.005$ for a paired t-test on the log-transformed response times).

3.4.1 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 et al., 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 3.3). For the confidence effect pattern obtained from the Global-Confidence Design Matrix (GC-DM), see supplementary figure C.3.

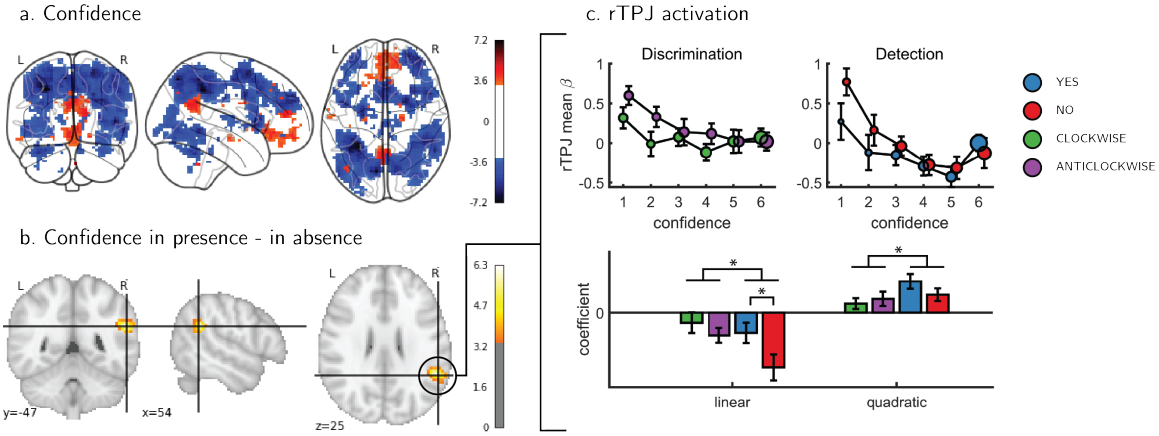


Figure 3.3: Univariate parametric effect of confidence. a) Glass brain visualization of global effect of confidence, thresholded at the single voxel level for visualization ($p<0.001$, uncorrected). Negative confidence effect appears in blue, and positive effect in red. b) Whole brain contrast between confidence in 'yes' (target-presemt) 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 3.3, 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 3.3). 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 C.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 3.4). 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; $t(34) = 2.10, p < 0.05$).

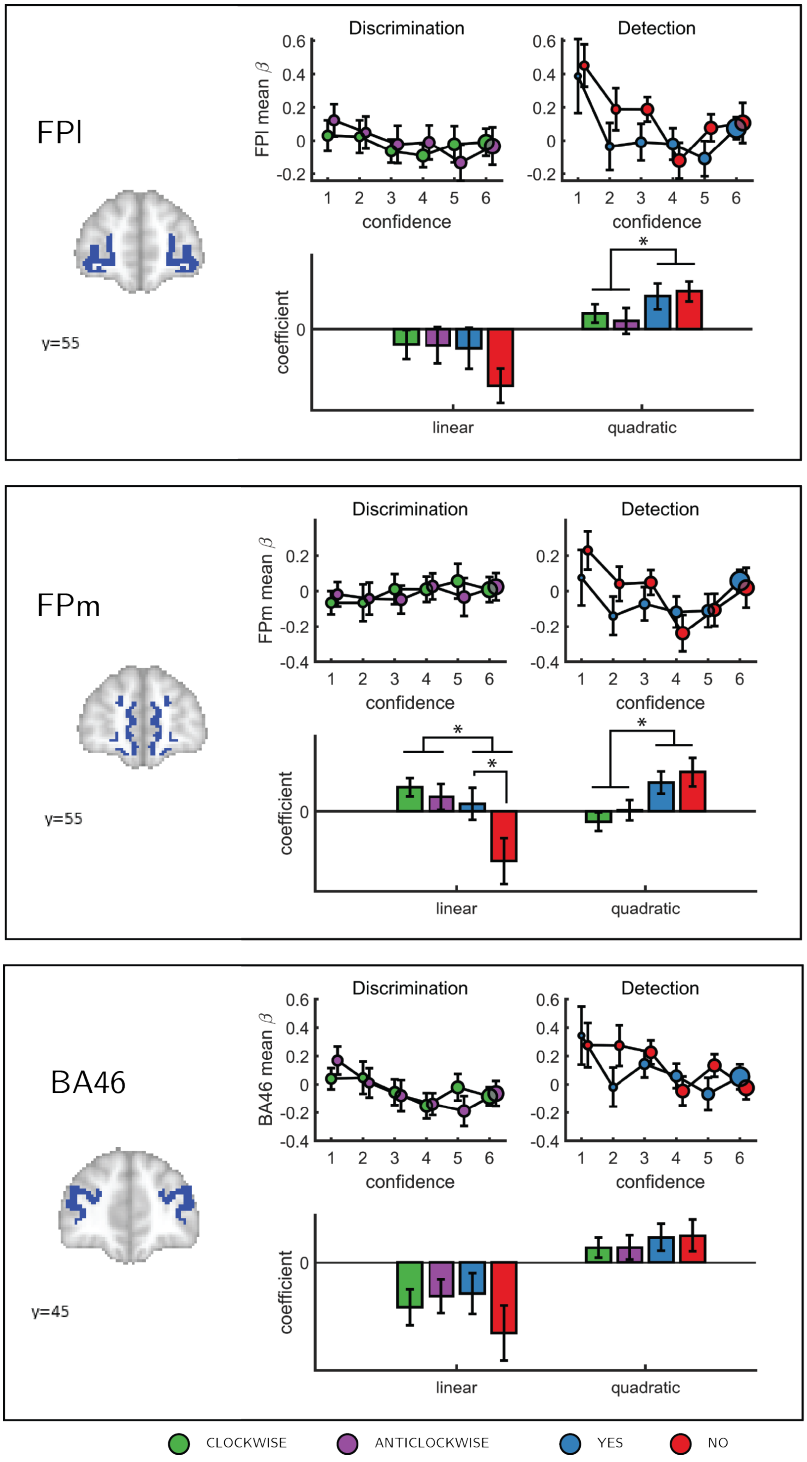


Figure 3.4: Confidence effect as a function of response in the frontopolar cortex separated into its three anatomical subcomponents: FPm, FPl, and BA 46. Same conventions as in Figure 3.3c * - p<0.05; uncorrected for multiple comparisons.

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 3.5). 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 3.5). 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 C.5 in the supplementary materials).

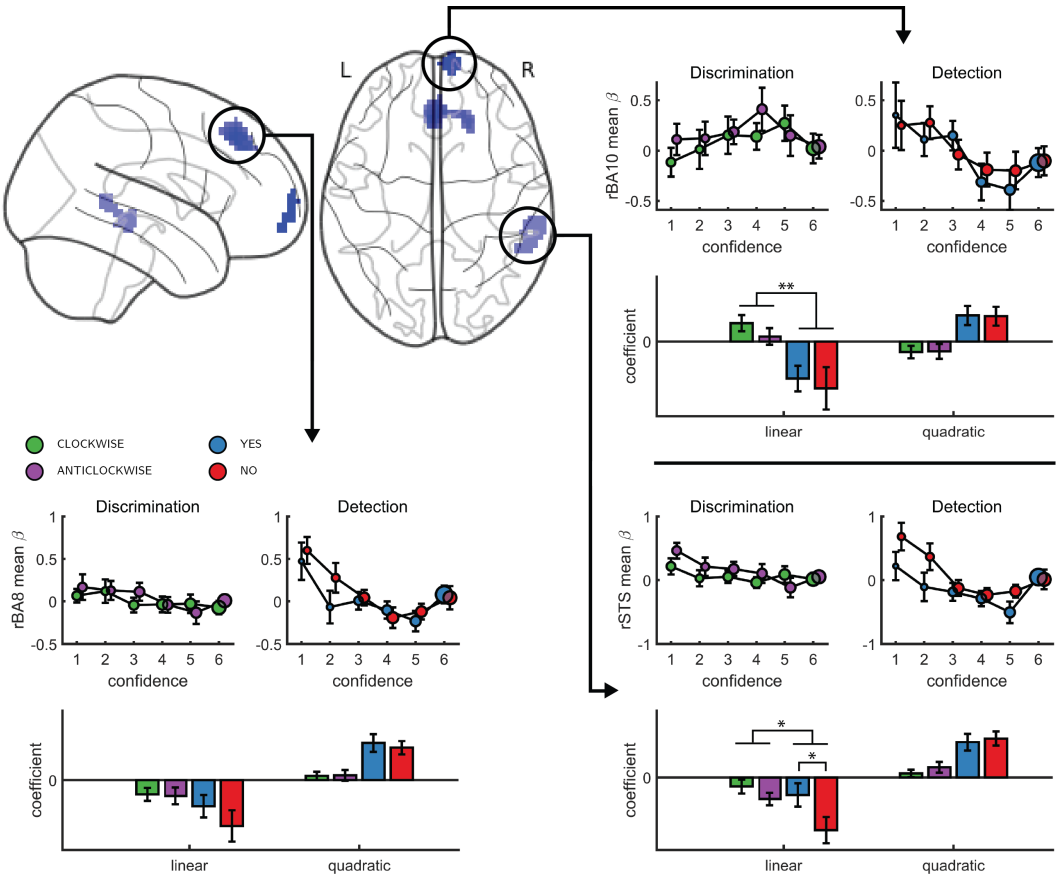


Figure 3.5: Left, top panel: a glass-brain representation of a contrast between the quadratic effects of confidence in detection and in discrimination, whole-brain corrected for family-wise error rate at the cluster-level ($p<.05$) with a cluster-defining threshold of $p<.001$, uncorrected). Remaining 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. * - $p<0.05$; ** - $p<0.01$

3.4.2 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 3.6): 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 C.8, C.9 and C.10), and Matlab

implementations are available in the following page.

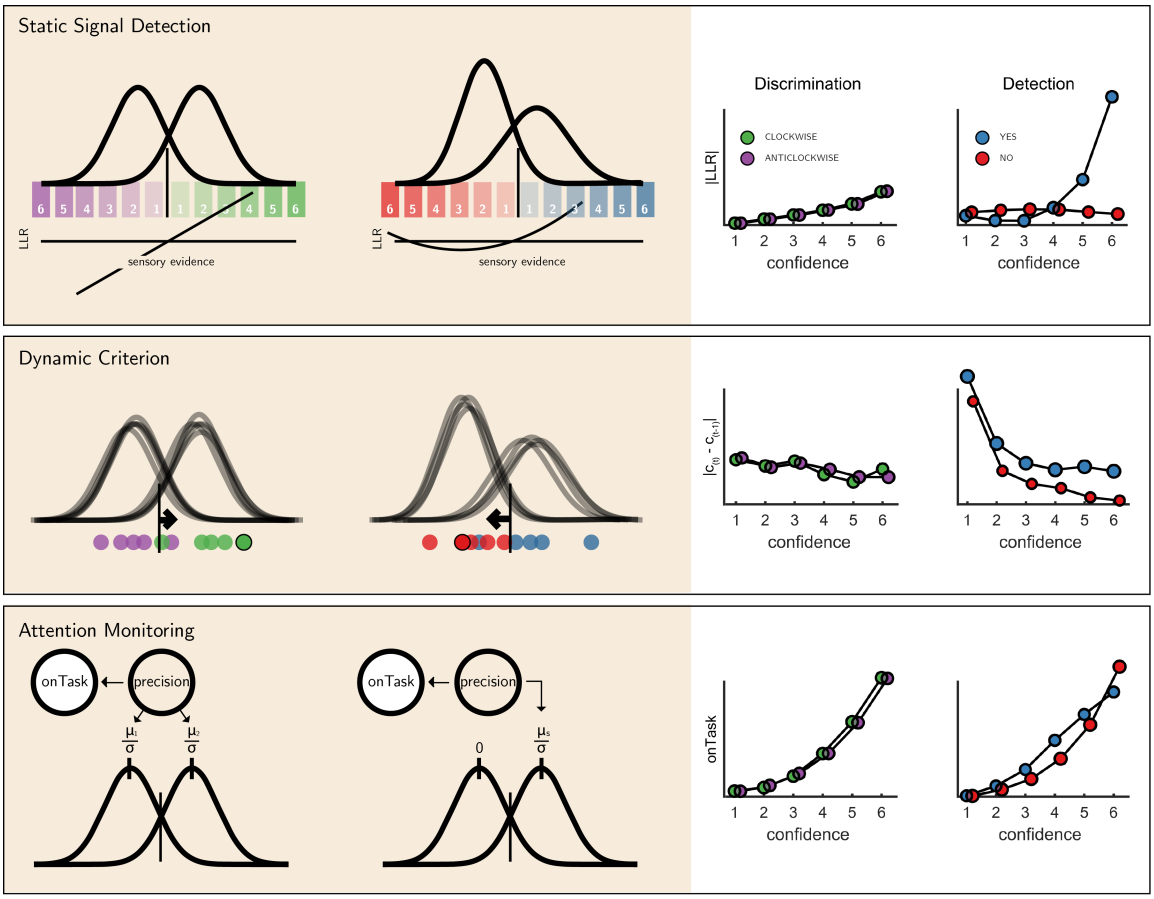


Figure 3.6: The three models (left) and their prediction for confidence effects (right). Top panel: In Signal Detection Theory, perceptual decisions and confidence ratings are generated by comparing the sensory evidence to a fixed set of criteria. In detection the 'signal' distribution is assumed to 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.

First, we consider the static Signal Detection Theory (SDT) model. In SDT models of confidence formation, the 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 with confidence in yes responses (see simulation results in Figure 3.6, 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 6—figure 1).

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 (for a formal description of the model, see Appendix section 10). 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 3.6, 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.

3.5 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 (see Appendix 4—figure 1), 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 C.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 et al., 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 et al., 2010; Kellij et al., 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 4

Paradoxical evidence weightings in confidence judgments for detection and discrimination

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

In two experiments we asked what sensory evidence is incorporated into decisions and confidence judgments in perceptual decisions about stimulus presence or absence (detection) and stimulus category (discrimination). We successfully replicated the positive evidence bias in discrimination confidence ratings: subjective confidence was boosted more by supporting evidence than it was undermined by conflicting evidence, in line with a detection disposition to the discrimination task. We further find that detection judgments show the same positive evidence bias as discrimination confidence ratings. Paradoxically, confidence ratings in detection present a discrimination-like evidence weighting, with equal weighting of positive and negative evidence. First-order perceptual decision making models fail to account for the entire set of findings.

4.1 Introduction

When considering two alternative hypotheses, the probability of a chosen hypothesis to be correct is not only a function of the likelihood of the observations under the chosen hypothesis, but also of the likelihood of the observations 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, in their subjective confidence ratings subjects put unproportional weight on positive evidence, giving rise to a *positive evidence bias* (Koizumi, Maniscalco, & Lau, 2015; 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, but also to the *sum evidence* for the two hypotheses (also termed *visibility*; Rausch, Hellmann, & Zehetleitner, 2018).

Focusing on sum rather than relative evidence is rational if subjects are rating their confidence not in the identity of the stimulus, but in the presence or absence of a signal. For example, when judging the direction of motion in a random dot kinematogram, if motion energy is high both to the left and to the right, confidence in the direction of motion should be low (low relative evidence), but confidence in the presence of coherent motion, regardless of its direction, should be high (high sum evidence). A positive evidence bias in discrimination judgments may indicate that participants are rating their confidence not in the accuracy of their choice, but in the presence of a signal.

This implied link between metacognitive evaluation and detection (judgments about the presence or absence of a signal) has led us to examine the contribution of perceptual evidence to decision and confidence in perceptual detection tasks. We were interested in three questions: first, when faced with a detection task where targets are drawn from two stimulus classes, would detection decision be sensitive to sum evidence (like in discrimination confidence), or to the relative evidence for presence for one category over the other? Second: would confidence in the presence of a target stimulus be susceptible to the same positive evidence bias as confidence in stimulus type? And finally, when making decisions about the absence of a signal, would confidence ratings be sensitive to some form of positive evidence for absence, or be entirely independent of sensory evidence?

In two experiments 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 their responses and confidence ratings, as well as markers of a processing asymmetry between detection 'yes' and 'no' responses (response time, general confidence, and metacognitive sensitivity). To anticipate our results, we fully replicated previous findings of a positive evidence bias in discrimination responses (Zylberberg et al., 2012). Paradoxically, although detection decisions were sensitive to sum evidence as expected, we found no positive evidence bias in confidence judgments following detection 'yes' responses. In Experiment 2, where reverse correlation revealed an accumulation of positive evidence for stimulus absence, we find no metacognitive sensitivity between the two detection responses. We discuss our findings as drawing a link between discrimination confidence ratings and detection responses, but not detection confidence ratings.

4.2 Experiment 1

4.2.1 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 subject recruiting system, 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 Experiment 1 largely followed the procedure described in Zylberberg et al. (2012), Experiment 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 response bias in 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. 4.1).

To control for response requirements, for 5 subjects the dots moved to the right or to the left, and for the 5 other 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 once in 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 et al. (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 patches of equal sizes of 56 dots each. Every other frame, the dots of one patch 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 patch moved coherently in one direction by a fixed distance of 0.33°, while the remaining dots in the patch moved in random directions by a fixed distance of 0.33°. On the next update, the patches 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.

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

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).

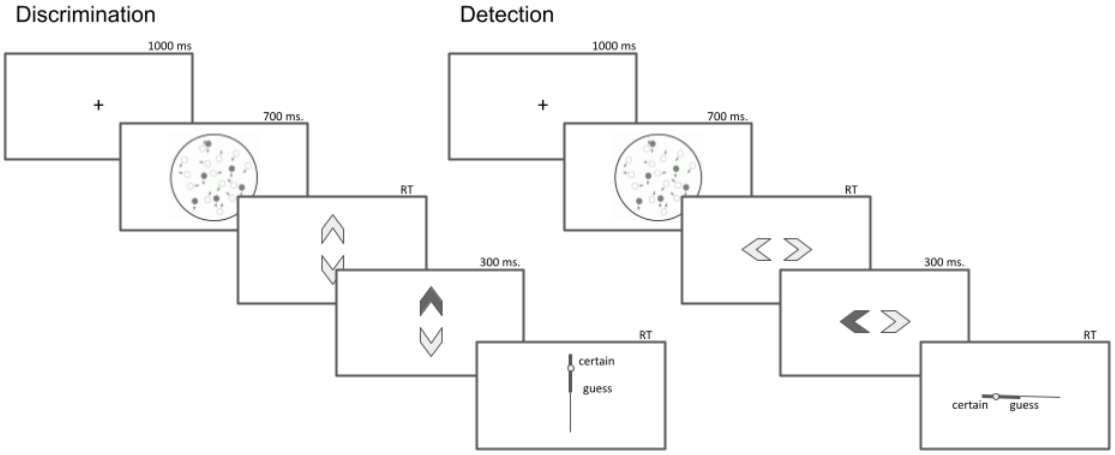


Figure 4.1: Task design for Experiment 1. In both tasks, participants viewed 700 milliseconds of a random dot motion array, after which they made a keyboard response to indicate their decision (motion direction in discrimination, signal absence or presence in detection), followed by a continuous confidence 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.

4.2.2 Analysis

Reverse correlation analysis

For the reverse correlation analysis, we followed a procedure similar to the one described in Zylberberg et al. (2012). For each of the four directions (right, 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 et al., 2012). The outputs of the two filters were squared and summed, resulting in a three-dimensional matrix with motion energy in the 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 trial-wise estimates of temporal fluctuations in the mean

and variance of motion energy for upward, downward, leftward and rightward motion. 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 et al. (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 et al., 2009).

4.2.3 Results

Response accuracy

Overall accuracy level 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 conflating true differences between discrimination and detection with more general difficulty effects, the first session was excluded from all subsequent analyses.

Overall properties of response 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], $t(9) = 1.00$, $p = .344$).

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. 4.2, upper panel). 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. 4.2, lower panel). 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], $t(9) = 0.91$, $p = .384$).

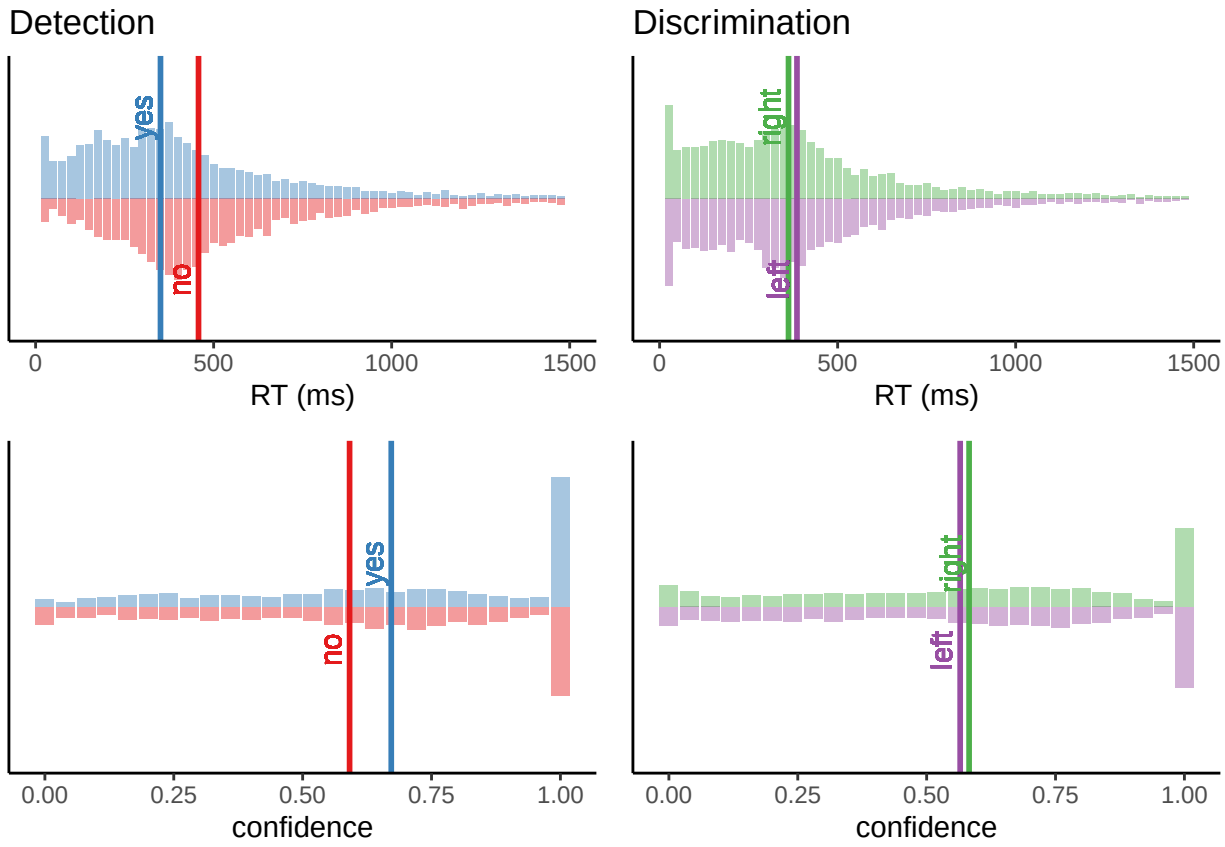


Figure 4.2: Response time (upper panel) and confidence (lower panel) histograms for the detection (left) and discrimination (right) tasks in Experiment 1. Vertical lines represent the median response time and the mean confidence rating for each response.

Response conditional ROC curves

Following Meuwese et al. (2014), we extracted response-conditional type-2 ROC (rc-ROC) curves for the two tasks. Unlike traditional type-I ROC curves that provide a visual representation 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], $t(9) = 3.28$, $p = .010$).

To better understand the origin of this difference between 'yes' and 'no' curves, we compared the detection auROC2 values with the average discrimination auROC2. We found both a significant increase in auROC2 for 'yes' responses ($M = 0.06$, 95% CI [0.01, 0.11], $t(9) = 2.80$, $p = .021$) and a marginally significant decrease in auROC2 for 'no' responses relative to discrimination ($M = -0.05$, 95% CI [−0.10, 0.00], $t(9) = -2.16$, $p = .059$). In other words, relative to our discrimination benchmark, metacognitive asymmetry in detection was driven by improved metacognitive insight into the accuracy of 'yes' responses, and degraded metacognitive insight into the accuracy of 'no' responses.

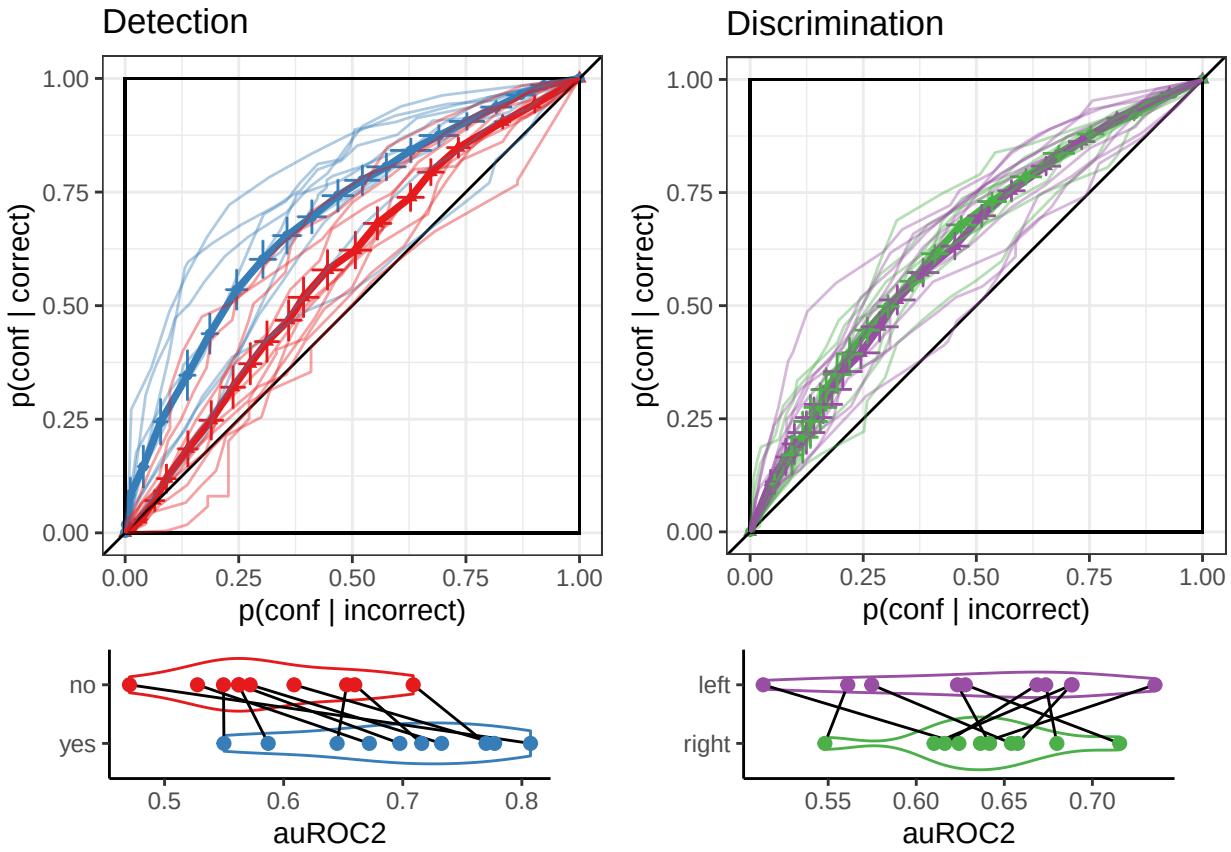


Figure 4.3: Response conditional ROC curves for the two tasks and four responses in Exp. 1. The area under the curve is a measure of metacognitive sensitivity, and the difference in areas between the two responses a measure of metacognitive asymmetry. Lower panel: distributions of the area under the curve for the four responses, across participants. Error bars stand for the standard error of the mean.

A difference in response-conditional auROC estimates can emerge from higher-

order differences in metacognitive monitoring for the two responses or from lower-level differences in the perceptual representations of signal and noise (such as in first-order signal detection models where the signal variance is higher; Maniscalco & Lau, 2014). Importantly, a difference can also emerge in first-order signal-detection models that assume equal variance, in the presence of a response bias or insufficient variance in confidence ratings. To test if the metacognitive asymmetry between 'yes' and 'no' responses could be accounted for a by an equal-variance SDT model, we simulated data that was identical to our empirical data except for confidence ratings in correct responses, which were chosen to perfectly agree with the assumptions of an equal-variance SDT model given participants' decision criterion, sensitivity, and their confidence in incorrect responses. We then compared subject-wise differences between the response-conditional auROCs with the differences in this simulated dataset (Mazor, Moran, & Fleming, 2021). The difference in differences was significant, indicating that the observed metacognitive asymmetry could not be accounted for by a first-order equal-variance SDT model ($M = 0.08$, 95% CI [0.02, 0.14], $t(9) = 2.96$, $p = .016$).

Reverse Correlation

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

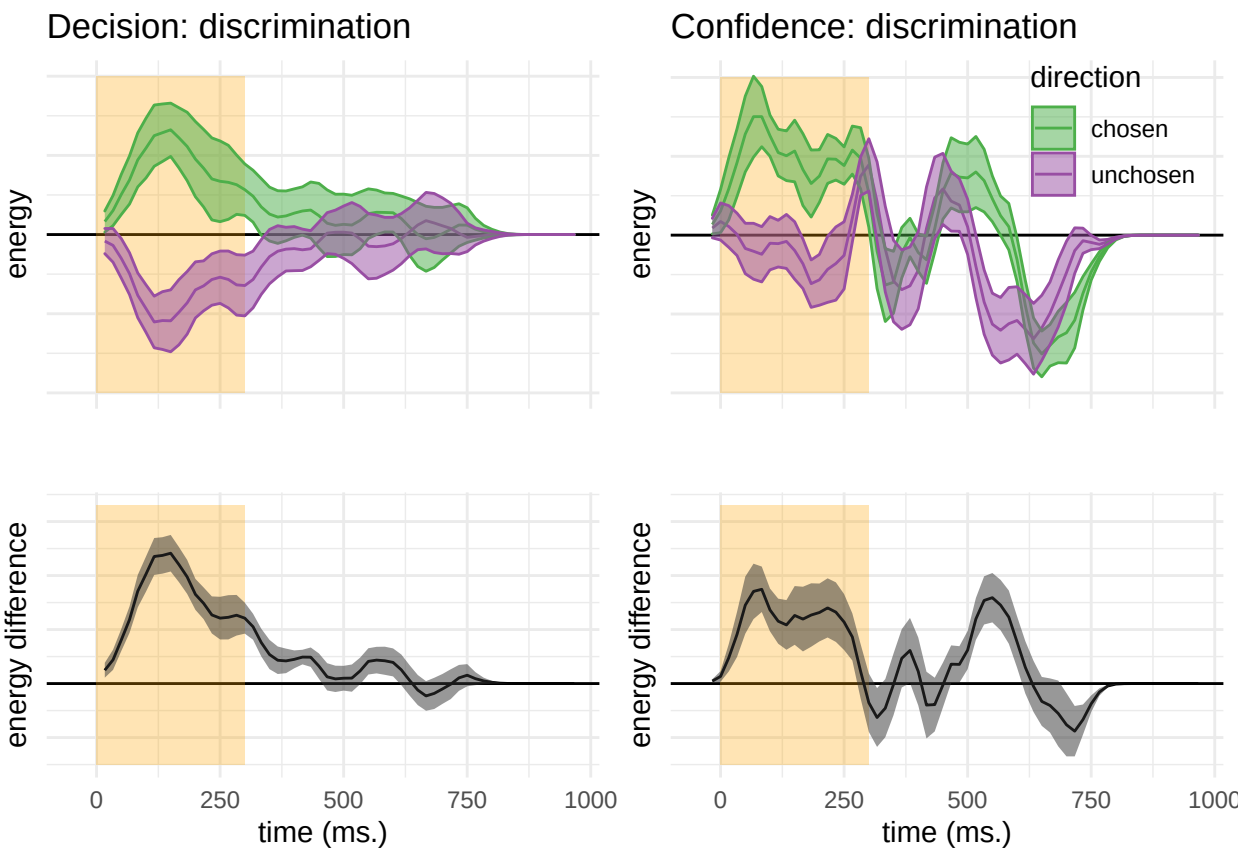


Figure 4.4: Decision and confidence discrimination kernels, Experiment 1. Upper left: motion energy in the chosen (green) and unchosen (purple) direction as a function of time. Lower left: a subtraction between energy in the chosen and unchosen directions. Upper right: confidence effects for motion energy 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

Discrimination Reverse correlation analysis quantified the effect of random fluctuations in motion energy on the probability of responding 'right' and 'left' (or 'up' or 'down'), and the temporal dynamics of decision formation. Similar to the results obtained by Zylberberg et. al., participants' decisions were sensitive to motion energy fluctuations during the first 300 milliseconds of the trial ($t(9) = 7.73$, $p < .001$; see Fig. 4.4, left panels). We note that 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 negative and positive evidence are defined based on participants' decision, making it impossible to test their contribution to decisions without engaging in circular inference. Instead, we tested the contribution of motion energy in the true and opposite directions (defined with respect to the stimulus, not the subject's decision) to discrimination decision. Fluctuations in motion energy in both

directions contributed significantly to discrimination decision ($t(9) = 8.38$, $p < .001$), with no significant difference between them ($t(9) = -0.65$, $p = .529$). To conclude, in agreement with the interpretation of Zylberberg et al. (2012), we observed no positive evidence bias in discrimination responses, even when positive and negative evidence were defined with respect to the stimulus itself.

We then turned to the contribution of motion energy to subjective confidence ratings. The median confidence rating in each experimental session was used to separate 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. 4.4, 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 again a successful direct replication of the Positive Evidence Bias observed in Zylberberg et al. (2012).

Detection We next turned to the effects of motion energy on detection responses and confidence ratings. Reverse correlation for detection introduces a challenge: while 'no' responses reflect a belief in the absence of any coherent motion, 'yes' responses can result from three different belief states: participants can detect motion in any of the two directions, or in 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 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 in their 'yes' responses.

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 ($t(9) = 1.23$, $p = .249$; see Fig. 4.5, left panel) or at any other time point. The effect of sum motion energy during the first 300 milliseconds on decision confidence was marginally significant ($t(9) = 2.15$, $p = .060$; see Fig. 4.5, right panel). Response-specific effects of sum motion energy on decision confidence were not significant for both responses.

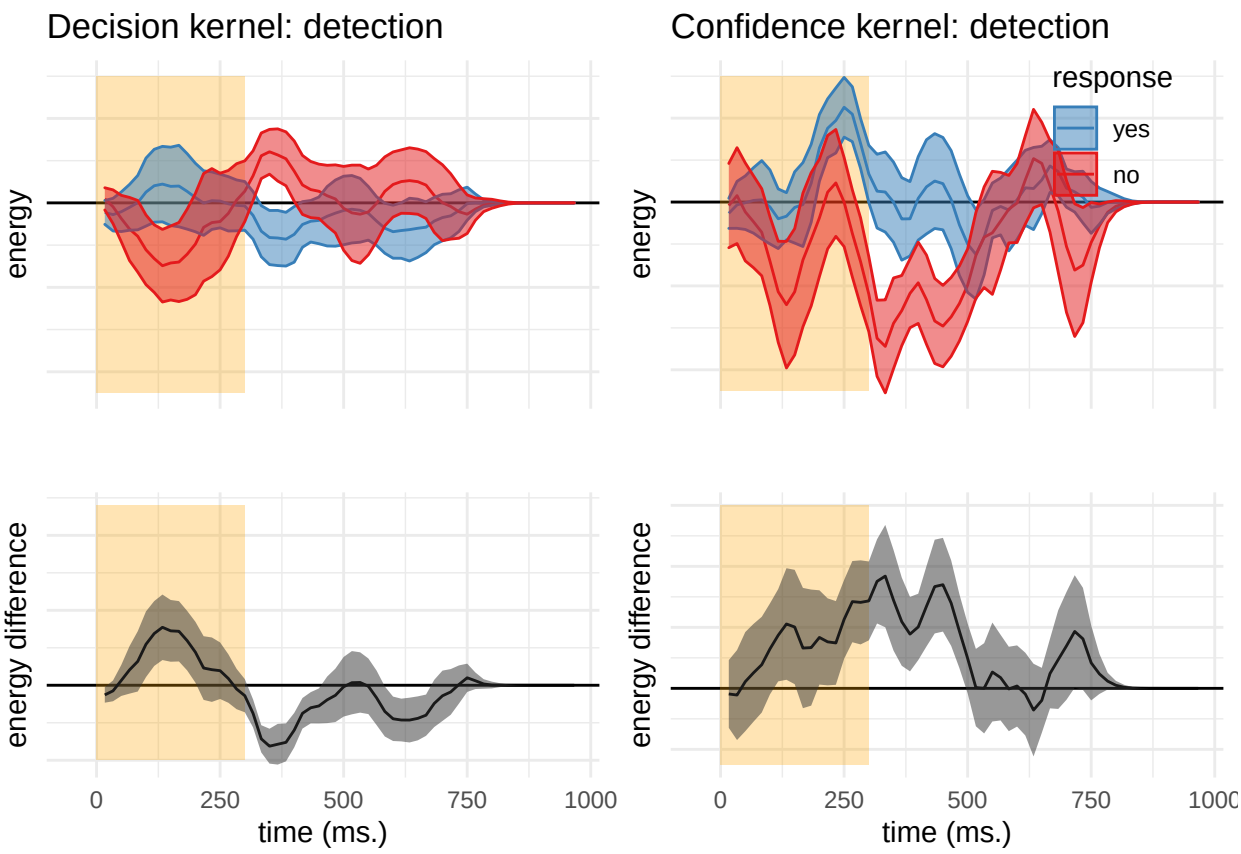


Figure 4.5: Decision and confidence detection kernels, Experiment 1. Upper left: sum motion energy along the relevant dimension in 'yes' (blue) and 'no' (red) responses as a function of time. Lower left: a subtraction between energy in 'yes' and 'no' responses. Upper right: confidence effects for motion energy in 'yes' and 'no' responses. Lower right: a subtraction between confidence effects 'yes' and 'no' responses. Shaded areas represent the the mean +- one standard error. The first 300 milliseconds of the trial are marked in yellow

Detection signal trials

A failure to find significant effects of sum motion energy on detection decision and confidence may be due to the fact that participants were sensitive to relative evidence (e.g., 'more dots are moving to the right') rather than to the sum motion along the relevant axis. However, as we mention above, for 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'. As a way to approximate participants' perception, we focused on detection signal trials. In these trials, 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.

Like discrimination decisions, detection decisions were most sensitive to perceptual evidence in the first 300 milliseconds of the trial (see Fig. 4.6, left panels). However, in contrast to discrimination, a positive evidence bias effect in detection 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 evidence for a positive evidence bias in confidence ratings ($t(9) = 0.11$, $p = .913$). To reiterate, while detection decisions were mostly sensitive to facilitating fluctuations in motion energy, confidence in detection 'yes' responses was equally sensitive to facilitating fluctuations in the true direction of motion, and to interfering fluctuations in the opposite direction 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 ($t(9) = -0.08$, $p = .941$).

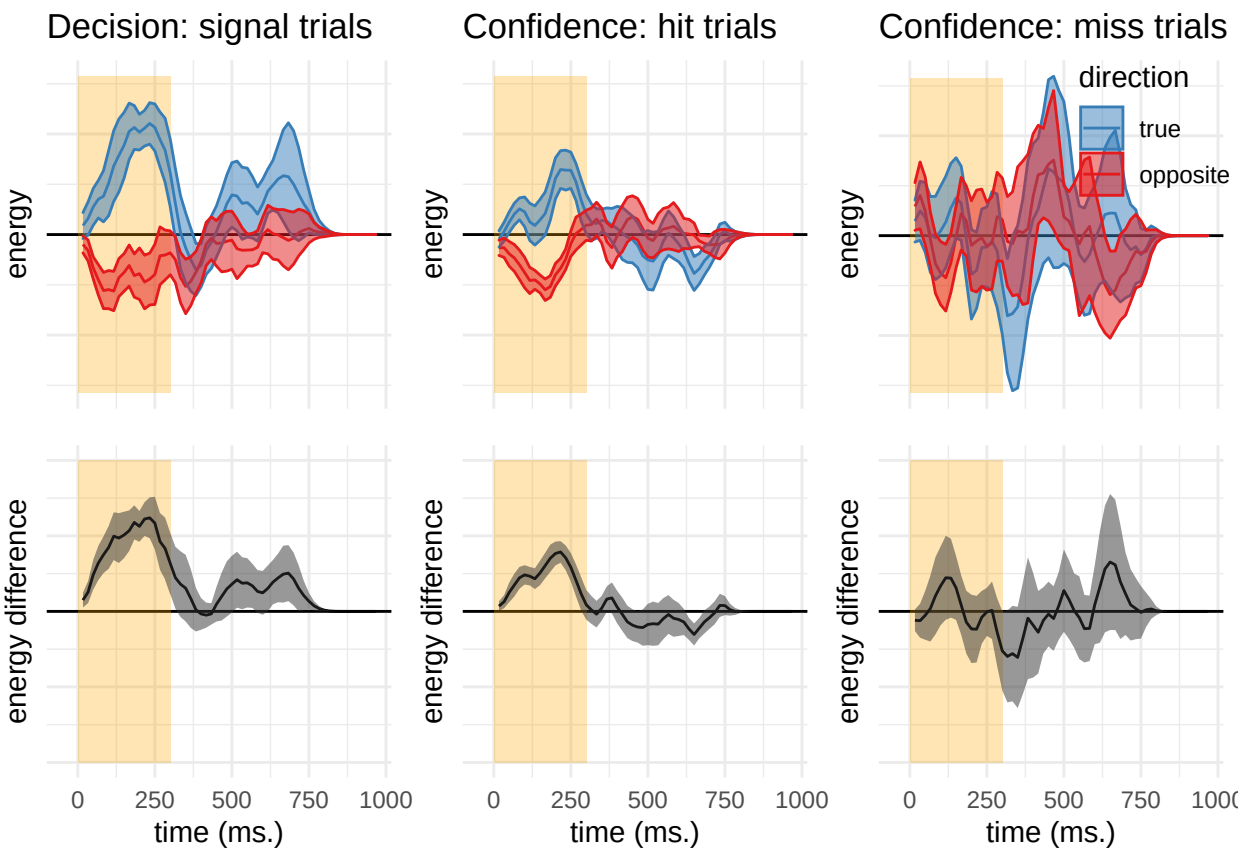


Figure 4.6: Decision and confidence detection kernels in signal trials, Experiment 1. Upper left: difference in motion energy between 'yes' and 'no' responses in the true (blue) and opposite (red) directions as a function of time. Upper middle and right: confidence effects for motion energy in the true and opposite directions for 'yes' and 'no' responses, respectively. Lower panels: the substraction 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

4.3 Experiment 2

In Exp. 1, we found that detection 'yes' responses are faster and are accompanied by higher subjective confidence than detection 'no' responses. We also replicated the metacognitive asymmetry between detection 'yes' and 'no' responses as measured with response-conditional ROC curves.

Examining random fluctuations in motion energy, we replicated the positive evidence bias in discrimination confidence, such that evidence in support of a decision was given more weight in the construction of confidence than evidence against it. This is consistent with the proposal that participants adopt a detection disposition when rating their confidence in discrimination responses. In detection, decision and

confidence were sensitive to fluctuations in motion energy at around the same time window as in discrimination. However, unlike discrimination, in detection a positive evidence bias was apparent in the decision, but not in the confidence kernels. Equal weighting of positive and negative evidence suggests that participants were rating their confidence not in the presence of a signal, but in its category. Furthermore, confidence in detection 'no' responses was not affected by fluctuations in motion energy.

In Experiment 2 we tested the robustness of these findings to a different type of stimuli (flickering patches) and mode of data collection (a ~10 minute online experiment). Specifically, our pre-registered objectives (see our pre-registration document: https://osf.io/8u7dk/) were to first, replicate the 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 positive or negative evidence on confidence in 'no' judgments.

4.3.1 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 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 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 consisted of two tasks (Detection and Discrimination) presented in separate blocks. A total of 56 trials of each task was 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 introduction section, which included instructions about the stimuli and confidence scale, four practice trials and four confidence practice trials. A second introduction section was presented before the second block. Introduction 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 given 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 Experiment 1 participants' perceptual thresholds for motion discrimination were highly similar, 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 allow for differences in task performance.

Trial structure In discrimination blocks, trial structure closely followed Experiment 2 from Zylberberg et al. (2012), with a few adaptations. Following a fixation cross (500 ms), a rapid serial visual presentation (RSVP) was be presented (12 frames, presented at 25Hz), consisting of two sets of four adjacent vertical gray bars, displayed to the left and right of the fixation cross (see Fig. 4.7). 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.

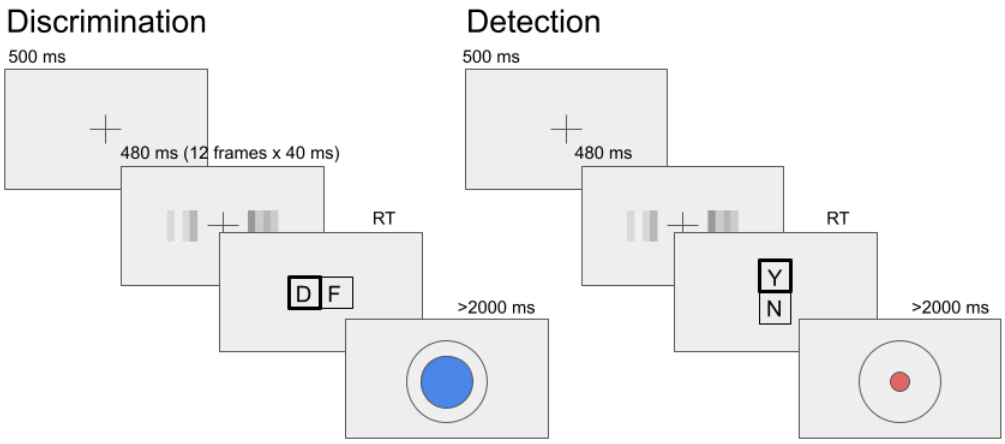


Figure 4.7: Task design for Experiment 2. In both tasks, participants viewed 480 milliseconds of two flicketing patches, after which they made a keyboard response to indicate which of the patches was bright (discrimination) or whether any of the patches was bright (detection).

Detection blocks were similar to discrimination blocks, with the exception that decisions were made about whether the average luminance of either of the two sets

was brighter than the gray backgroud, or not. In 'different' trials, 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 ('y' for 'yes' and 'n' for 'no'). Confidence ratings and feedback were as in the discrimination task.

4.3.2 Results

Response accuracy

Overall accuracy level 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 Experiment 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], $t(101) = -0.62$, $p = .537$).

Participants were also 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. 4.8, upper panel). 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. 4.8, lower panel). 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], $t(101) = -0.03$, $p = .974$).

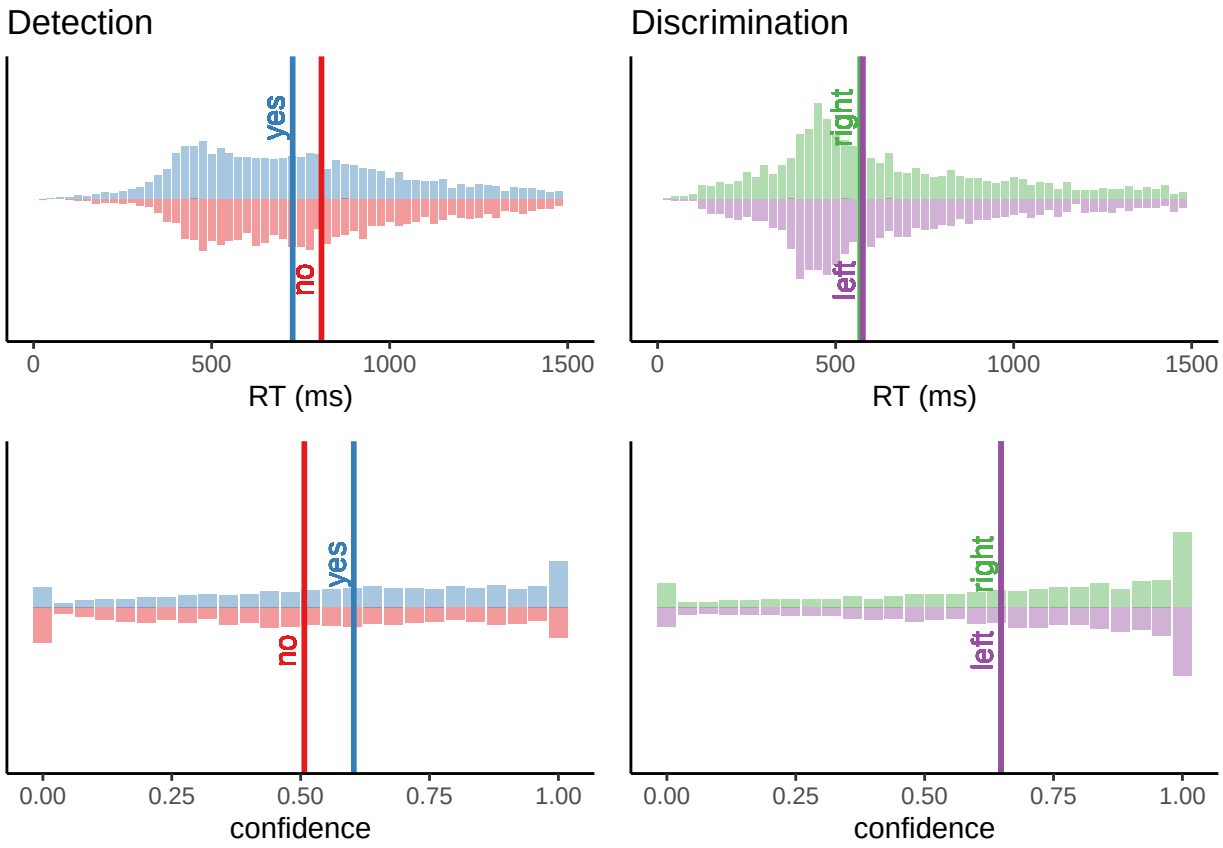


Figure 4.8: Response time (upper panel) and confidence (lower panel) histograms for the detection (left) and discrimination (right) tasks in Experiment 2. Vertical lines represent the median response time and the mean confidence rating for each response.

Response conditional ROC curves

In contrast to the results of Experiment 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. 4.9). In the Discussion, we discuss a candidate explanation for this null finding. Importantly, similar metacognitive sensitivity for 'yes' and 'no' responses should not affect the interpretation of our reverse correlation findings.

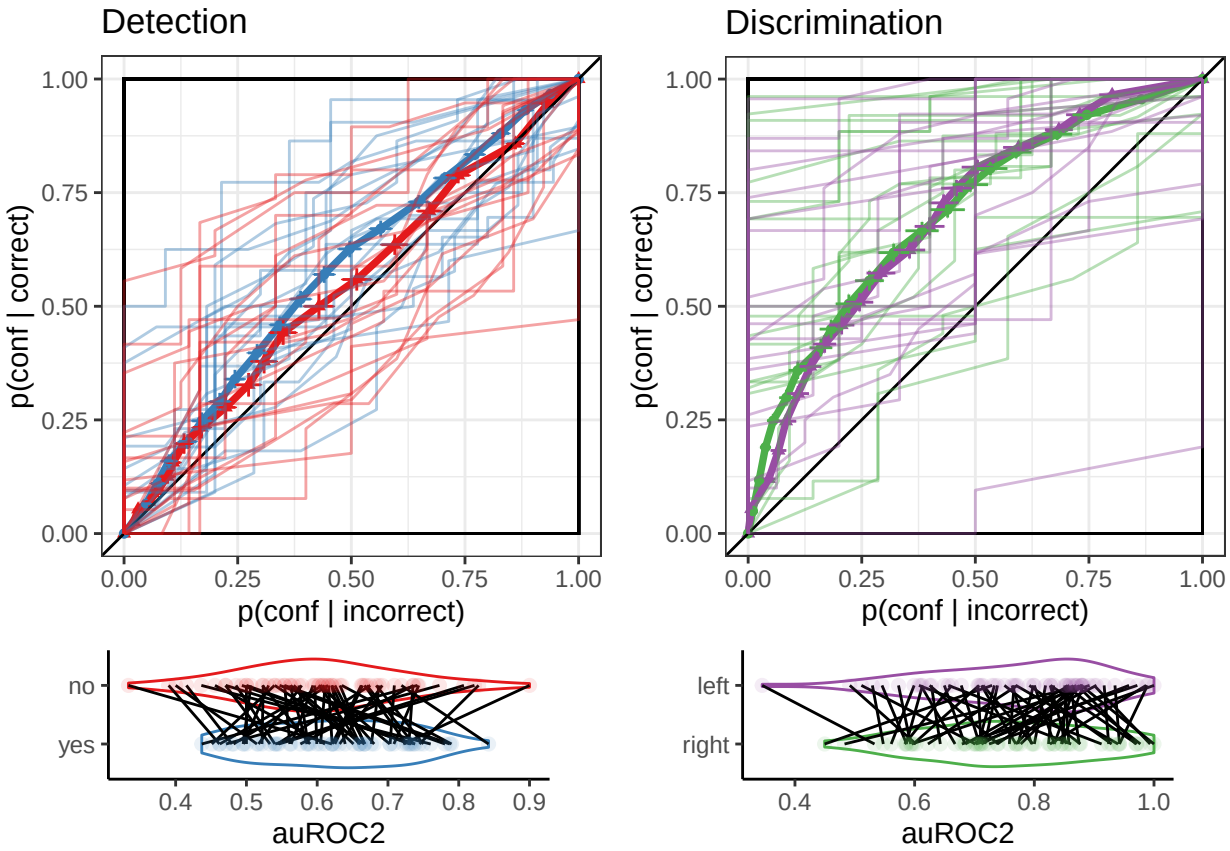


Figure 4.9: Response conditional ROC curves for the two tasks and four responses in Exp. 2. The area under the curve is a measure of metacognitive sensitivity. Lower panel: distributions of the area under the curve for the four responses, across participants. Error bars stand for the standard error of the mean.

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 of Exp 2. in Zylberberg et al. (2012).

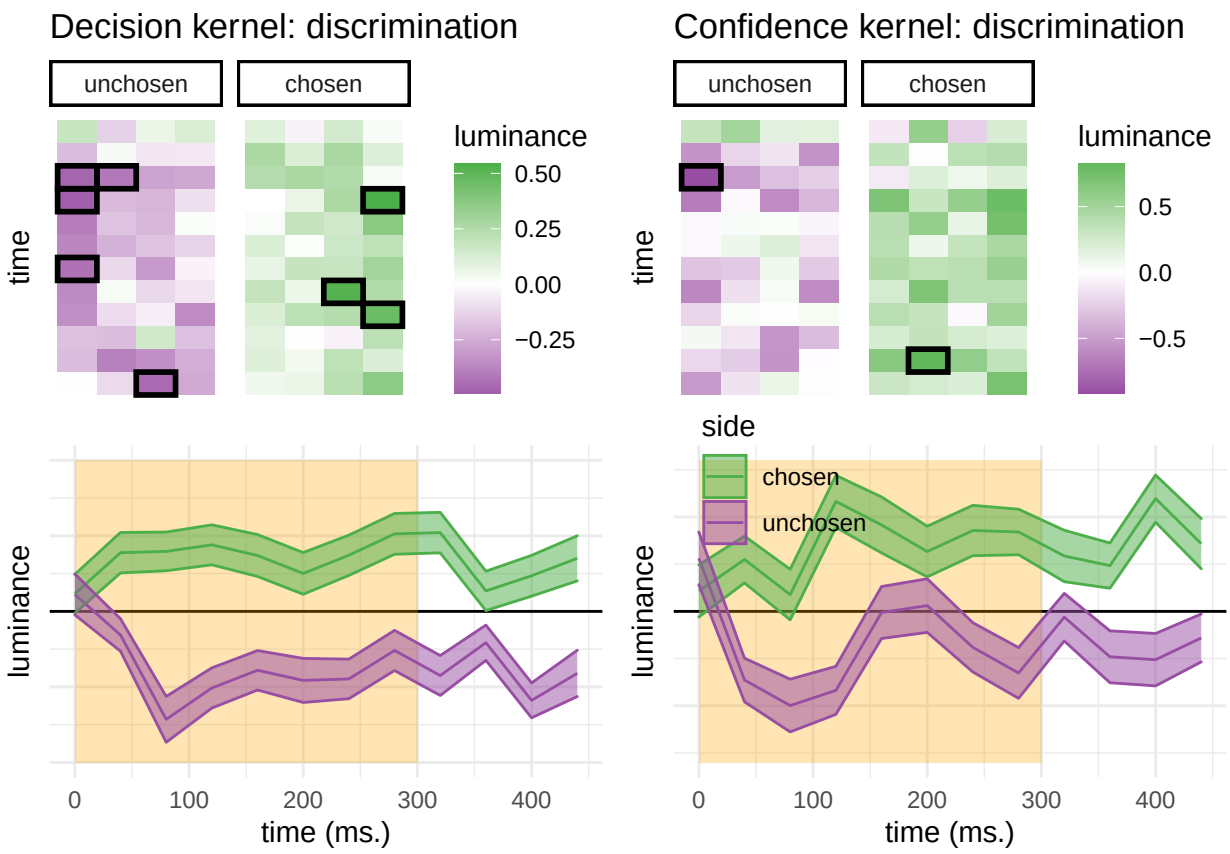


Figure 4.10: Decision and confidence discrimination kernels, Experiment 2. Upper panels: decision (left) and confidence (right) kernels for the flickering patch stimuli. Black frame signify a significant effect at the 0.05 significance level controlling for family-wise error rate across the 48 (12 timepoint x 4 positions) comparisons. Lower panels: decision and confidence kernels, averaged across the four bars to yield a single timecourse for the chosen (green) and unchosen (purple) stimuli. Shaded areas represent the the mean +- one standard error. The first 300 milliseconds of the trial are marked in yellow

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 motion energy fluctuations during the first 300 milliseconds of the trial ($t(101) = 10.98$, $p < .001$; see Fig. 4.10, left panels). As per our comment in section 4.2.3, in order to test for decision biases we need to divide evidence not based on participants' decision, but based on the true signal. Participants' decisions were significantly more sensitive to fluctuations in luminance in the foil compared with the signal stimulus within the first 300 miliseconds of the trial ($t(100) = -2.29$, $p = .024$).

Discrimination confidence We observed a significant effect of motion energy on confidence within the first 300 milliseconds of the stimulus ($t(100) = 7.14$, $p < .001$; see Fig. 4.10, right panels). Replicating Zylberberg et al. (2012), this effect was significantly stronger for motion energy in the chosen direction, compared to the unchosen direction ($t(100) = 2.56$, $p = .012$).

Detection decisions We pooled luminance values from both right and left stimuli and contrasted the resulting values as a function of detection response. The sum luminance had a significant effect on participants' responses during the first 300 milliseconds ($t(101) = 6.10$, $p < .001$; see Fig. 4.11, left panel), suggesting that participants were sensitive to sum evidence (overall luminance) in their detection responses.

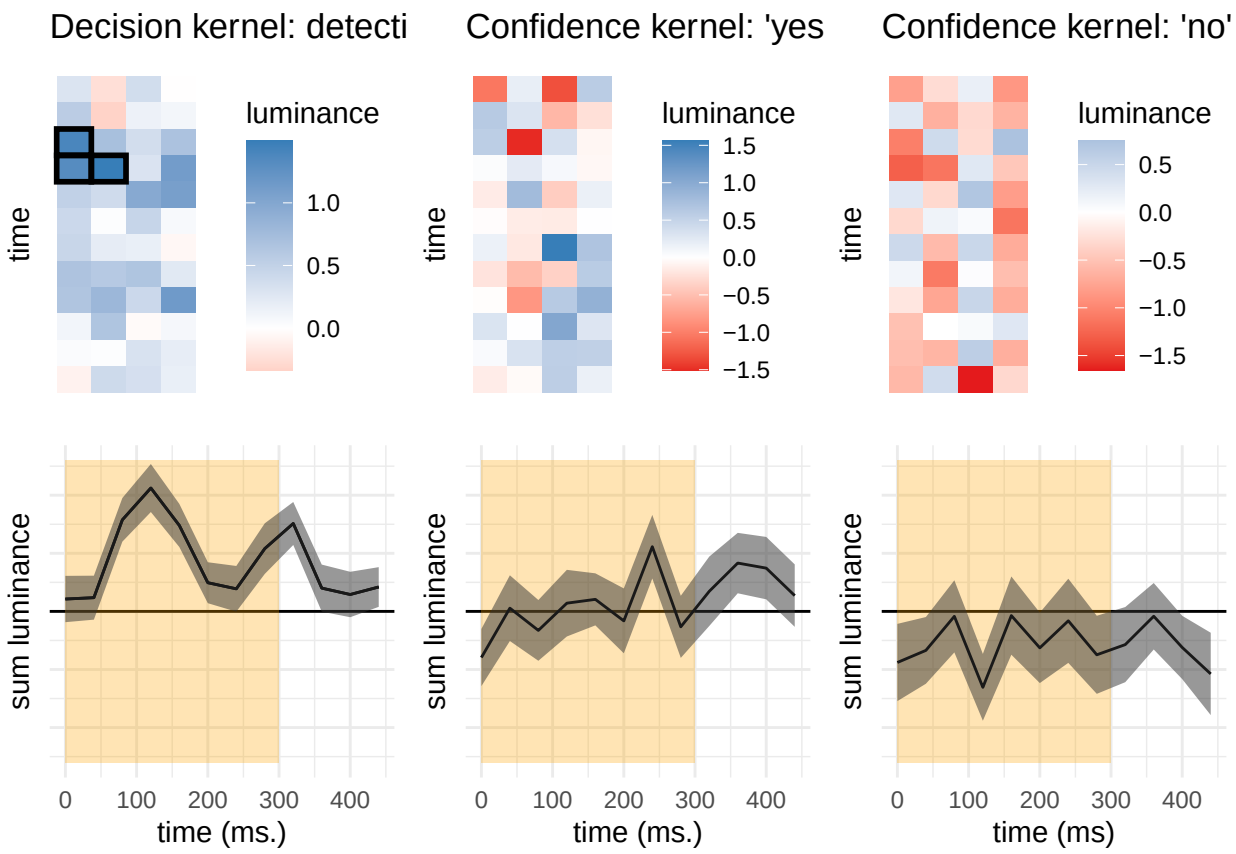


Figure 4.11: Decision and confidence detection kernels, Experiment 2. Upper panels: decision (left) and confidence (right) kernels for the flickering patch stimuli, showing the effect of overall luminance (across both stimuli) on decision and confidence. Black frame signify a significant effect at the 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. Shaded areas represent the the mean +- one standard error. The first 300 milliseconds of the trial are marked in yellow

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 effect on confidence in 'no' responses ($t(99) = -2.43$, $p = .017$; see Fig. 4.11, middle and right panels). As we show below, confidence in 'yes' responses was sensitive to the relative evidence for the two stimulus categories, rather than to the overall luminance of the screen. Our next analysis of detection signal trials diverged from our pre-registered plan. For the pre-registered analysis, see Appendix section B.1.

4.3.3 Detection signal trials

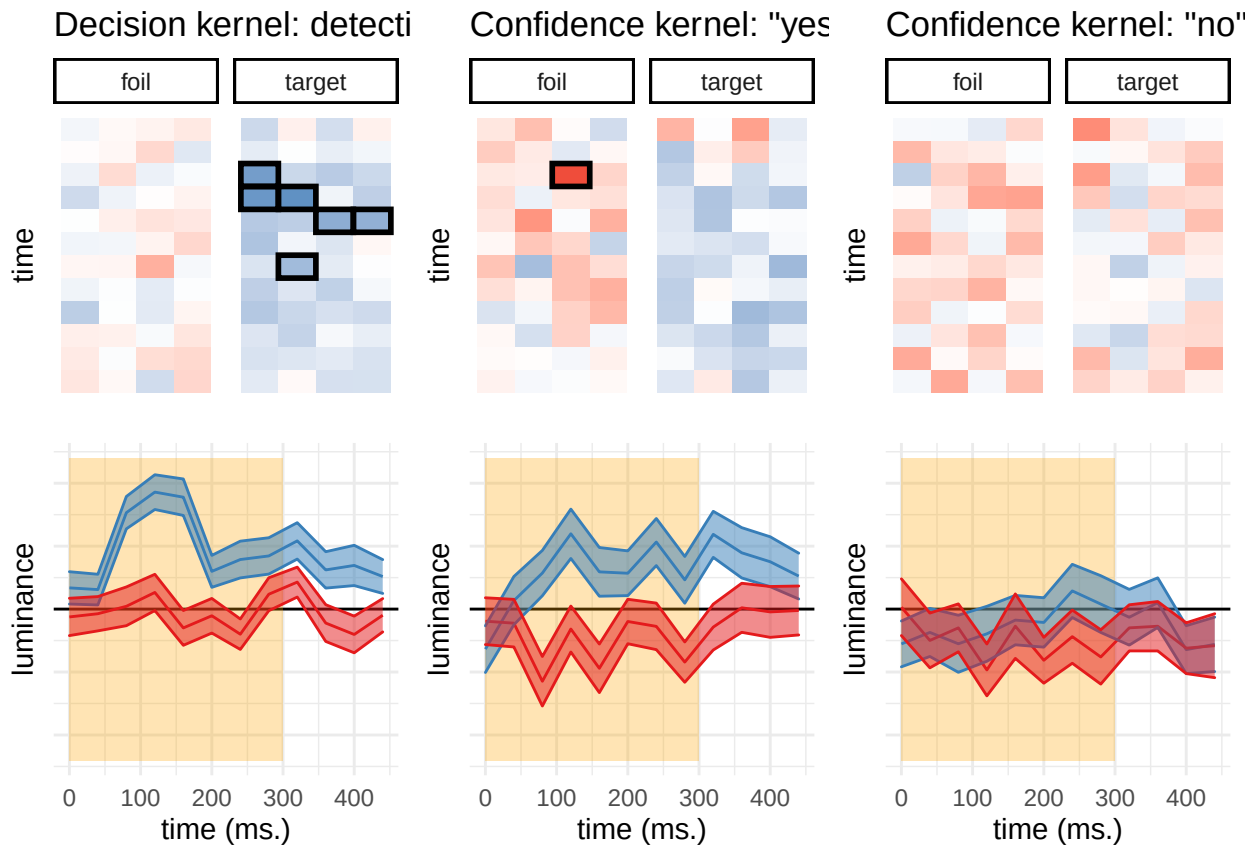


Figure 4.12: Decision and confidence kernels for detection signal trials, Experiment 2. Upper left: mean difference in luminance between 'yes' and 'no' responses for the target stimulus and foil stimuli. Upper middle and right panels: mean effect of luminance on confidence in the target and foil stimuli, in 'yes' and 'no' responses. Lower panels: the effects of luminance on decision and confidence, averaged across the four spatial locations. Shaded areas represent the the mean +- one standard error. The first 300 milliseconds of the trial are marked in yellow

We next focused on detection signal trials. In these trials, we could separate stimuli to a signal channel (the bright stimulus) and a noise channel (the foil), and ask how random variability in luminance in each channel affected detection decision 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 sensitive to positive noise in the bright patch, but not to negative noise in the foil 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; $t(99) = 5.08$, $p < .001$). Similar to the results of Exp. 1, detection confidence was not susceptible to a positive evidence bias ($t(99) = -0.12$,

$p = .901$). To reiterate, while detection decisions were mostly sensitive to facilitating noise, confidence in detection 'yes' responses was equally sensitive to facilitating noise in the target stimulus, and to interfering noise in the foil stimulus.

Consistent with the results of Exp. 1, confidence in 'miss' trials was independent of the contrast in luminance between the right and left stimuli ($t(98) = 1.26$, $p = .210$). However, as described in section 4.3.2, confidence in 'no' responses was sensitive to the overall luminance of the display. A negative effect of luminance on confidence in 'no' responses was significant for the foil stimulus ($t(98) = -2.64$, $p = .010$), and marginally significant for the target stimulus ($t(98) = -1.67$, $p = .099$). Importantly, for both stimuli higher confidence was associated with lower luminance values, consistent with our observation that confidence in detection 'no' responses was based on the overall darkness of the display, rather than on relative evidence.

4.4 Discussion

In two experiments, we compared participants' decisions and confidence ratings in discrimination and detection, matched for difficulty (Exp. 1) and signal strength (Exp. 2). In order to measure the contribution of perceptual evidence to confidence in detection and discrimination confidence ratings, we followed Zylberberg et al. (2012) and applied reverse correlation to noisy stimuli in perceptual decision making tasks. We fully replicated the main results of Zylberberg and colleagues: decision and confidence were affected mostly by perceptual evidence in the first 300 milliseconds of the trial, peaking at around 200 milliseconds. We also successfully replicated the positive-evidence bias: confidence in the discrimination task was more affected by supporting evidence than by conflicting evidence, giving rise to a 'positive evidence bias'. A positive evidence bias in discrimination confidence judgments may indicate that participants adopt a detection disposition in their metacognitive monitoring, and focus on sum evidence rather than relative evidence.

In both experiments, evidence accumulation for detection responses had a similar temporal profile to that of discrimination. However, detection decisions but not confidence ratings 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. In both experiments, relative evidence contributed to decision confidence in 'yes' responses, but was ignored in 'no' responses. Finally, in Experiment 2, but not in Experiment 1, sum evidence (the overall luminance of the display) significantly contributed to confidence in 'no' responses. Below we explore the predictions of several Bayes-rational models and their alignment with our observations.

4.4.1 Model 1: a rational agent + symmetric evidence structure

The first model made optimal decisions based on the likelihood ratio between the two hypotheses. This model had full access to the stimulus. Stimuli were modeled as

ordered pairs of numbers, corresponding to the two sensory channels (for example, right and left motion, or right and left flickering patch). For simplicity, we ignored the temporal and spatial dynamics of evidence accumulation in our simulations, and focused on the general patterns of evidence weightings instead. In noise trials, both numbers were modeled as sampled from a normal distribution with mean 0 and standard deviation 1 ($E_n \sim \mathcal{N}(0,1)$). In signal trials, one of the two numbers was sampled from a normal distribution with mean 1 ($E_s \sim \mathcal{N}(1,1)$). The agent observes the two numbers, and decides (based on the likelihood ratio, and having full access to the true underlying distributions) if a stimulus was present or not (detection), or which of the two numbers was sampled from the signal distribution (detection). Their confidence is then proportional to the log likelihood ratio between the two hypotheses (signal presence of absence, or signal 1 or 2).

This model makes accurate predictions for the contribution of positive and negative perceptual evidence to discrimination and detection decisions: equal in discrimination, but asymmetric for detection (see Fig. 4.13. However, its predictions for confidence ratings are the exact opposite of what we observe in our data. The model predicts a positive evidence bias in detection confidence ratings, but we find symmetrical confidence kernels for detection confidence. In discrimination, where the model predicts equal contribution of positive and negative evidence to confidence, we find a significant positive evidence bias.

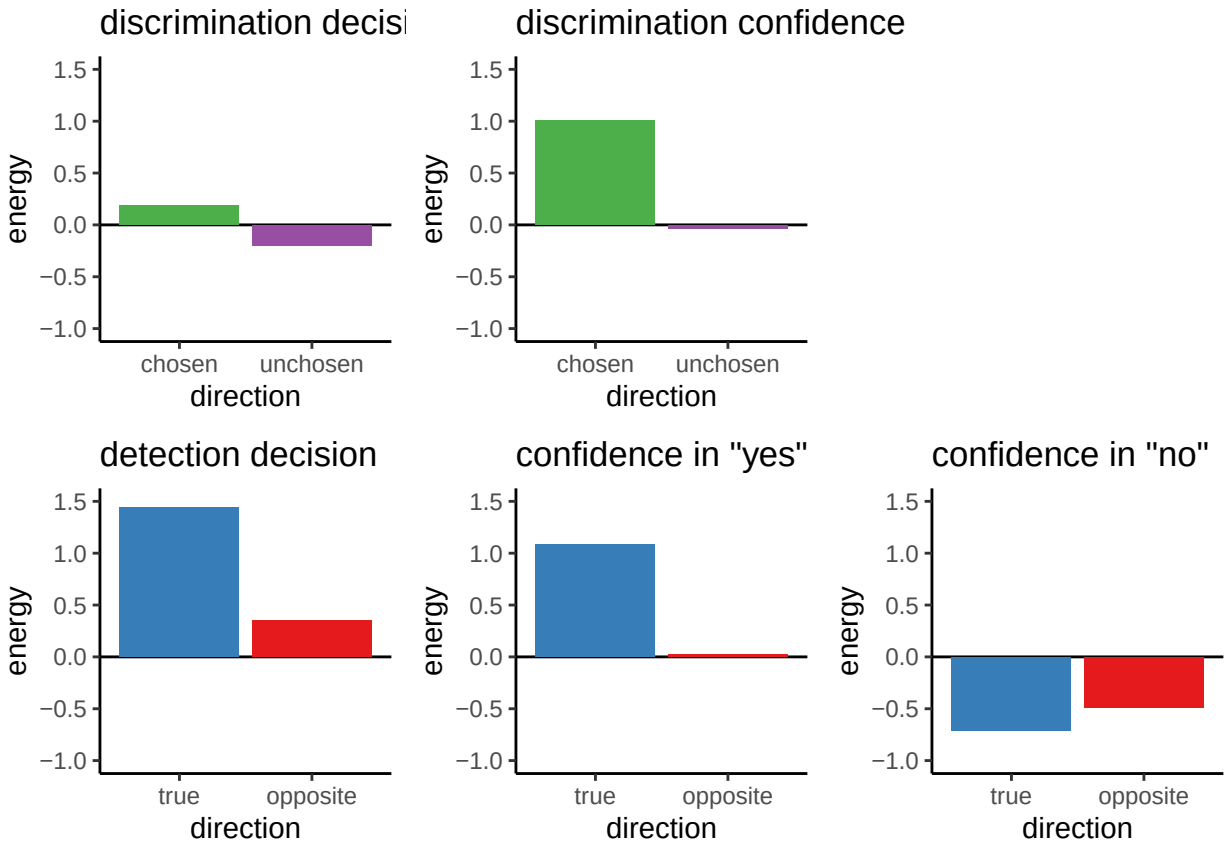


Figure 4.13: Simulated reverse-correlation analysis in Model 1. A bias emerges in detection, but not in discrimination confidence ratings - the opposite of what we observe.

4.4.2 Model 2: a rational agent + symmetric evidence structure

One possible driver of the positive evidence bias in confidence ratings is higher informational value in signal than in noise, such that 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 3, sampling the value 6 is much more informative than sampling the value -2, because the first is only likely if sampled from the signal distribution (likelihood ratio $>$ 1,000,000), but the second is likely under both distributions (likelihood ratio $=$ 1). 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 et al. (2012), the number of

coherently moving dots was itself randomly determined, sampled from a Gaussian distribution once in 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 based on the Weber-Fechner law (Fechner & Adler, 1860) and from the coupling between firing rate and firing rate variability implied by the Poisson form of neuronal firing distributions.

To obtain qualitative predictions, we simulated an unequal-variance first-order SDT model (full simulation details, including the source python code are available in appendix **??**). This model was identical to model 1 with one exception. In this model the artificial agent had access only to a degraded version of the two sensory samples, corrupted by additional noise. To model the unequal variance nature of the perception of signal and noise, this perceptual noise was sampled from a normal distribution with mean 0 and a standard deviation proportional to the magnitude of the sensory sample ($x' = x + \epsilon; \epsilon \sim \mathcal{N}(0, 0.5 \times x)$). The 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. **??**. Simulated agents were more sensitive to variations in the signal channel for deciding whether a signal was present or not, and when rating their confidence in discriminating between two stimulus classes. However, in contrast with the observed data, our unequal-variance model also predicted a positive evidence bias in detection confidence ratings and an effect of relative evidence on confidence in 'no' responses, which we do not observe in the actual data.

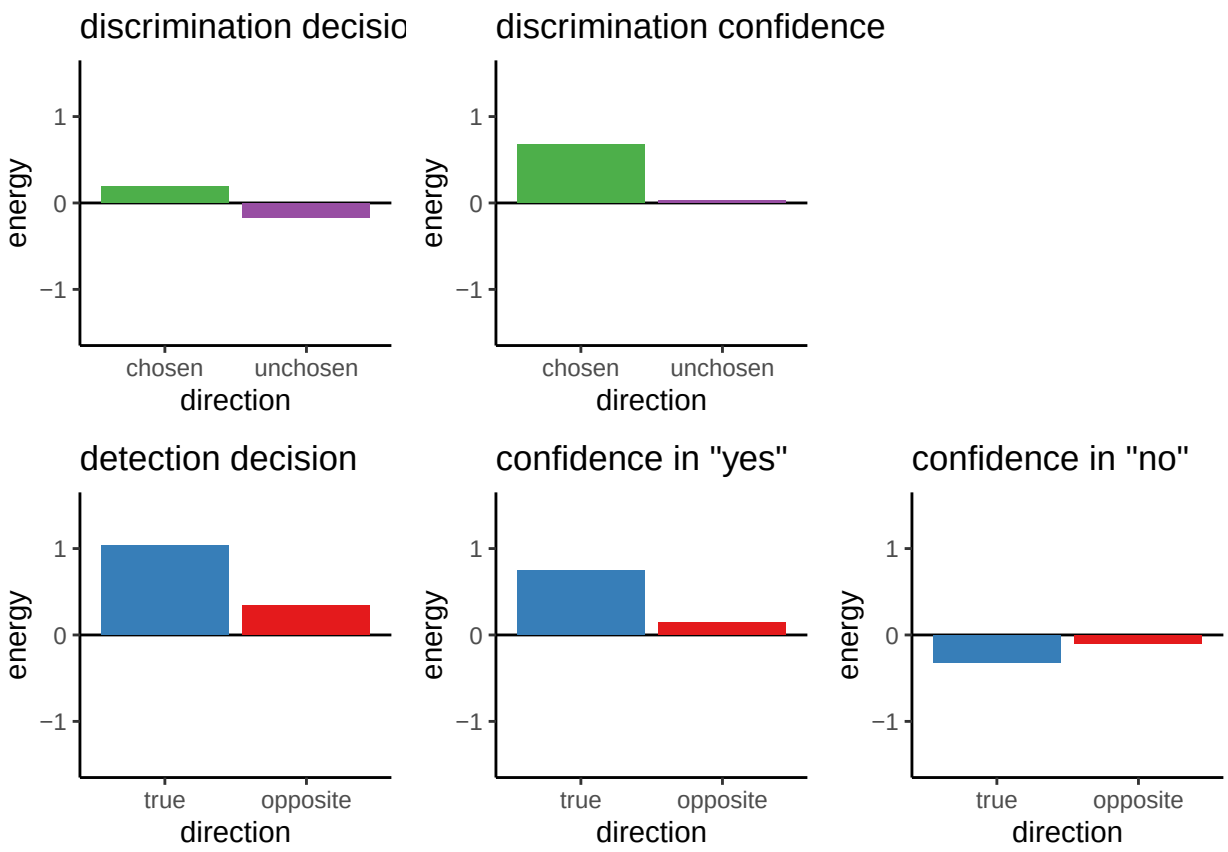


Figure 4.14: Simulated reverse-correlation analysis in Model 2. A bias emerges in detection as well as in discrimination confidence ratings, in contrast to our finding of symmetrical confidence kernels in detection.

4.4.3 Model 3: confidence decision cross

Models 1 and 2 described the behaviour of a rational agent but were unsuccessful in accounting for the mismatch between decision and confidence kernels. Model 3 drops the rationality assumption. This model is identical to Model 1 when it comes to the modeling of perceptual samples and the decision process. However, when coming to rate its confidence in a discrimination judgment, this model extracts the Log Likelihood Ratio not between stimulus category 1 and 2, but between signal presence or absence. Similarly, confidence in discrimination judgments is based on the Log Likelihood Ratio between the presenec of stimulus 1 or 2.

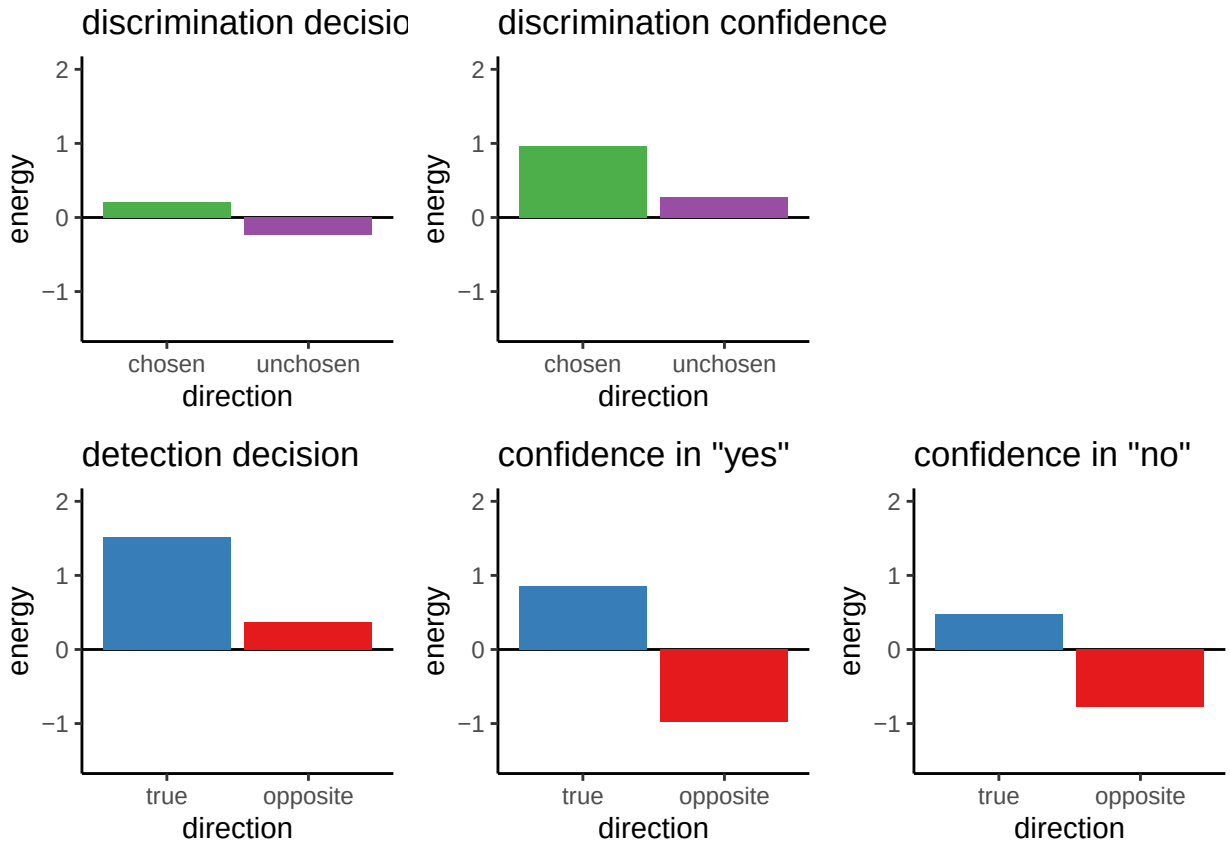


Figure 4.15: Simulated reverse-correlation analysis in Model 3. A positive evidence bias emerges in discrimination confidence ratings, and a negative evidence bias emerges in detection confidence ratings. This is in contrast to our finding of symmetrical confidence kernels in detection.

4.4.4 Evidence for absence

The results of the two experiments were highly similar, with two exceptions. First, the observed metacognitive asymmetry between confidence judgments for detection responses in Experiment 1 was not replicated in Experiment 2. In the second experiment, participants had similar metacognitive insight into their judgments about target presence and absence. Second, in Exp. 1 we found no effect of stimulus energy on confidence judgments in detection 'no' responses, whereas in Experiment 2 participants were more confident in the absence of the stimulus when overall stimulus energy was low. We suggest that these two observations may be related, and that the difference may lie in the availability of evidence for absence in the two experiments.

In Exp. 2, signal presence was defined as one of the flickering patches being brighter than the gray background. This meant that participants could be highly confident in the absence of a signal when both stimuli were particularly dark. This is what we observe in our reverse correlation analysis of detection 'no' responses (Fig. 4.11 and Fig. 4.12, right panels). In contrast, in Exp. 2 the presence of a signal could

mean coherent motion to one of two opposite directions. This means that evidence for absence was never available: the opposite of the presence of rightward motion is leftward motion, not random motion. Indeed, motion energy had no effect on confidence in 'no' responses in Exp. 1 (Fig. 4.5 and Fig. **??**, right panels).

The availability of positive evidence for signal absence may have boosted metacognitive sensitivity for detection 'no' responses in Exp. 2. Interestingly, however, even in Experiment 2, overall confidence in absence was lower than in presence with a similar effect size to that of Exp. 1 (mean differences of 0.08 and 0.10 of the confidence scale in Exp. 1 and 2, respectively), and 'yes' responses were faster on average (median differences -85.37 and -77.12). This may hint to the fact that RT and confidence differences between judgments of presence and absence are unrelated to the informational asymmetry between evidence for presence and for absence.

In summary, in two experiments we replicated the positive evidence bias for discrimination confidence judgments and found a similar bias in detection decisions. A first-order unequal variance framework accounted for this, but failed to account for the absence of a positive evidence bias for confidence judgments in signal presence: participants were more confident in the presence of a signal not only when the true signal was stronger, but also when the opposite signal was weaker. Our findings hint at a qualitative difference in the way subjects evaluate evidence for presence, absence, stimulus class.

In both experiments, detection 'yes' responses were faster on average, and accompanied by higher levels of subjective confidence compared with detection 'no' responses. In contrast, discrimination responses were similar at the group level. These behavioural asymmetries are in line with the classic interpretation of detection responses: 'yes' responses reflect the successful accumulation of evidence for signal presence, and 'no' responses reflect a failure to accumulate such evidence rather than the successful accumulation of evidence for signal absence.

General Discussion

In this thesis I investigated inference about absence in visual perception, and its relation with self-modeling and default-mode reasoning. In chapter 1 I focused on the first few-trials in a visual search task to trace the origins of the metacognitive knowledge that allows subjects to efficiently decide that visual items are missing from a display. In Chapter 4 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 **??** 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 @(ref:asymmetry) 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 will provide a summary of the results from all chapters. I will then evaluate my original proposal, that inference about absence critically relies on self-knowledge, in light of my findings. Specifically, I will examine alternative interpretations and first-order accounts of inference about absence. Before concluding, I will briefly describe two directions for future research that build on and extend my work here.

4.5 Summary of results

4.6 What I didn't find

4.6.1 Chapter 1: no correlation with explicit metacognition

4.6.2 Chapter 2: no effect of confidence in signal presence

mention project with Roy

4.6.3 Chapter 3: small differences in brain activity between inference about absence and presence

4.6.4

4.7 Future directions

4.7.1 Failures of a self-model

4.8 Conclusion

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).

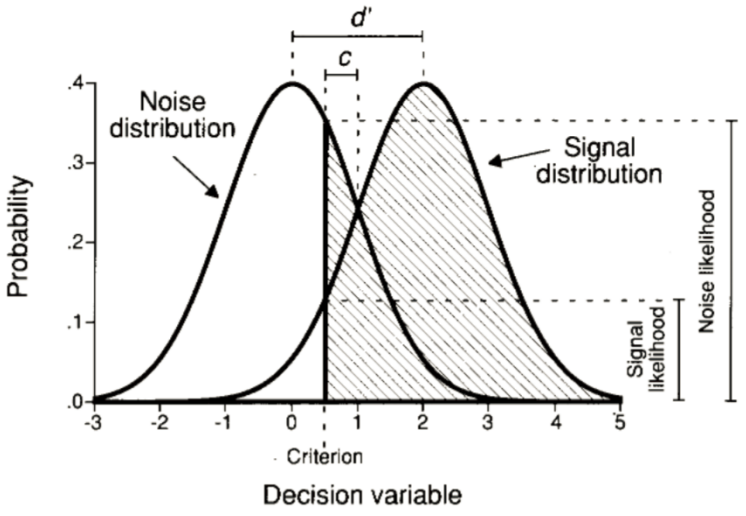


Figure A.1: Distribution of the decision variable across noise and signal trials, showing *d'*, *c*, and the likelihoods. Figure from @stanislaw1999calculation

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).

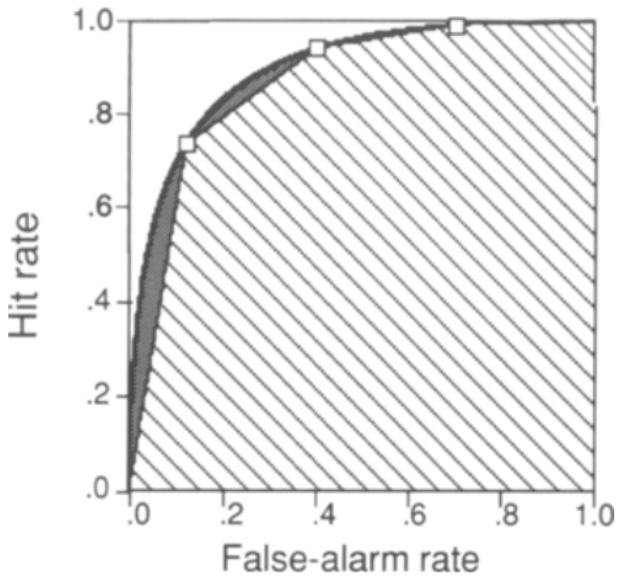


Figure A.2: Receiver Operating Characteristic (ROC) curve. Three points on the ROC curve are shown (open squares). The area under the curve, as estimated by linear extrapolation, is indicating by hatching; the true area includes the gray regions. Figure from @stanislaw1999calculation

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 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.

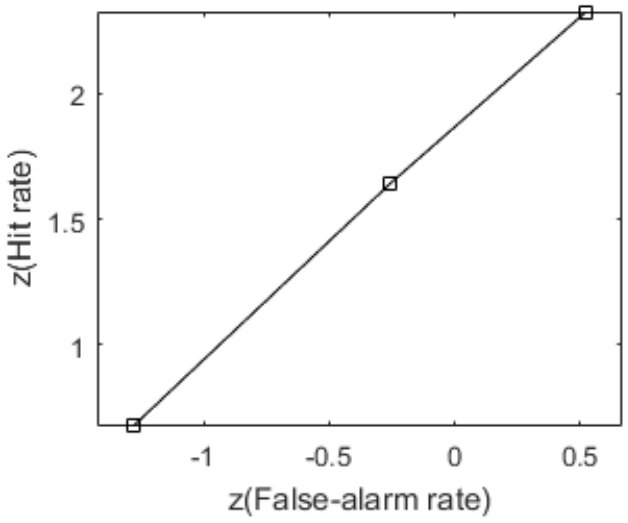


Figure A.3: zROC curve

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 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).

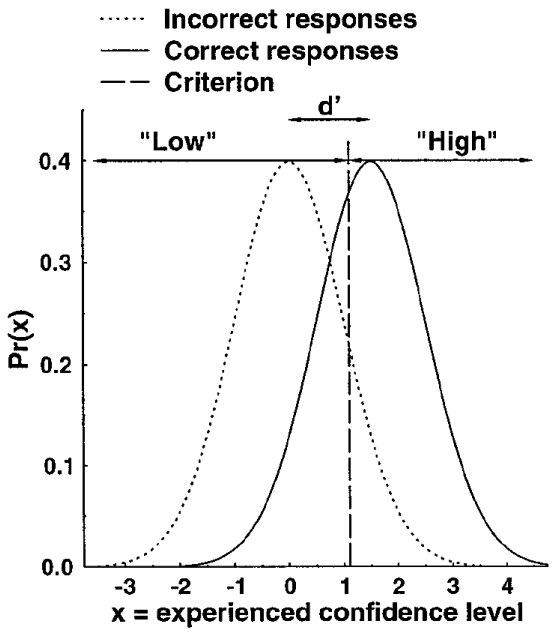


Figure A.4: A second order SDT model: confidence judgments are assumed to result from a process that uses an internal variable to separate correct from incorrect responses. Figure from @kunimoto2001confidence

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`.

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

Appendix B

Supp. materials for ch. 2

B.1 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 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.

B.1.1 Exp. 1

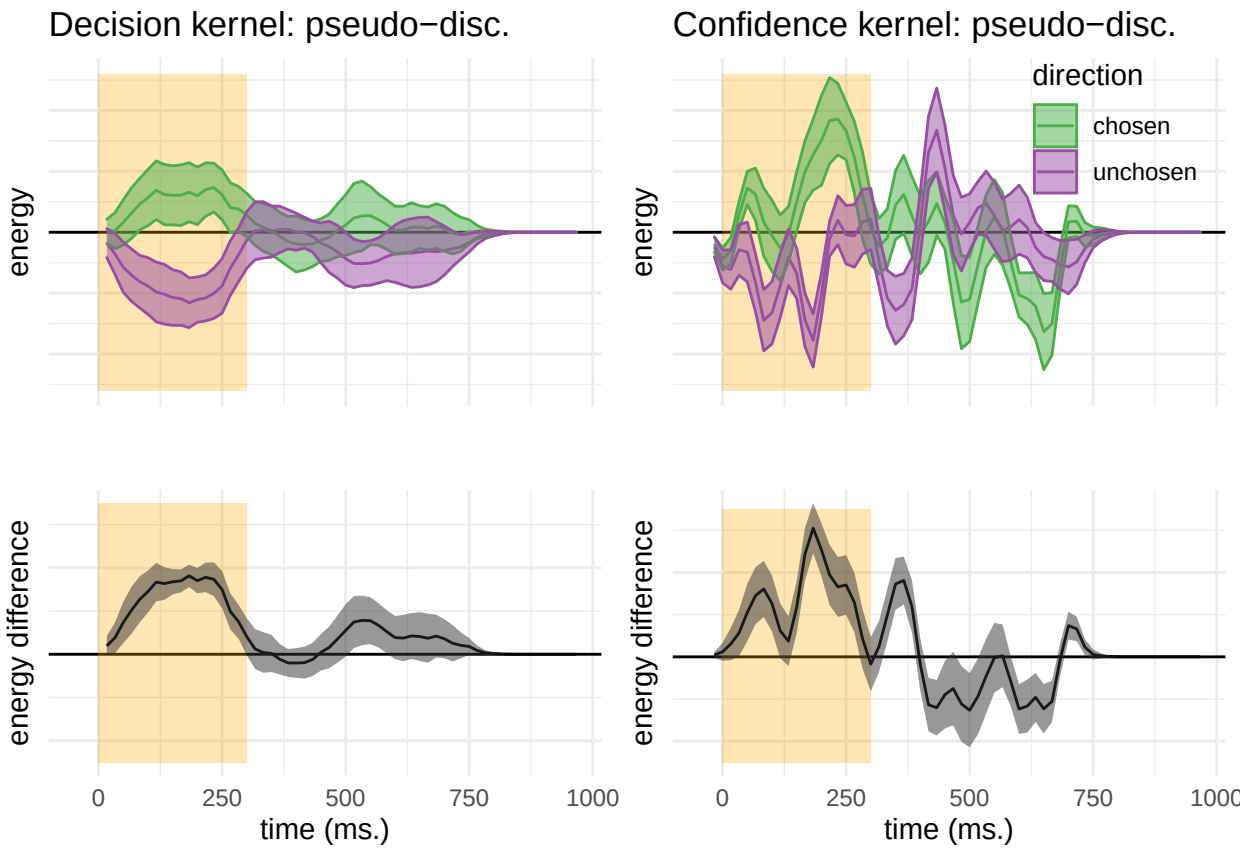


Figure B.1: Decision and confidence pseudo-discrimination kernels, Experiment 1. Upper left: motion energy in the "chosen" (green) and "unchosen" (purple) direction as a function of time. Bottom left: a subtraction between energy in the "chosen" and "unchosen" directions. Upper right: confidence effects 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

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 ($t(9) = 3.26$, $p = .010$). 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.

B.1.2 Exp. 2

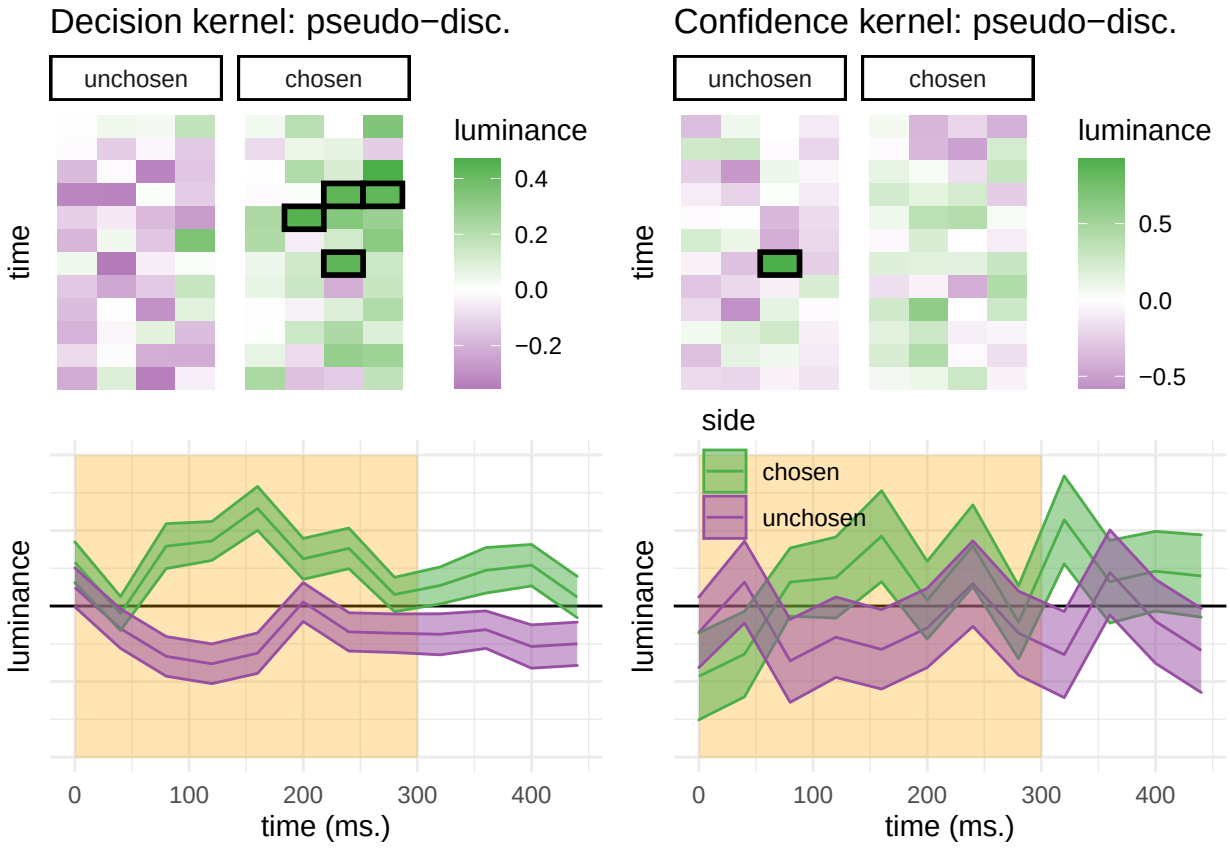


Figure B.2: Decision and confidence pseudo-discrimination kernels, Experiment 2. Upper left: luminance in the "chosen" (green) and "unchosen" (purple) stimulus as a function of time and spatial position. Bottom left: decision kernel averaged across the four spatial positions. Upper right: confidence 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 ($t(99) = -0.66$, $p = .512$).

B.2 Unequal-variance model

B.2.1 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.

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

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:

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

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

Inference

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

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

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

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

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

        self.df = pd.DataFrame()
```

```
        self.mu = mu
        self.sigma = sigma
        self.noise_factor = noise_factor

      ## The agent has full access to valid expectations about stimulus energy in

        self.signalHist = pd.DataFrame()
        self.signalHist['input'] = [np.random.normal(self.mu[1], self.sigma**2) fo
        self.signalHist['variance'] = self.signalHist.apply(lambda row: abs(row.in
        self.signalHist['percept'] = self.signalHist.apply(lambda row: row.input+n
        hist = np.histogram(self.signalHist.percept.to_numpy(), bins=1000)
        self.signal_dist = stats.rv_histogram(hist);

        self.noiseHist = pd.DataFrame()
        self.noiseHist['input'] = [np.random.normal(self.mu[0], self.sigma**2) for
        self.noiseHist['variance'] = self.noiseHist.apply(lambda row: self.noise_f
        self.noiseHist['percept'] = self.noiseHist.apply(lambda row: row.input+np.
        hist = np.histogram(self.noiseHist.percept.to_numpy(), bins=1000)
        self.noise_dist = stats.rv_histogram(hist)

    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.binomia

        self.getMotionEnergy()

        self.extractLLR()

        self.makeDecision()

        self.rateConfidence()

        self.df['correct'] = self.df.apply(lambda row: row.direction==row.decision

        #energy in chosen direction
        self.df['E_c'] = self.df.apply(lambda row: row.E_r if row.decision=='r' el

        #energy in unchosen direction
        self.df['E_u'] = self.df.apply(lambda row: row.E_l if row.decision=='r' el

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

```
                                self.sigma**2), axis=1)

        self.df['E_l'] = self.df.apply(lambda row: np.random.normal(self.mu[1] if row.di
                                self.sigma**2), axis=1)

    # how it appears to subjects
        self.df['E_ra'] = self.df.apply(lambda row: row.E_r+np.random.normal(0, abs(row.

        self.df['E_la'] = self.df.apply(lambda row: row.E_l+np.random.normal(0, abs(row.

    def extractLLR(self):

        # extract the Log Likelihood Ratio (LLR) log(p(Er|r))-log(p(Er|l)) + log(p(El|
        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 just chooses the direction associated with hi
        self.df['decision'] =  self.df.apply(lambda row: 'r' if row.LLR>0 else 'l', axis

    def rateConfidence(self):

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

B.2.2 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.

$$LLR = log(0.5 \times p([x_t'^1, x_t'^2]|stim = [x^s, x^n]) + 0.5 \times p([x_t'^1, x_t'^2]|stim = [x^n, x^s])) - logp([x_t'^1, x_t'^2]|stim = [x^n, x^n] \tag{B.7}$$

$$decision_t = \begin{cases} 1, & \text{if } LLR > 1. \\ 2, & \text{else.} \end{cases} \tag{B.8}$$

$$confidence_t = |LLR| \tag{B.9}$$

```
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.binomia

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

        self.getMotionEnergy()

        self.extractLLR()

        self.makeDecision()

        self.rateConfidence()

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

         #energy in true direction
        self.df['E_t'] = self.df.apply(lambda row: row.E_r if row.direction=='r' e

        #energy in opposite direction
        self.df['E_o'] = self.df.apply(lambda row: row.E_l if row.direction=='r' e

    def getMotionEnergy(self):
     # sample the motion energy for left and right as a function of the true dire
        self.df['E_r'] = self.df.apply(lambda row: np.random.normal(self.mu[1] if
                                self.sigma**2), axis=1)

        self.df['E_l'] = self.df.apply(lambda row: np.random.normal(self.mu[1] if
                                self.sigma**2), axis=1)

    # how it appears to subjects
        self.df['E_ra'] = self.df.apply(lambda row: row.E_r+np.random.normal(0, ab

        self.df['E_la'] = self.df.apply(lambda row: row.E_l+np.random.normal(0, ab

    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 highe
        self.df['decision'] =  self.df.apply(lambda row: 'p' if row.LLR>0 else 'a', axis

    def rateConfidence(self):

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

Appendix C

Supp. materials for ch. 3

C.1 Confidence button presses

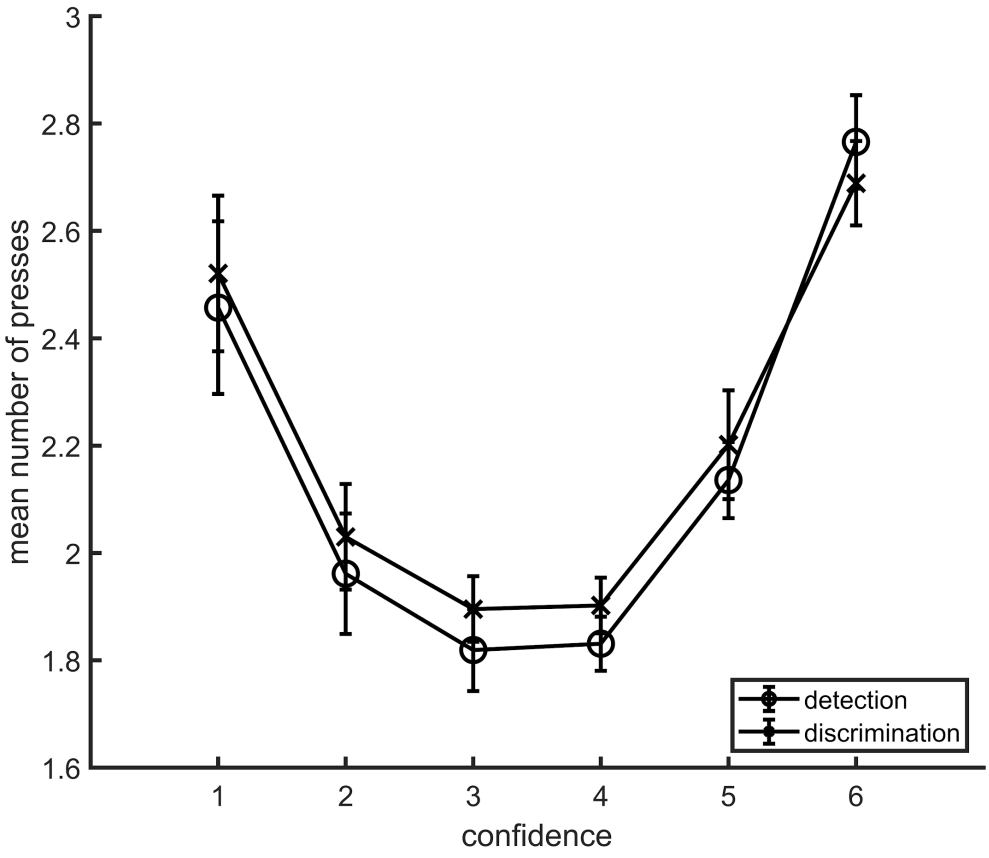


Figure C.1: Average number of button presses for each confidence level, as a function of task. More button presses were needed on average to reach the extreme confidence ratings, hence the quadratic shape. No difference between the two tasks was observed in the mean number of button presses for any of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

C.2 zROC curves

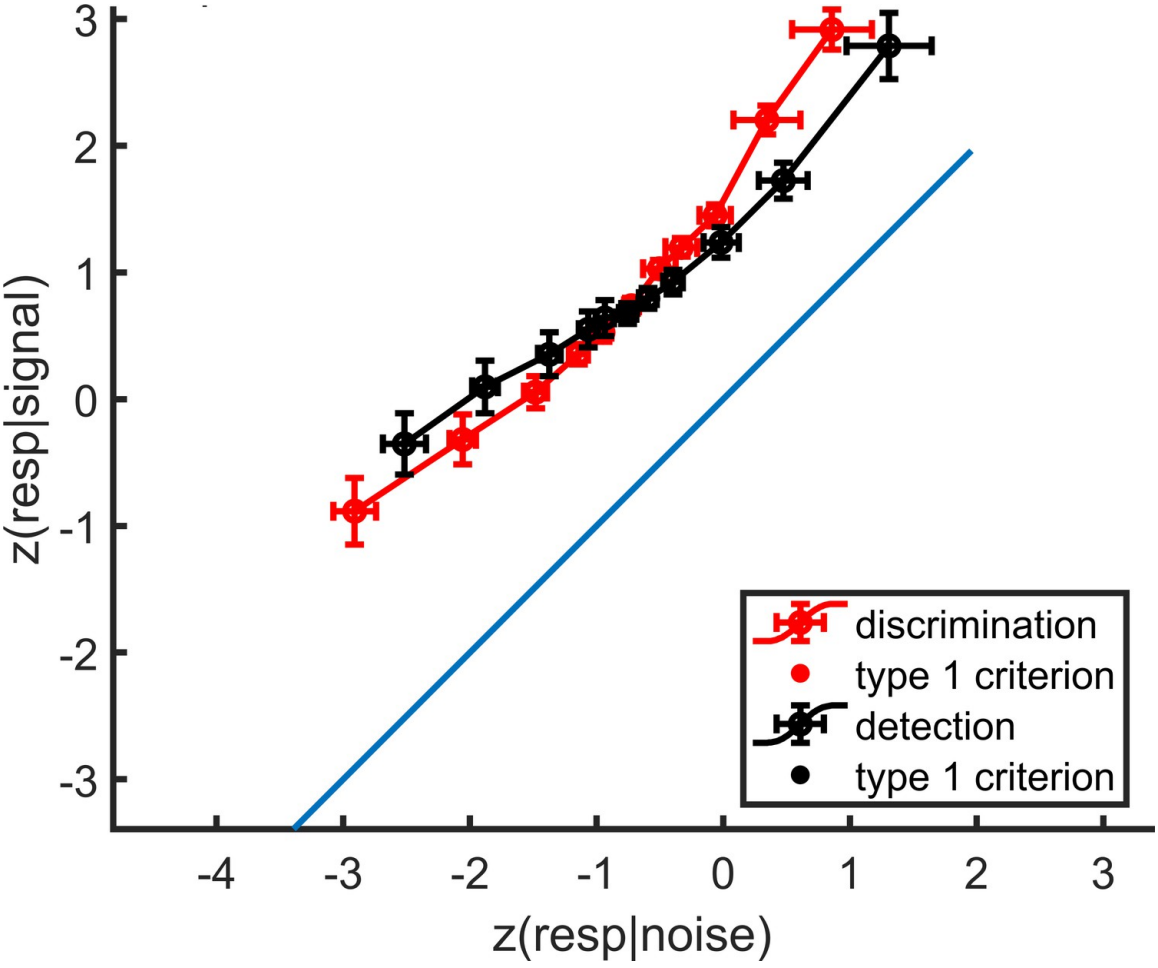


Figure C.2: mean zROC curves for the discrimination and detection tasks. As expected in a uv-SDT setting, the discrimination curve is approximately linear with a slope of 1, and the detection curve is approximately linear with a shallower slope. Error bars represent the standard error of the mean.

C.3 Global confidence design matrix

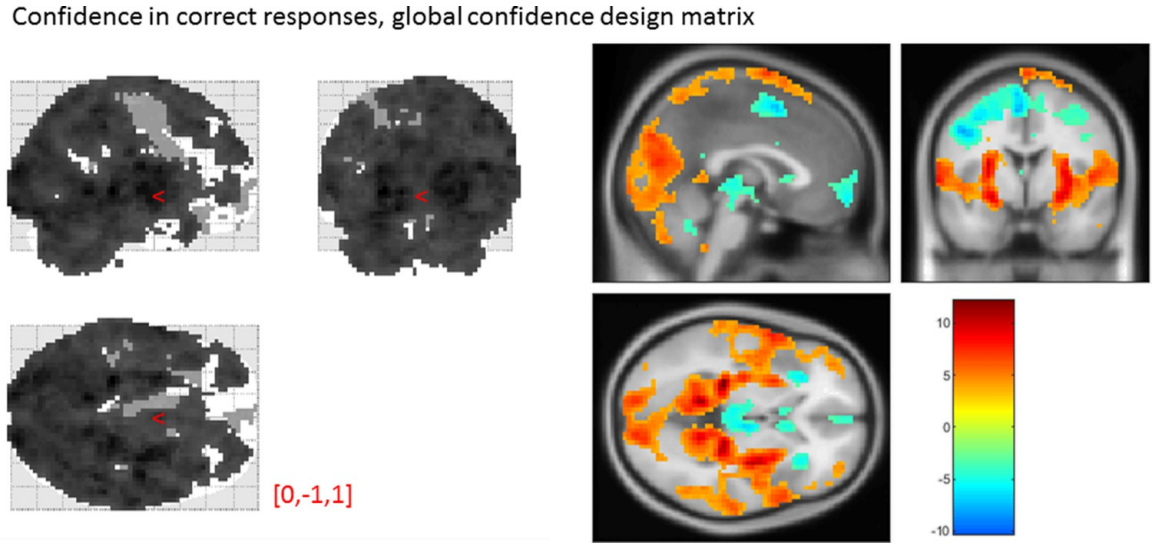


Figure C.3: Effect of confidence in correct responses, from the global-confidence design matrix. Uncorrected, thresholded at p<0.001. Left: glass brain visualization of the whole brain contrast. Right: yellow-red represent a positive correlation with subjective confidence ratings, and green-blue represent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

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

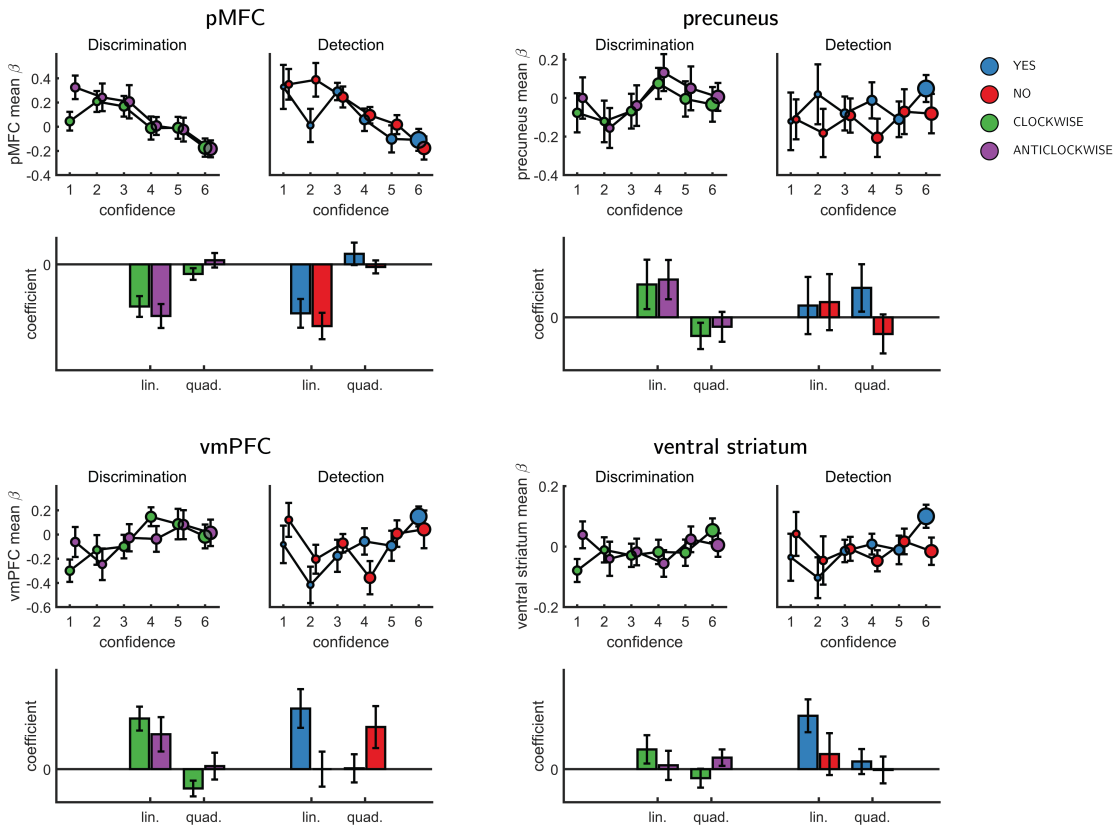


Figure C.4: Effect of confidence in all 4 ROIs, as a function of task and response, as extracted from the categorical design matrix. No significant interaction between the linear or quadratic effects and task or response was observed in any of the ROIs.

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

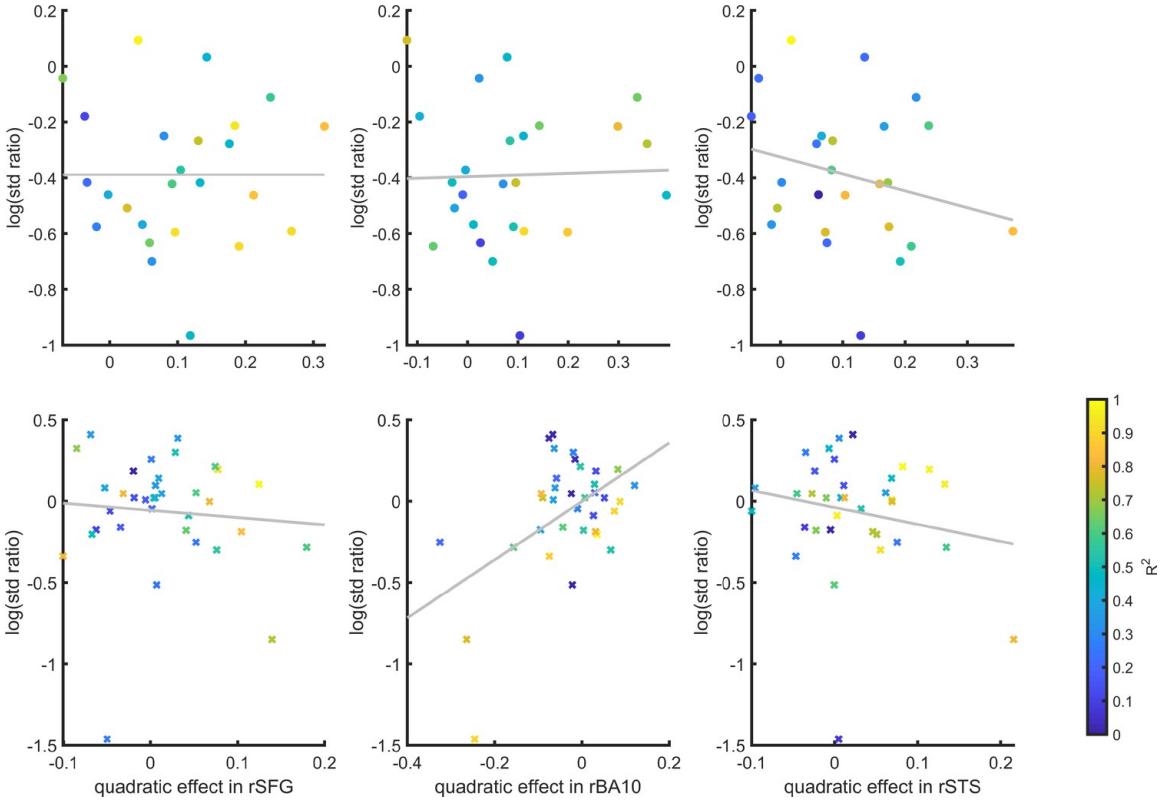


Figure C.5: Inter-subject correlation between the quadratic effect in the right hemisphere clusters and the ratio between the detection (top panel) and discrimination (lower panel) distribution variances, as estimated from the zROC curve slopes in the two tasks. Marker color indicates the goodness of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

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

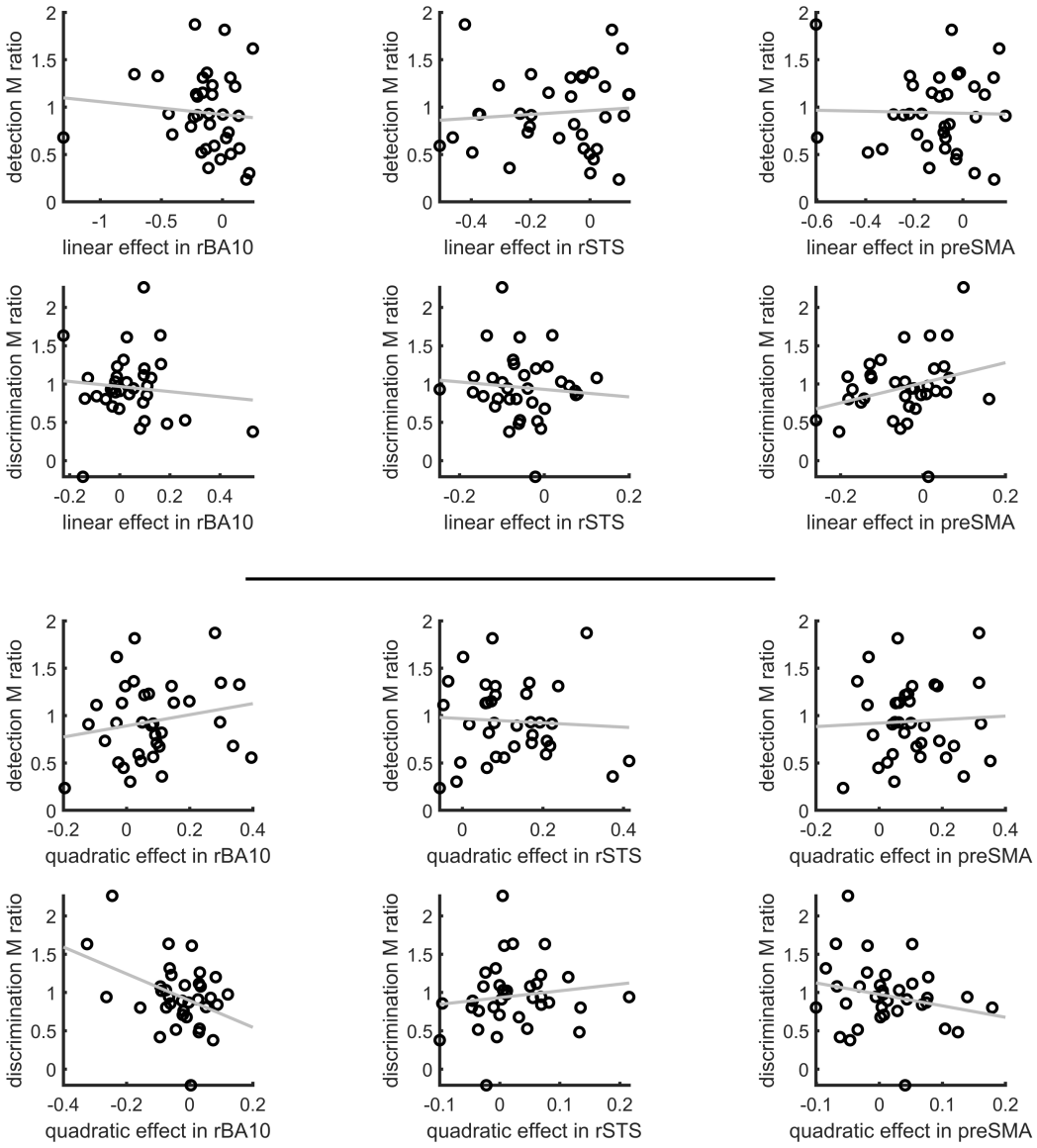


Figure C.6: Inter-subject correlation between the linear (upper panel) and quadratic (lower panel) effects in the right hemisphere clusters and metacognitive efficiency scores (measured as M ratio = meta-d'/d', Maniscalco and Lau, 2012).

C.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).

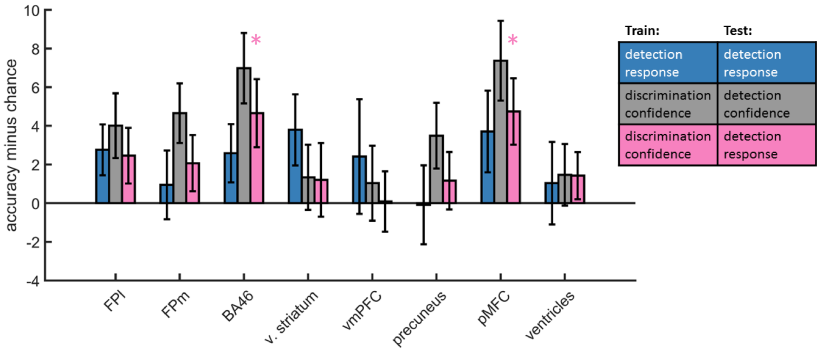


Figure C.7: Accuracy minus chance for classification of response in detection (yes vs. no; blue), and from a cross-classification between tasks: confidence in detection and confidence in discrimination (gray), and confidence in discrimination and decision in detection (pink).

C.8 Static Signal Detection Theory

C.8.1 Discrimination

Generative model

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

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

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

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.

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

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.

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

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

C.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, p. 48).

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

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

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

Inference

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

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

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

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

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

C.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.

C.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.

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

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

C.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.

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

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.

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

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

C.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).

C.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.

$$p(attended_t) = 0.5 \tag{C.16}$$

The attentional state determines the means of sensory distributions.

$$\mu_t = \begin{cases} 0.5, & \text{if cw and } \neg attended_t. \\ -0.5, & \text{if acw and } \neg attended_t. \\ 2, & \text{if cw and } attended_t. \\ -2, & \text{if acw and } attended_t. \end{cases} \tag{C.17}$$

$$x_t \sim \mathcal{N}(\mu_t, 1) \tag{C.18}$$

However, they don't have direct access to their attentional state, but only to a noisy approximation of the probability that they were attending.

$$onTask_t \sim \begin{cases} Beta(2,1), & \text{if } attended_t. \\ Beta(1,2), & \text{if } \neg attended_t. \end{cases} \tag{C.19}$$

Inference

Inference

Participants are then assumed to use their knowledge about the $onTask$ variable when making a decision and confidence estimate.

$$\begin{aligned} p(x_t|\text{cw}) &= p(attended_t|onTask_t)\phi(x_t, 2, 1) + p(\neg attended_t|onTask_t)\phi(x_t, 0.5, 1) \\ &= onTask_t\phi(x_t, 2, 1) + (1 - onTask_t)\phi(x_t, 0.5, 1) \end{aligned} \tag{C.20}$$

$$\begin{aligned} p(x_t|\text{acw}) &= p(attended_T|onTask_t)\phi(x_t, -2, 1) + p(\neg attended_t|onTask_t)\phi(x_t, -0.5, 1) \\ &= onTask_t\phi(x_t, -2, 1) + (1 - onTask_t)\phi(x_t, -0.5, 1) \end{aligned} \tag{C.21}$$

$$LLR = log(p(x_t|\text{w})) - log(p(x_t|\text{acw}) \tag{C.22}$$

$$decision_t = \begin{cases} \text{cw}, & \text{if } LLR > 0. \\ \text{acw}, & \text{else.} \end{cases} \tag{C.23}$$

$$confidence_t \propto |LLR| \tag{C.24}$$

C.10.2 Detection

Generative model

In detection, attentional states only affect the signal distribution, as noise is always centred at 0.

$$\mu_t = \begin{cases} 0, & \text{if a and } \neg attended_t. \\ 0.5, & \text{if p and } \neg attended_t. \\ 0, & \text{if a and } attended_t. \\ 2, & \text{if p and } attended_t. \end{cases} \tag{C.25}$$

$$x_t \sim \mathcal{N}(\mu_t, 1) \tag{C.26}$$

Inference

$$p(x_t|\text{p}) = p(attended_t|onTask_t)\phi(x_t, 2, 1) + p(\neg attended_t|onTask_t)\phi(x_t, 0.5, 1)$$
$$= onTask_t\phi(x_t, 2, 1) + (1 - onTask_t)\phi(x_t, 0.5, 1) \tag{C.27}$$

The likelihood of observing x_t if no stimulus was presented is independent of the attention state.

$$p(x_t|\text{a}) = p(attended_t|onTask_t)\phi(x_t, 0, 1)) + p(\neg attended_t|onTask_t)\phi(x_t, 0, 1)$$
$$= \phi(x_t, 0, 1) \tag{C.28}$$

$$LLR = log(p(x_t|\text{p})) - log(p(x_t|\text{a}) \tag{C.29}$$

$$decision_t = \begin{cases} \text{p}, & \text{if } LLR > 0. \\ \text{a}, & \text{else.} \end{cases} \tag{C.30}$$

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.

$$confidence_t \propto |LLR| \tag{C.31}$$

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, *2*(2), 284–299.

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.

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.

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.

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.

Champely, S. (2020). *Pwr: Basic functions for power analysis*. Retrieved from https://CRAN.R-project.org/package=pwr

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.

De Cornulier, B. (1988). Knowing whether, knowing who, and epistemic closure. *Questions and Questioning*, 182–192.

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.

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.

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.

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.

D'Zmura, M. (1991). Color in visual search. *Vision Research*, *31*(6), 951–966.

Ekstrøm, C. T. (2019). *MESS: Miscellaneous esoteric statistical scripts*. Retrieved from https://CRAN.R-project.org/package=MESS

Ellison, A., & Walsh, V. (1998). Perceptual learning in visual search: Some evidence of specificities. *Vision Research*, *38*(3), 333–345.

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.

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.

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual Review of Anthropology*, *40*, 379–398.

Geng, J. J., & Vossel, S. (2013). Re-evaluating the role of tpj in attentional control:

Contextual updating? *Neuroscience & Biobehavioral Reviews*, *37*(10), 2608–2620.

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. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8–20.

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.

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.

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.

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2019). Lab. Js: A free, open, online study builder.

Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2015). Global statistics are not neglected. *Journal of Vision*, *15*(4), 7–7.

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).

Kanai, R., Walsh, V., & Tseng, C.-h. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*(4), 1045–1057.

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.

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.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340.

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).

Lange, K., Kühn, 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).

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159.

Lee, S. M., & McCarthy, G. (2016). Functional heterogeneity and convergence in the right temporoparietal junction. *Cerebral Cortex*, *26*(3), 1108–1116.

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*. T. Tegg; Son.

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., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance sdt model. In *The cognitive neuroscience of metacognition* (pp. 25–66). Springer.

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. (n.d.). Inference about absence as a window into the mental self-model.

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 1 registered report: Metacognitive asymmetries in visual perception.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–131.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., Lange, F. P. de, & 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.

Moorselaar, D. van, Lampers, E., Cordesius, E., & Slagter, H. A. (2020). Neural mechanisms underlying expectation-dependent inhibition of distracting information. *Elife*, *9*, e61048.

Moorselaar, D. van, & Slagter, H. A. (2019). Learning what is irrelevant or relevant: Expectations facilitate distractor inhibition and target facilitation through distinct neural mechanisms. *Journal of Neuroscience*, *39*(35), 6953–6967.

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*, 2360–17.

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, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

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.

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., & 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*.

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.

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.

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/

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.

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.

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.

Sainsbury, R. (1971). The "feature positive effect" and simultaneous discrimination learning. *Journal of Experimental Child Psychology*, *11*(3), 347–356.

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.

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.

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.

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.

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.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401.

Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, *255*(5), 114B–125.

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.

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.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Ushey, K., Allaire, J., & Tang, Y. (2020). *Reticulate: Interface to 'python'*. Retrieved from https://CRAN.R-project.org/package=reticulate

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception & Psychophysics*, *56*(5), 495–500.

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. (n.d.). *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

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. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*.

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., & 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.

Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4582–4591).

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.

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.