

# Suboptimality and Efficiency in human decision-making



Vickie Chui Lam, Li  
Wolfson College  
University of Oxford

A thesis submitted for the degree of  
*Doctor of Philosophy*  
Trinity 2019

# Acknowledgement

First, I would like to thank Croucher Foundation for providing me with the financial support. Without Croucher Foundation's support, I would not have been able to afford this intellectual pursuit and may not have the opportunity to have any impacts to the scientific world.

I would like to thank my collaborators: Professor Joshua Solomon, Santiago Herce Castanon, Lizzie Michael, Hildward Vandormael and Jan Balaguer for their valuable contribution to some of the work reported in this thesis. I would like to thank all the academics who gave me suggestions on my academic work, and gave me opportunities and encouragements because of my work.

I would like to express my gratitude towards my supervisor Chris Summerfield. If I had never met Chris, I would not have been in this lab and been able to do all the things I now know how to do. You took the risk to take in someone like me, who knew very little about experimental psychology, to join your lab. During my PhD, you have held more optimism on my projects and more confidence in me than myself. Thank you very much for investing your time and effort in me.

To the past and present members of the Summerfield lab, thank you for providing me with a scientifically exciting environment. All of you have been incredibly kind to me and offered help and advice when I needed it. I would like to give special thanks to these people. Thank you, Santiago Herce Castanon and Jan Balaguer, you taught me to code and to think creatively about problems. You were also patient listeners of my problems when I needed help. Thank you Hamed Nili, who have made the lab a joyful place, and I will never forget our in-depth discussions on scripts and pca. Thank you, my friends, especially Mavis Wong, Kewe Hui, Joanne Lai, Felix Lai, Jacqueline Ding, you all have been loyal and loving to me.

I would like to thank the Vandormael family for treating me as their family member. Thank you, Hil, for telling me that I am a great human being when I doubted myself, for telling

me that I am strong and independent when I feel that I could not get past the obstacles, for sharing your shoulder when I feel homesick, and for many more that I cannot write all of them down.

Lastly, I would like to thank my family and dedicate this thesis to them. When I chose to do a doctoral degree rather than taking up the responsibility for the family, my family supported me unconditionally because they understand this is what I wanted. When I felt sad, they comforted me from afar. When I felt lonely, they reassured me that I am not alone. I feel nothing more than proud and grateful to be the daughter of my parents and the sister of Ricky, who have provided me 100% of what they can offer me.

## Contribution

The majority of work described in Chapter 3 was published in PLoS Computational Biology: Li, V., Castañón, S. H., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. PLoS computational biology, 13(8), e1005723. I, Santiago Herce Castañón., Hildward Vandormael and Chris Summerfield contributed analyses ideas. Chris Summerfield and I wrote the original draft. I collected and analysed the data. J.A.S. and Chris Summerfield oversaw the development of the project.

The work described in Chapter 4 (and part of chapter 5) was published in PNAS: Li, V., Michael, E., Balaguer, J., Castañón, S. H., & Summerfield, C. (2018). Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making. *Proceedings of the National Academy of Sciences*, 115(38), E8825-E8834. I, Elizabeth Michael and Chris Summerfield designed the experiment. Jan Balaguer, Santiago Herce Castañón, Hildward Vandormael, Hannah Tickle and Keno Juechems performed data collection of the fMRI data. Jan Balaguer contributed fMRI analysis tools. I collected all of the behavioural data. Chris Summerfield and I analysed the data and wrote the article.

# Abstract

A good decision in isolation may be a bad decision in other conditions. Existing normative theories have been used to benchmark human performance. This is a compelling feature as they provide an understanding of how our decision unfolds in an orderly and rational way. These theories often assumed that decisions are made independent of contexts. However, decisions we have to make in the real-world depends on the context in which they occur. In the past few decades, psychologists and economists notably reveal that human decisions deviate from normative principles when the decisions are made in a context, and protest against one another about whether humans are suboptimal under different considerations of suboptimality. The goal of this thesis is to take these “suboptimal” decisions as a tool to understand what contributes to these suboptimalities. In this thesis, I refine these normative models to show that under reasonable assumptions, biases by decision context can be optimal; and (ii) describe a model embodying an efficiency principle that captures these biases by context across an unusually wide range of settings. In particular, I will focus on how a system that maximises computational efficiency with respect to behavioural needs and environmental statistics can explain: *i)* why do humans discard useful information during summary judgements, *ii)* take into account irrelevant information in perceptual judgement tasks, cognitive control, and value-based judgement tasks, and lastly *iii)* which efficient algorithm can approximate human performance in a trinary choice task.

# Table of Content

|                                                                                                                                       |            |
|---------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>Chapter 1 - Optimality and rationality .....</b>                                                                                   | <b>8</b>   |
| <b>1.1. Ideal observer approach.....</b>                                                                                              | <b>8</b>   |
| 1.1.1. Signal-detection theory .....                                                                                                  | 10         |
| 1.1.2. Bayesian decision theory .....                                                                                                 | 12         |
| <b>1.2. Rational economic decisions .....</b>                                                                                         | <b>19</b>  |
| 1.2.1. Luce's choice axiom .....                                                                                                      | 20         |
| 1.2.2. Von Neumann-Morgenstern's Expected Utility Theory .....                                                                        | 22         |
| 1.2.3. The uprising of bounded rationality .....                                                                                      | 24         |
| <b>Chapter 2 - What is efficiency? .....</b>                                                                                          | <b>28</b>  |
| <b>2.1. Different types of efficiency .....</b>                                                                                       | <b>28</b>  |
| 2.1.1. Information efficiency .....                                                                                                   | 28         |
| 2.1.2. Representational efficiency .....                                                                                              | 31         |
| 2.1.3. Computational efficiency.....                                                                                                  | 36         |
| <b>2.2. Efficiency-inspired models.....</b>                                                                                           | <b>40</b>  |
| 2.2.1. Efficient (en)coding and Bayesian decoding.....                                                                                | 41         |
| 2.2.2. Representational efficiency in visual working memory.....                                                                      | 50         |
| 2.2.3. Adaptive Gain control.....                                                                                                     | 57         |
| <b>Chapter 3 - Robust averaging protects decisions from noise in neural computations .....</b>                                        | <b>61</b>  |
| <b>3.1. Introduction .....</b>                                                                                                        | <b>61</b>  |
| <b>3.2. Methods.....</b>                                                                                                              | <b>63</b>  |
| 3.2.1. Design and behavioural analysis.....                                                                                           | 63         |
| 3.2.2. Modelling.....                                                                                                                 | 66         |
| <b>3.3. Results.....</b>                                                                                                              | <b>75</b>  |
| 3.3.1. Human behaviour .....                                                                                                          | 75         |
| 3.3.2. Model fitting.....                                                                                                             | 78         |
| 3.3.3. Linking decision policy to performance.....                                                                                    | 88         |
| <b>3.4. Discussion.....</b>                                                                                                           | <b>89</b>  |
| <b>Chapter 4 - Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making .....</b> | <b>93</b>  |
| <b>4.1. Introduction .....</b>                                                                                                        | <b>93</b>  |
| <b>4.2. Methods.....</b>                                                                                                              | <b>98</b>  |
| 4.2.1. Subjects.....                                                                                                                  | 98         |
| 4.2.2. Design and behavioural analysis.....                                                                                           | 98         |
| 4.2.3. Adaptive gain model.....                                                                                                       | 101        |
| 4.2.4. Conflict models.....                                                                                                           | 103        |
| 4.2.5. Simulations of perceptual and economic decisions. ....                                                                         | 104        |
| 4.2.6. fMRI data analyses.....                                                                                                        | 106        |
| 4.2.7. Model Comparisons.....                                                                                                         | 108        |
| <b>4.3. Results.....</b>                                                                                                              | <b>108</b> |

|                    |                                                                                                                    |            |
|--------------------|--------------------------------------------------------------------------------------------------------------------|------------|
| 4.3.1.             | Adaptive gain simulations.....                                                                                     | 108        |
| 4.3.2.             | The effect of distractor variance. ....                                                                            | 111        |
| 4.3.1.             | Repulsion effects in a conflict task. ....                                                                         | 115        |
| 4.3.2.             | Functional brain imaging. ....                                                                                     | 118        |
| <b>4.4.</b>        | <b>Discussion.....</b>                                                                                             | <b>125</b> |
| <b>Chapter 5 -</b> | <b>Compare and contrast biologically-plausible explanations on multiattribute<br/>multialternative choice.....</b> | <b>129</b> |
| <b>5.1.</b>        | <b>An overview of multialternative decisions .....</b>                                                             | <b>129</b> |
| 5.1.1.             | Decoy effects .....                                                                                                | 130        |
| 5.1.2.             | Individual differences on the impact of context.....                                                               | 131        |
| <b>5.2.</b>        | <b>An overview of different classes of models in explaining context-dependent decisions .....</b>                  | <b>134</b> |
| 5.2.1.             | Explaining context-dependent decisions using diffusion processes.....                                              | 134        |
| 5.2.2.             | Context-dependent decisions as a consequence of a noisy decision process.....                                      | 135        |
| 5.2.3.             | Efficiency-principled models in value-based decisions.....                                                         | 137        |
| <b>5.3.</b>        | <b>A qualitative comparison of the predictions of efficiency-based models .....</b>                                | <b>147</b> |
| 5.3.1.             | Parametric influence of a two-dimensional decoy. ....                                                              | 147        |
| <b>5.4.</b>        | <b>A quantitative comparison of efficiency-based models .....</b>                                                  | <b>154</b> |
| 5.4.1.             | Parametric multiattribute multialternative experiment .....                                                        | 154        |
| 5.4.2.             | Results .....                                                                                                      | 160        |
| 5.4.3.             | Discussion. ....                                                                                                   | 177        |
| <b>Chapter 6 -</b> | <b>General Discussion.....</b>                                                                                     | <b>182</b> |
| <b>6.1.</b>        | <b>Summary.....</b>                                                                                                | <b>182</b> |
| <b>6.2.</b>        | <b>General Discussion and Future Direction .....</b>                                                               | <b>183</b> |
| <b>Reference</b>   | <b>.....</b>                                                                                                       | <b>188</b> |
| <b>Appendix</b>    | <b>.....</b>                                                                                                       | <b>197</b> |

## Chapter I - Optimality and rationality

---

The goal of this Chapter is to provide a general overview on the normative theories the literature has used for explaining perceptual decisions and economics decisions.

---

### I.1. Ideal observer approach

---

*This section first describes what is an ideal observer approach, then it discusses the contribution of ideal observer approach on visual science. I will take “tilt illusion” as an example to illustrate how the ideal observer approach results in a scientific progress for understanding human vision.*

---

Deriving optimal principles for empirical phenomena is a common practice across the behavioural and life sciences (Schoemaker, 1991). For example, in macroeconomics, it is assumed that individuals, households and companies behave rationally, and thus that free markets will converge to equilibria where the aggregate supply meets the aggregate demand for a product or service. To take an example from biology, optimal foraging theory describes how animals should choose among patches of food to maximise their calorific intake per unit time (Charnov & others, 1976).

In psychology and neuroscience an optimality principle known as the *ideal observer framework* has shaped our understanding of how humans and other animals make decisions. An ideal observer is defined as “a hypothetical device that performs optimally in a perceptual task given the available information” (Geisler, 2011, p. 771) or “a model that includes a full generative model of an observer’s internal responses, mathematically specifies all of the components of the decision-making process, and postulates a locally optimal decision rule.” (Rahnev & Denison, 2018, p. 47). When specifying an ideal observer model, the experimenter typically assumes that the observer has access to an internal model that fully specifies the (stimulus generating) task structure, and that it takes decisions that minimise a cost function, typically by seeking to maximise accuracy or reward (Geisler, 2011). In other words, the ideal observer creates a normative benchmark

model to compare to the descriptive performance of a human. When human performance approaches the ideal performance, we can conclude that the observer is sensitive to the relevant task variables, makes correct assumptions about the generative process, and wishes to maximise positive outcomes. When human performance deviates from the ideal performance, it provides an opportunity to relax the optimality assumption in the ideal observer model, and to devise new hypotheses about the agent's behavior.

For illustration, consider the task of judging the orientation of a central grating that is flanked by a ring of distracter gratings. In this case we might consider the ideal observer to be an artificial image processor with multiple units, each of which faithfully and independently encodes the orientation of a single grating. This processor is coupled to a "decoder" that reads out the orientation of these units, combining them with a specific pattern of weights. Here, because the task is to judge the orientation of the central grating, the decoder should give a weight of 1 to the unit processing the central grating and a weight of zero elsewhere. The outputs of this system will match perfectly with the true or physical orientation of the stimulus.

However, the idealized model above is not a good description of human performance. In the tilt illusion, discovered by Gibson (1937), when the tilts of central and flanking distracters are similar, the perceived orientation of a central grating is biased away from ("repulsed by") that of the flanking gratings. Blakemore, Carpenter, and Georgeson (1970) proposed that a repulsive tilt illusion effect arises from lateral inhibition between neighbouring neurons in the primary visual cortex. In terms of our model, thus, the processing units for each individual grating are not independent but rather share inhibitory connections. This means that the output of the unit coding for the central units is biased by that coding for the surrounding units, leading to deviations in the reported orientation.

The tilt illusion also illustrates how the biases that occur in human perception can be complex. When the orientation of the central and surround gratings are highly discrepant, the orientation central grating is perceived as more similar to (not more

different to) the surround orientation (an attractive, rather than repulsive effect) (O'Toole & Wenderoth, 1977). A simple inhibition account fails to explain the attractive tilt illusion effect. Subsequently, different mechanisms have been hypothesized to explain the tilt illusion effect (e.g., Schwartz, Sejnowski, & Dayan, 2009; Solomon & Morgan, 2006; Wei & Stocker, 2015). For example, Solomon and Morgan (2006) proposed that there is disinhibition in addition to lateral inhibition in the neural circuit, neurons shift their orientation preference towards the surround orientations etc. The tilt illusion effect illustrates nicely how a discrepancy between the human observer and the ideal observer (i.e. the independent processor example) results in an adjustment the ideal observer framework, which in turns lead to development of new hypotheses to understand the cause of suboptimal decisions or biased judgments.

#### 1.1.1. Signal-detection theory

---

*This section describes the components of a commonly applied ideal-observer model – Signal detection theory, and how these components contribute to perceptual decisions.*

---

One of the earliest ideal observer models is known as Signal Detection Theory (SDT) (Green & Swets, 1966). SDT begins with the intuition that decisions are irreducibly corrupted by noise. In perceptual decision making literature, two main sources of noise (or uncertainty) have been identified. One is external noise, where variability comes from the environment itself (and may be controlled by the experimenter) such as the contrast of a grating or Gabor patch. The other is internal noise, where the stochasticity within the brain, such as the intrinsic variability in neural responses, leads to variable choices even to the same stimulus. These sources of noise entail that even an optimal participant can make different choices on exactly the same trial.

According to SDT, performance on a perceptual categorisation task will be influenced by the relative levels of signal and noise available in the stimulus (signal-to-noise ratio). Consider a forced choice detection task where participants have to decide whether

the imperative stimulus contains a low-contrast Gabor or not. On each trial, the observer receives incoming evidence  $e$ . The variable  $e$  might be thought of as the (noisy) neural firing evoked by the stimulus in a pool of neurons. Over repeated trials, the observer forms an internal model of two probability distributions:  $P(e|S+N)$  and  $P(e|N)$  i.e. the likelihood that the firing pattern  $e$  comes from both stimulus + noise or just noise alone. The observer can use these distributions to infer the most likely cause of  $e$  by comparing the two likelihoods. To make a decision, the decision maker simply needs to compute the likelihood ratio (LR) of these two quantities  $LR = \frac{P(e|S+N)}{P(e|N)}$  and compare it against a response criterion  $c$ , such that she would respond that a Gabor is present only when  $LR \geq c$ . Figure 1.1 shows four possible response classifications under the SDT framework: correctly detect the presence of a Gabor (hit), fail to detect a Gabor (miss), correctly detect the absence of a Gabor (correct rejection) or falsely detect a Gabor when none is present (false alarm). The proportion of trials that falls into these four different response types depend on two factors: (i) deviation of the likelihood ratio from zero, which depends on the overlap between the  $S+N$  and  $N$  distributions (i.e. the observer's sensory sensitivity:  $d'$ ), and (ii) the decision rule or response criterion the participant adopts. To set the response criterion in a way that maximises reward, an observer needs to consider the relative fraction of  $S+N$  and  $N$  trials in the experiment, as well as the cost function, which is given by the payoff matrix for each trial type.

To describe the participants' performance, we can compute a statistic known as  $d'$ , which is derived from the probability of making a hit vs. a false alarm. A different description of performance is given by the ROC curve (Receiver Operating Characteristic curve). A ROC curve shows the false alarm rate versus the hit rate under different response criteria (Tanner Jr & Swets, 1954). In other words, a ROC curve provides information regarding the observer's sensitivity. When the signal+noise distribution is indistinguishable from the noise only distribution (i.e., 100 % overlap, i.e.  $d' = 0$ ), the hit rate should always match with the false alarm rate regardless of where the response criterion is placed, resulting in a ROC curve that lies on the identity line. However, as the participant's

sensitivity becomes higher i.e., the two distributions are far apart, the probability that a subject would make a hit is higher than the probability of making a false alarm, resulting in a curvilinear ROC curve that lies above the identity line.

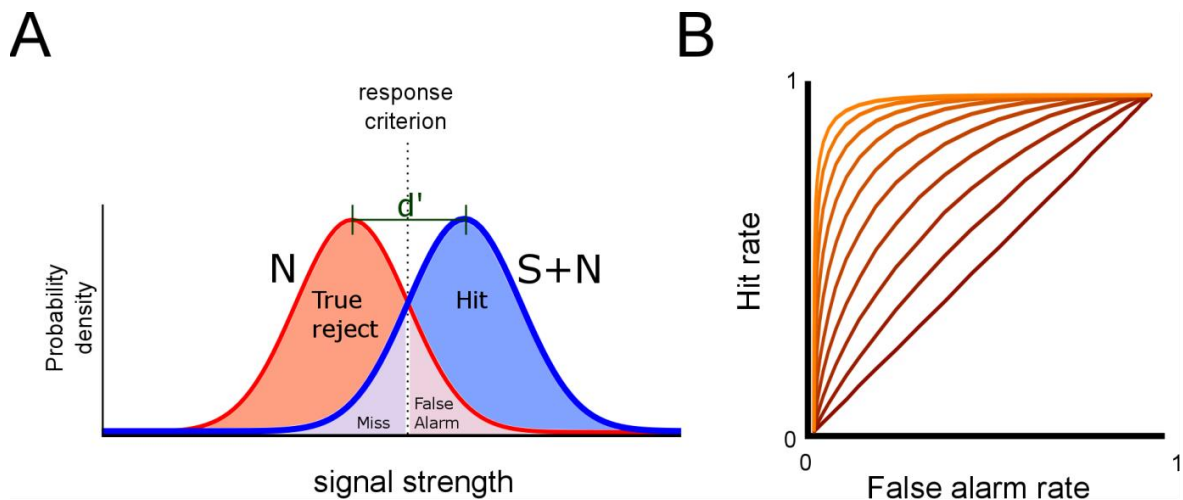


**Figure 1.1. Hypothetical probability distributions and ROC curves**

(A) The red and blue distributions represent the internal signal associated with the trials where a noisy stimulus was present (S+N) and the trials where the stimulus only contains noise (N). In SDT, performance on the task depends on how overlapping the two distributions are, and how the occurrence of different types of error depends on the response criterion placed by the participant. The participant's sensitivity ( $d'$ ) is measured by the difference of the inverse normal probability on the hit rate (proportion of hits on trials where the stimulus is truly present) and the inverse normal probability on the false alarm rate (proportion of false alarms on trials where the stimulus is not present). A higher  $d'$  means that the signal is more readily detected from the noise. On the contrary, a  $d' = 0$  means that the subjects cannot distinguish the signal from noise. (B) A ROC curve represents the hit rate vs the false alarm rate. Elbowed ROC curves (orange lines) that are more deviated from the diagonal ROC curve (brown lines) means that the sensitivity of this subject is higher.

### 1.1.2. Bayesian decision theory

*Going beyond Signal Detection theory, Bayesian decision theory formalizes how the observer's prior has an impact on their decisions. This section first describes what Bayesian decision theory is, then I discuss how do Bayesian decision theory explain*

*perceptual illusions. Lastly, I discuss the drawbacks of using Bayesian decision theory to understand optimality in human decision-making.*

---

Optimal decisions require the observer to incorporate prior knowledge about what is most likely to occur. The idea that perception is jointly determined by our current sensation and our prior expectations (formed by our past experience) can be traced back to Hermann von Helmholtz, who believed that perception is an inference over the causes of sensation (Von Helmholtz, 1867). Von Helmholtz’s theory of “unconscious inference” is an intellectual precursor to contemporary Bayesian decision theory. As discussed in the previous sections, sensory inputs are inherently noisy, resulting in decision uncertainty. Bayesian theories of perception propose that the goal of our sensory system is to infer the hidden state or the true generative distribution that gives rise to these noisy signals (Knill, Kersten, & Yuille, 1996; Yuille & Kersten, 2006). Bayesian decision theory (BDT) is a generalization of SDT that provides the optimal standard for how an ideal observer should respond given some beliefs about the world and the statistical likelihood of an outcome given the action (Mamassian, Landy, & Maloney, 2002; Yuille & Kersten, 2006). BDT can be applied to a variety of decision problems with differing complexity, but let us consider here a simple perceptual problem: estimating the orientation of a tilted grating. According to BDT, the optimal choice can be defined using Bayes’ rule (Figure 1.2):

$$p(\theta|e) = \frac{p(e|\theta)p(\theta)}{p(e)}$$

Where  $p(\theta|e)$  refers to the posterior distribution over possible orientations: the tilt  $\theta$  given the noisy spike rate from a pool of neurons “ $e$ ”. The posterior distribution denotes our belief distribution concerning the true state (orientation in this example) given what we have observed. The posterior probability is comprised of three components: the likelihood function  $p(e|\theta)$  and the prior  $p(\theta)$ . The prior encodes an estimate of the likely probability over different orientations prior to the observation. For example, I might have a prior belief that the grating is tilted clockwise without yet having seen the grating. The likelihood function specifies the probability distribution of our internal signals or

observations given the physical orientation of the grating. Lastly,  $p(e)$  corresponds to the probability distribution of the observation (e.g., the probability of different spike rates), which can be computed by marginalising the likelihoods weighted by the prior.

The general form of BDT here allows the observer to estimate the most likely orientation of the grating. This estimate is derived from the posterior density function over all possible orientations. To derive this estimate from the posterior, it is necessary to specify a cost function, which indicates the penalty incurred when the estimate deviates from the ground truth. In the case of the current example, one possibility is to define the loss  $\mathcal{L}(\hat{\vartheta}, \theta)$  weighted by our posterior belief about the orientation, marginalised over all possible orientations :

$$\mathbb{E}[\mathcal{L}(\hat{\vartheta}, \theta)] = \int_{-\infty}^{\infty} \mathcal{L}(\hat{\vartheta}, \theta) p(\theta|e) d\theta$$

The best estimate  $\vartheta^*$  is simply the one that minimises the expected loss:

$$\vartheta^* = \underset{\hat{\vartheta}}{\operatorname{argmin}} \mathbb{E}[\mathcal{L}(\hat{\vartheta}, \theta)]$$

The actual specification of the loss function (or the estimator) can be in other forms and is subjective to the decision problem of interests (e.g., mutual information in efficient coding hypothesis section or  $L_p$  norm in the Bayesian Observer Model section). Take a simple example of a binary loss function: there is a loss = 1 (0 otherwise) whenever the orientation estimate does not match with the true orientation:

$$\mathcal{L}(\hat{\vartheta}, \theta) = \begin{cases} 0 & ; \quad \hat{\vartheta} = \theta \\ 1 & ; \quad \text{otherwise} \end{cases}$$

It turns out that using a binary loss function is equivalent to taking the orientation estimate that lies on the mode of the posterior distribution (this is also called an orientation estimate that is generated by a maximum a posteriori estimator).

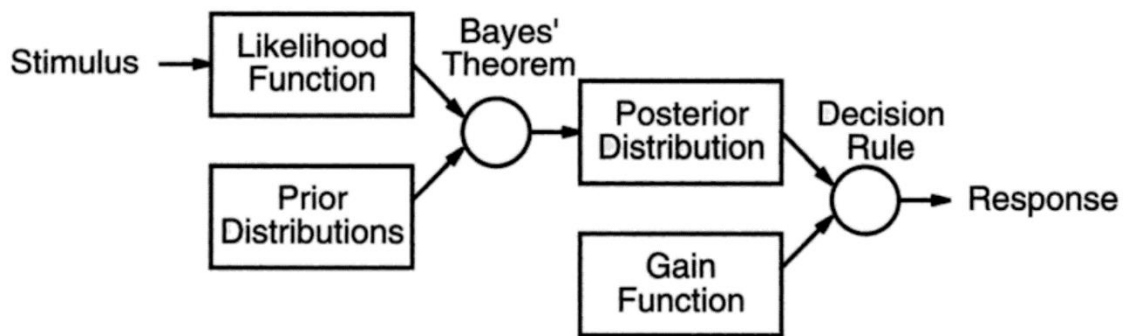


**Figure 1.2. Steps for Bayesian modelling**

Figure from (Mamassian et al., 2002). In this example, a gain function is used. Equivalent to the loss function specified in the main text, the gain function (negative loss) serves as a multiplicative scaling factor to the posterior distribution.

**Understanding perceptual illusions through BDT.** Bayesian decision theory offers a broad theory of cognition: perception, decisions and motor control can be understood as inference under uncertainty. There is ample evidence demonstrating that human behaviour matches the ideal performance as described by the BDT. For example, Körding & Wolpert (2004) tested whether humans combine sensorimotor information optimally according to Bayes' rule. Participants performed a motor task that involved pointing to a visual target with a cursor in a virtual reality setup. When the participant moves the cursor with their finger, the cursor disappears and displaces laterally from the true finger location. The extent of the lateral shift was randomly drawn from a prior distribution that is normal in shape with a mean of 1 cm to the right of veridical. When the participant is halfway through the target, visual feedback (i.e., likelihoods) with various degrees of variabilities (no, medium, large, full uncertainty) is provided to give participants a noisy estimate of their true finger position. If participants are Bayes optimal and can represent the level of sensory uncertainty accurately, they should rely heavily on the prior when the sensory uncertainty increases (e.g., 0 variabilities visual feedback condition), i.e., compensate 1 cm caused by the displacement from the prior, and should compensate to an increasing extent for the lateral shift as the visual feedback becomes more reliable. This is exactly how

participants behaved, implying that participants have learned the prior and that they are able to incorporate the information in a Bayes optimal way.

This work, and other findings too numerous to summarise, suggest that human behavior approximates the ideal standard described by BDT. However, humans frequently exhibit perceptual illusions, as alluded to above with the example of the tilt illusion, suggesting that the brain's interpretation of the sensory input deviates from the true retinal image (Gregory, 1968). How can we reconcile these two findings?

In BDT, one of the conditions required for forming optimal perceptual decisions is that an observer needs to use an appropriate prior for interpreting sensory signals, i.e. one that matches the generative model that describes real-world statistical regularities (Kersten & Yuille, 2003). Thus, perceptual illusions may be “optimal” if humans are using real-world prior beliefs about stimulus probabilities when performing artificially controlled experiments in the laboratory (Geisler & Kersten, 2002; Mamassian et al., 2002).

For instance, some studies suggest that human perception can be explained by known properties of the environment (e.g., Adams, Graf, & Ernst, 2004; Mamassian & Landy, 1998; Stocker & Simoncelli, 2006; Weiss, Simoncelli, & Adelson, 2002). Take 3D interpretations as an illustration. Humans have the ability to interpret 2D drawings as 3D. For example, the 2D shaded shapes in Figure 1.3 are perceived as convex or concave. This commonly is explained by an overlearned prior that “light comes from above” (Adams et al., 2004; Ramachandran, 1988). Mamassian and Landy (1998) probed how 2D line drawings are interpreted as 3D shapes by asking subjects to report whether a curved line drawing that can occur in different orientations (Figure 1.3B left) has a saddle-shaped (hollow-like; Figure 1.3B upper-right) or egg-shaped (bulged-up; Figure 1.3B lower-right) surface. Human reports can be explained by a Bayesian observer who has priors that match with real-life experience of the world, for example, most surfaces are convex, and surface contours are consistent with those that would be observed when the stimulus is viewed from above. Another well-known perceptual illusion example is that the faces in Figure

1.3C are both perceived as convex even though the shading and shadows of both faces are opposing. In fact, the face on the left is hollow, but since we have a strong prior that faces are convex, both faces result in similar perceptions (Gregory, 1968).

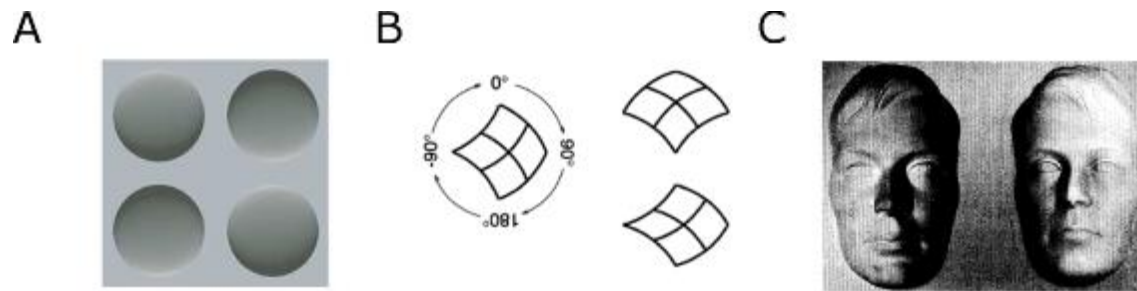


**Figure 1.3. Interpreting 2D drawings as 3D shapes**

(A) Figure copied from (Adams et al., 2004) (B) Figure copied from (Mamassian et al., 2002) (Left panel) line drawings stimulus used in the viewpoint experiment. On each trial, a rotated line drawing with a specific orientation is displayed. Human observers have to report if the line drawing is more (upper-right) egg-shaped or (lower-right) saddle-shaped. (C) Figure copied from (Gregory, 1997)

One drawback of the view described above is that the prior responsible for the bias is proposed post-hoc, rather than being predicted from some measurable property of the world. As an alternative, recent studies have tried to measure an observer's internal prior from image statistics, and asked whether this can explain human perception (Girshick, Landy, & Simoncelli, 2011; Golz & MacLeod, 2002). One prominent example is the oblique effect – where human observers and animals display elevated acuity for tilts at cardinal orientations (vertical and horizontal orientations) as opposed to oblique orientations (non-cardinal orientations), as well as displaying a perceptual bias where orientations appear as closer to cardinal orientations than oblique orientations (Appelle, 1972; Solomon & Morgan, 2006, 2009; Tomassini, Morgan, & Solomon, 2010). Girshick et al. (2011) asked whether a Bayesian observer who uses image statistics as the prior would recover the same type of perceptual biases and error variability exhibited in humans. They first ran a psychophysical task where participants see two arrays of oriented Gabor patches placed on the left and right hand side of the screen. Participants were asked to report whether

the array on the right-hand side is tilted more clockwise or counterclockwise of the oriented Gabor patches on the left. They recovered prior density functions, under the assumption that the discrimination threshold is influenced by subjects' internal priors. These recovered prior density functions fit well with the previous observation that cardinal orientations are more prevalent. They further analysed the local orientation statistics of natural scenes and used these as the prior of a Bayesian observer for this perceptual discrimination task. They showed that such a Bayesian observer is able to predict perceptual biases and acuity as demonstrated by human observers, providing a solid evidence on the relationship between environmental priors and perception.

**Criticism of the Bayesian approach.** Despite Bayesian theories' success in explaining diverse perceptual effects, one major criticism on the Bayesian framework is that the success of Bayesian theories rests on the potentially arbitrary choice of priors and likelihoods (Bowers & Davis, 2012; Marcus & Davis, 2013). Simply put, a Bayesian model risks being unfalsifiable as one can always choose an ad hoc prior that fits the data well (Jones & Love, 2011). Some suboptimal decisions such as perceptual illusions discussed above, as well as perceptual adaptation effects (where the perception of an imperative stimulus deviates from its true identity by appearing closer to the adapted stimulus) can be framed as an optimal decision once an appropriate prior has been taken into account. Apart from the potentially unjustifiable flexibility, a Bayesian approach that attributed the data wholly by a prior might lack explanatory power since it failed to provide the exact reason why should humans adopt those priors and why would humans deviate from optimality (Summerfield & Li, 2018).

There are a number of potential replies to this critique. For example, the Girshick et al. study illustrates how the prior can be derived empirically, sidestepping the criticism of limitless flexibility in model specification (Geurts et al., 2018; Summerfield & Li, 2018). As a result, proponents of the Bayesian approach believe that Bayesian models are falsifiable (Gershman, 2018; Griffiths, Chater, Norris, & Pouget, 2012). Another reply is that the main function of Bayesian models is not to demonstrate or falsify optimality; rather

they are a functional tool for researchers to specify the prior, likelihoods, loss functions and the decision rules given reasonable assumptions about a particular task. (Mamassian et al., 2002; Rahnev & Denison, 2018).

Despite Bayesian inference being statistically optimal, Bayesian inference is often an inefficient computation for the brain since it requires the brain to consider an unlimited number of possible hypotheses about the world at once (Summerfield & De Lange, 2014). Deriving the posterior distribution is often mathematically intractable since the denominator of the Bayes rule or the normalising constant requires one to integrate the prior-weighted likelihoods from all possible hypotheses, making it impossible to compute when the hypothesis space is huge (Dasgupta, Schulz, & Gershman, 2017). However, one can approximate the posterior distribution using different approximation methods, e.g., sampling-based methods. In the unlikely case where the brain can sample infinite samples to approximate the posterior distribution, optimal Bayesian inference is possible. Recently, Sanborn and Chater (2016) argued that the brain resolves the intractable inference problem using sampling, and behavioural biases such as base rate neglect can be explained by the fact that limited samples are used for approximating (an inaccurate) posterior. Dasgupta and colleagues (2017) showed how an algorithmic formulation of sampling approximations could explain multiple cognitive effects during decision-making (e.g., anchoring, subadditivity, etc). However, it remains unclear how the brain can perform sampling or optimal Bayesian inference at the implementation level.

## **1.2. Rational economic decisions**

---

*The goal of this section is to provide an overview of the equivalent ideal observer approach in economic decisions i.e. –the rational approach. I will first give an introduction on the commonly used choice axioms in the rational approach. Then, I will describe how humans violate these axioms. Lastly, I will describe the bounded rationality approach as an alternative way to approach irrational economic decision-making.*

---

In the field of economics, humans are typically assumed to be ‘rational’ in that they make decisions that maximise their expected utility. The approach here is very similar to that previously discussed for perceptual decision making. Rational models present an idealized standard for exploring economic phenomena, but do not always describe well how individual human make economic choices. Humans can be shown to make choices that are not in their economic interest, i.e. that do not lead to maximization of monetary gain or leave them open to financial exploitation. This rational approach has given rise to much controversy, which is summarized below in some depth.

In the following sections, I will first give an introduction to the commonly used choice axioms in behavioural economics— Luce’s choice axiom and von Neumann-Morgenstern axioms in the rational approach. Finally, I will describe how humans violate these axioms and describe the bounded rationality framework as an alternative way to approach human decision-making.

### 1.2.1. Luce’s choice axiom

---

*The Luce’s choice axiom proposes that the presence of alternative options in a choice set should not change the relative preference over two existing options.*

---

A choice axiom formulated by Luce (1959) described that rational choice behaviour should follow a principle where the probability of choosing an option should be the joint probability of choosing the option from a set and the probability of choosing that set from a larger set:

$$p_T(R) = p_S(R)p_T(S)$$

Where  $R$  is an option within the set  $S$ , and the set  $S$  is within a bigger set  $T$ .  $p_T(R)$  refers to the probability of choosing  $R$  from the set  $T$ . For instance, the probability that I will order dumplings for lunch is equal to the joint probability that I will choose the dumplings among other food items from the menu of restaurant A and the probability that I go to restaurant A. This choice axiom is in fact fundamental to many computational

models that involve action selection. Take the simplest case where the set  $S$  only contains two options:  $R$  and  $P$ , the probability of choosing  $R$  from set  $S$  would be the ratio between the value of option  $R$  and the value of the set  $S$  (i.e., the sum of value for option  $R$  and option  $P$ ).

$$\frac{p(R, P)}{p(P, R)} = \frac{p_S(R)}{p_S(P)} = \frac{v(R)/v(S)}{v(P)/v(S)}$$

Where  $p(R, P)$  refers to the probability of choosing  $R$  over  $P$ ,  $v(R)$  and  $v(s)$  refer to the value of option  $R$  and set  $S$  respectively. As a consequence, other alternatives in the set  $S$  should not impact the relative probability of choosing  $R$  over  $P$ , a.k.a. The “constant ratio rule” (Luce, 1959; Pleskac, 2012). This also implies the following features in rational choice behaviour: (i) the probability of choosing an option is proportional to its value or response strength for that option, (ii) the pairwise probability between options within a set should be multiplied to 1. This will be more clear when we consider the case that there are three options  $\{O, P, R\}$  in set  $S$ . The product of the pairwise probabilities between options (i.e. a probability which indicates how valued one option is over another option) is multiplied to one (‘product rule’):

$$\frac{v(O)/v(S)}{v(P)/v(S)} \cdot \frac{v(P)/v(S)}{v(R)/v(S)} \cdot \frac{v(R)/v(S)}{v(O)/v(S)} = 1$$

$$P(O, P) \cdot P(P, R) \cdot P(R, O) = 1$$

Lastly, (iii) alternative options should not change the *relative* choice probability of two options (i.e., Independence from Irrelevant Alternatives ‘IIA’). The commonly used choice function – the Softmax rule from Sutton and Barto (2011) is an instantiation from this choice axiom where the choice probability of an option is the ratio of the exponentially transformed value for that alternative and the sum of the exponentially transformed value for all alternatives:

$$p(R) = \frac{e^{v(R)}}{\sum e^{v(S)}}$$

Where  $v(R)$  refers to a scalar that scales proportionally with the value of option  $R$ . In conclusion, the assumption that choices should be independent of irrelevant contextual factors is a premise for many theoretical frameworks.

### 1.2.2. Von Neumann-Morgenstern's Expected Utility Theory

---

*The theory describes a few features of rational choice behaviour: completeness, transitivity, continuity and independence and it is subjected to diminishing marginal utilities on subjective values.*

---

Imagine two people who won the first and the second prize of a lucky draw at a café respectively. The first winner of the lucky draw receives as much free coffee as they want for one year. The second winner gets to drink 4 cups of coffee for free every day for one year. Does this mean that the first winner should be infinitely happier with the prize (i.e., happier than the second winner)? Intuition suggests that the answer is likely to be no, because the first winner is unlikely to drink more than four coffees a day and thus having unlimited coffee provides no extra benefit. Bernoulli was the first person who proposed the concept of “expected utility” – an intangible quantity that describes the average degree of satisfaction one could get from consuming goods (Bernoulli, 2011). In particular, it is stated that the psychological “utility” of some monetary value does not increase proportionally with its objective value. The example of the lucky draw illustrates the law of diminishing marginal returns – the extra utility one could obtain from having an extra consumption of the good increases at a decreasing rate – implying a concave utility function. The invention of utility allowed economists to quantify subjective preferences, and thus to derive a set of rules so that an agent is maximising the expected value of a decision.

Expected utility theory as formalised by von Neumann & Morgenstern (1944) suggests that a rational agent should follow four choice axioms. They first assume that any set of utilities could be represented by some numerical values allowing an ordinal relationship between utilities. For describing these four choice axioms, assume there is a

choice set with two utilities  $m$  and  $n$ , where  $m > n$ . The two utilities can be represented by a utility function  $v$  that is assigned a numerical value so that  $v(m) > v(n)$ .

The first axiom is that of *completeness*, which states that any utility function must hold either one of the relations:

$$m > n, \quad m < n, \quad m = n$$

The second axiom is that of *transitivity*. A choice set with three utilities  $l, m, n$  must follow this transitivity rule: if  $l > m, m > n$ , then it must be the case that  $l > n$ . The third axiom is that of *continuity*. Given that  $l < m < n$ , there exists a probability ( $0 < p < 1$ ) such that  $p \cdot l + (1 - p) \cdot n < m$ . This means that given  $l$  is preferred over  $m$  and  $n$ , regardless of how valued  $n$  is, if the probability of receiving  $n$  is very small, it will not change the preference for  $l$  over  $m$ . The last axiom is that of *independence* (i.e., also IIA). Given the three utilities  $l, m, n$  and  $0 < p < 1$ ,

$$p \cdot l + (1 - p) \cdot n > p \cdot m + (1 - p) \cdot n \Leftrightarrow l > m$$

This axiom states that if  $l$  is preferred over  $m$ , the preference between  $l$  and  $m$  should not change as a function of the utility  $n$  and the probability  $p$ .

Let us say we are deciding which of two lotteries is preferred: lottery A [50% ( $p$ ) chance of winning \$100 ( $l$ ) and 50% ( $1 - p$ ) chance of losing \$100 ( $n$ )] and lottery B [50% ( $p$ ) chance of winning \$50 ( $m$ ) and 50% ( $1 - p$ ) chance of losing \$100 ( $n$ )]. It should be obvious that lottery A is better than lottery B as it has a greater chance of earning more money. However, imagine a second scenario: lottery A [1% chance of winning \$100 and 99% chance of losing \$50] and lottery B [1% chance of winning \$50 and 99% chance of losing \$50]. It remains the case that lottery A is better than lottery B even though the probability ( $p$ ) of winning the lottery is less fair compared to scenario 1 and the losing values ( $n$ ) of both A and B have changed. This example implies that the probability  $p$  and utility  $n$  have no impact on the preference.

### 1.2.3. The uprising of bounded rationality

---

*The observation that humans are risk-averse and loss-averse has motivated the formulation of a new theory – Prospect theory, which incorporates the role of “context” in decision making. Next, I describe the bounded rationality approach, which provides a reason on why humans use “context” for decision-making. I concluded this section with the idea that “irrationality” can be seen as an outcome of a system that maximises its processing efficiencies with limited capacity.*

---

**Prospect Theory.** Prospect theory (Figure 1.4.) suggests a model that can account for risk aversion and loss aversion. In Prospect theory, a gain or loss  $x = \{x_1, x_2, \dots, x_n\}$  with probability  $p = \{p_1, p_2, \dots, p_n\}$ , has value  $V$  given by:

$$V = \sum_{i=0}^n \pi(p_i) v(x_i)$$

Where  $\pi(p)$  and  $v(x)$  refers to the weighted perceived probability of outcome  $x$  and the value function for  $x$  (Kahneman & Tversky, 2013). The origin of the value function in Figure 1.6 is placed at a subjective reference point for the agent.

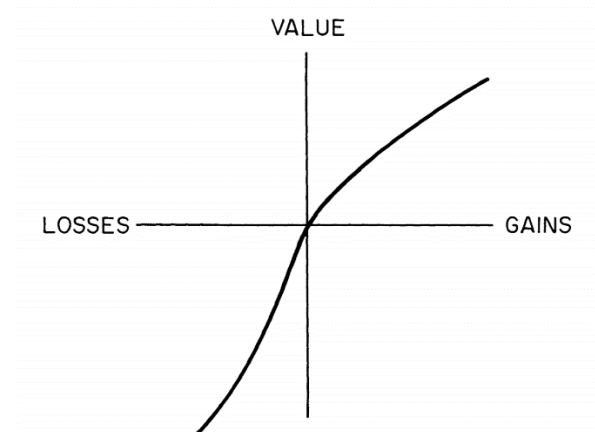


**Figure 1.4. A hypothetical value function from Prospect theory**

Figure copied from (Tversky & Kahneman, 1991). The value function for gains is concave and convex for losses. Importantly, the slope for the function in the gain domain is shallower than the slope of the function in the loss domain for a loss averse agent, implying that the rate of increase in value by 1 unit of gain is smaller than the rate of decrease in value by 1 unit of loss. For someone

to accept a fair chance gamble like the one suggested in the main text, the dealer would need to offer a greater monetary gain for people to “counteract” for people’s loss aversion.

The reference-point is variable can be influenced by multiple factors, such as the current wealth of the individual. To capture the finding that individuals are risk averse in the gain domain and risk seeking in the loss domain, the value function is concave and convex in the two domains respectively. Lastly, the value function of losses is steeper than that of gains for any individuals who are loss averse. A mathematical formulation of the nonlinear mapping of the value function and the probability weighting function is proposed by Tversky and Kahneman (1992). A value function for monetary value  $x$  can be described by:

$$v(x) = x^\alpha \text{ if } R_x \geq 0$$

$$v(x) = -\lambda(-x)^\beta \text{ if } R_x < 0$$

Where  $\alpha, \beta$  are free parameters describing the characteristic pattern of diminishing marginal utility and loss aversion;  $R_x$  corresponds to the reference point. The probability weighting function takes the following form:

$$\pi(p) = \frac{p^\gamma}{(p^\gamma + (1 - p)^\gamma)^{1/\gamma}}$$

This formulation allows one to fit the  $\gamma$  separately to account for different risk averse or risk seeking preferences.

**Bounded rationality.** One plausible hypothesis for why human behaviour deviates from normative principles is that biological processing systems have intrinsic costs and constraints; researchers often fail to include these when formulating normative choice theories. A new theoretical framework known as bounded rationality was proposed to shift the theoretical focus from “making optimal decisions” to “making satisficing decisions” (Simon, 1955, 1956). Unlike the normative rationality approach, bounded rationality proposes that the choices people make are not only determined by the properties of the

environment but also by the knowledge that the decision-makers *do and do not have* about the world, and their *ability and inability* to evoke this knowledge when it is relevant (Simon, 2000). Another way of considering bounded rationality is that it considers decisions as “optimisation under constraints”, where constraints are often defined by the cognitive limitations of the human mind, such as attention and memory capacities (Gigerenzer & Selten, 2002). In the past decade, many theories have built upon “bounded rationality” to explain different economic phenomena – ranging from individual choice behaviour in microeconomics to consumption and saving behaviour in macroeconomics.

For example, Kahneman and Egan (2011) proposed that judgements are often performed by “system 1”, e.g. the system that is responsible for automatic and fast operations at the expense of creating some systematic errors. These errors can be corrected by system 2, e.g. the system that is responsible for slow and deliberative operations. A common strategy of system 1 is that it creates heuristics that are useful for making judgements, even though this will result in some biases. In particular, decisions between lotteries such as those that are described in the previous section involve operations similar to perception (which is assumed to adopt system 1), i.e. they are reference-dependent. We know from human perception that our sensation adapts its sensitivity to the contextual environment. For example, two objects with the same luminance might appear brighter or duller depending on whether we are in a dark or bright room respectively. Intuitively, evaluation of potential decision outcomes from economic prospects can also be reference-dependent – e.g., depending on the initial state of wealth. For example, a millionaire who earns \$10 from trading stock might feel that the stock is doing badly. The same \$10 earning might mean a huge gain for a homeless person. This reference-dependence is the major advance of Prospect theory over earlier models. However, context-dependent judgments are indeed suboptimal according to the principles of axiomatic rationality. Under the axiom of transitivity, i.e., when option A is preferred to option B, and option B is preferred to option C, one should prefer option A over option C. However, imagine a person who violates this axiom and ends up choosing C instead of A.

In other words, this person is open to exploitation, whereby this person is guaranteed to lose from a transaction when he is offered to trade away a better option A for a worse option C (e.g. a “Dutch book”). Then why are people generally risk averse and loss averse?

Stewart and colleagues proposed a theory called Decision by Sampling (DbS). This theory can explain many context-dependent effects including concave utility functions, loss aversion and non-linear weighting of probabilities. It can also explain other effects not emphasized here such as hyperbolic temporal discounting, whereby the value of a reward is discounted if the reward is not received immediately and this discounting is faster for early delay periods compared to longer delay periods. DbS assumes that the subjective valuation of a prospect is constructed online by sampling instances from memory (Noguchi & Stewart, 2018; Stewart et al., 2006). Because our memory is formed by past encounters with stimuli, and thus our memory of the value of goods should on average reflect their true value. Stewart and colleagues approximated the distribution of gains and losses by measuring the frequencies of individual bank credits and item prices from a UK supermarket. These distributions follow the power law where smaller gains and losses are more probable. As a result, our sensitivity to smaller gains and losses should be greater (i.e., steeper utility functions). This is consistent with the observation that not only the perceptual sensitivity for low-level features follows the Weber-Fechner law (there is a greater sensitivity for changes in small stimulus magnitudes than large stimulus magnitudes), the sensitivity for processing numerical values also follows Weber’s law (Dehaene, 2003). Thus, in DbS, loss aversion can be a byproduct of the natural statistics of values in the world. In the next Chapter, I will discuss how the natural statistics of the world is relevant for constructing an efficient neural coding scheme.

## Chapter 2 - What is efficiency?

---

*Three forms of efficiencies have been defined across fields to understand human cognition. The goal of this chapter is to provide an overview on the three types of efficiency: information efficiency, representational efficiency and computational efficiency. I will first provide some background on these efficiencies, then I will examine computational models that have assumed the three types of efficiency to show how visual representations are formed.*

---

### 2.1. Different types of efficiency

#### 2.1.1. Information efficiency

---

*This section describes the efficient coding hypothesis, a theory of neural computation that is grounded in information theory.*

---

**Efficient coding hypothesis.** The notion that neural information processing systems have evolved for maximise coding efficiency can be dated back to the 1950s when cognitive scientists and neuroscientists asked “how is the perceptual system optimised to process the enormous flow of sensory information? (Attneave, 1954; Barlow, 1961). Perception can be characterised as comprising two processing stages whereby sensory information is first encoded and represented by the brain, and then decoded during perceptual inference (Figure 2.1A). A neural system cannot just encode all sensory information perfectly due to metabolic or energy constraints (Laughlin, 2001). As a consequence, Barlow proposed that the limitation in neural resources lead to a sensory feature coding scheme that maximises efficiency during sensory representation (Barlow, 1961). The efficient coding hypothesis Barlow has considered is grounded in information theory, a mathematical theory of the limitation in communications across operation channels. Hence, this thesis considers Barlow’s definition of efficient coding (Barlow, 1961) or, more specifically – the “redundancy reduction hypothesis” as *Information Efficiency*.

Like any information system, our perceptual system faces an irreducible upper limit on information transmission, because of the biophysical constraints imposed by the biology of the brain. Therefore, the goal of the nervous system is to recode the sensory input to maximise the amount of information transmitted by increasing the independence in the signal while minimising the redundant signals (“Redundancy reduction hypothesis”,).

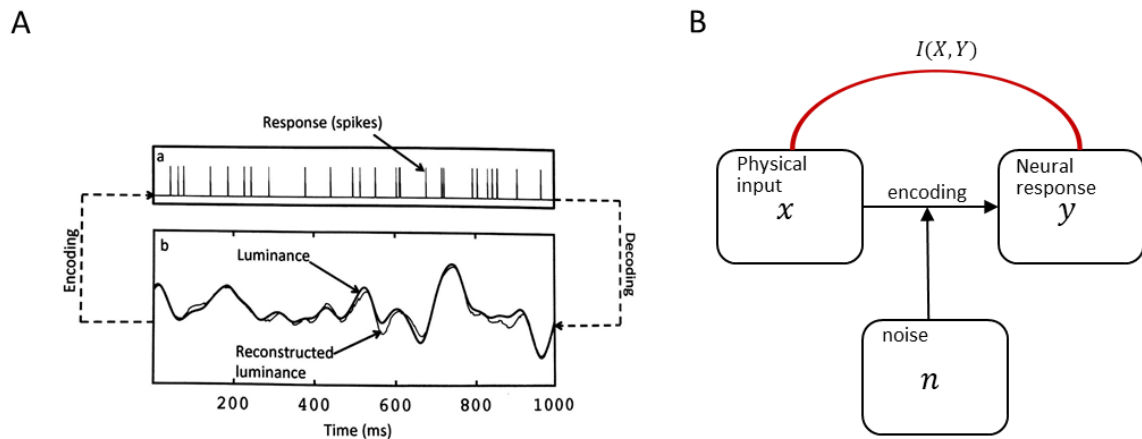


**Figure 2.1. Perception as an information transmission system**

(A) Figure copied from (Stone, 2018). Two processing stages of perception. The true luminance (thick black line) from a scene gets encoded or presented by the brain through neural spikes. The brain then needs to interpret this sensory signal as the luminance we observed (thin black line). (B) Schematic flowchart showing the visual processing stages given the input ( $x$ ) to the neural response ( $y$ ).  $I(x, y)$  refer to the mutual information of two variables:  $x$  and  $y$ . During encoding stages, noise might come in and corrupt the communication message. The goal of our visual system is to minimise redundancy in the system; that is to have the maximal mutual information between the input and neural response, given the maximum information neural codes can afford to transmit.

The concept of redundancy reduction originates from the information theory proposed by Shannon and Weaver (1963). For example, when we observe a visual scene, the sensory information ( $x$ ) is encoded in a perceptual signal ( $y$ ); Figure 2.1B. The degree of redundancy in this case is defined as the degree of dependency between input  $x$  and output  $y$ :  $1 - \frac{I(x,y)}{C}$ , where  $I(x, y)$  refers to the mutual information between  $x$  and  $y$ , and  $C$  refers to the capacity of the communication channel. The mutual information  $I(x, y)$  is

defined by the difference in the uncertainty over  $x$  (i.e., entropy on  $x$ :  $H(x)$ ) and the uncertainty over  $x$  given that  $y$  is fully known (i.e., conditional entropy on  $x$  over  $y$ :  $H(x|y)$ ). In other words, it represents how much uncertainty one could reduce about the visual scene by knowing the identity of the percept. Mutual information is also symmetric, that is, it is equivalent to the difference in the uncertainty over  $y$  and the uncertainty over  $y$  given we know about  $x$ :  $H(y) - H(y|x)$ . The channel capacity  $C$  is defined as the maximal mutual information a channel can transmit. Efficiency can be achieved by maximising the uncertainty over physical input  $x$  (or the neural response  $y$ ), so that the potential reduction in uncertainty is the greatest. Theoretical maximal uncertainty can be achieved when the probability of  $x$  (e.g., pixel values) or  $y$  (e.g., spike rate) is close to uniform or by minimising the uncertainty over  $x$  by knowing about  $y$  (or over  $y$  by knowing about  $x$ ). For example, when the spikes  $y$  are triggered noiselessly by the visual scene  $x$ , then the uncertainty over the visual scene is 0 if we know about the spikes. One cause of the redundancy is that there is an unequal probability of occurrence in the signal transmitted (Barlow, 2001). In other words, according to Barlow's efficient coding hypothesis, our neural system should account for the regularities in the environment by removing the statistical redundancy or increasing the independence between neural response to the visual stimuli (Simoncelli & Olshausen, 2001). Schwartz and Simoncelli (2001) showed that a model that normalises the visual filter response based on the responses of neighbouring filters removes response redundancies of neurons. This model can predict the neural response in the primary visual cortex and primary auditory cortex to natural images and sounds. Other studies have also shown that responses in the primary visual cortex have a low level of redundancies to natural stimuli (Reinagel, 2001; see also sparse codes in Section Representational Learning). While efficient coding appears to be a sensible coding mechanism in a biological brain, Brette (2019) has argued that a perfect efficient coding case is paradoxical since physical input that has its statistical redundancy removed means that the signal transmitted is indistinguishable from random. It is not clear how the brain would be able to decode such a message.

In fact, Machens and colleagues showed that neural efficiency depends not only on the statistics of the environment but also on the behavioural relevance to the task (Machens, Gollisch, Kolesnikova, & Herz, 2005), implying the importance of adaptive coding in the brain that is not only for the service of efficient coding, but also for the purpose of being able to act on the environment (see Section Adaptive gain model).

### 2.1.2. Representational efficiency

---

*This section describes two definitions of representational efficiency. The first definition of representational efficiency is an extension on the efficient coding hypothesis. The efficient coding hypothesis made no assumptions on how well should the information be encoded. Therefore, this definition of representational efficiency concerns with how well information is encoded given the constraint on the capacity limit of an information system (rate-distortion theory). The current thesis focuses on this definition of representational efficiency. The second definition of representational efficiency concerns with representational learning: how does an information processing system form representations. I will focus on discussing how the architecture of the system contribute to better representations of visual information. The current thesis does not investigate representational learning. Nevertheless, I included it to provide a comprehensive review on the literature.*

---

There are multiple conventional definitions on *Representational efficiency* depending on the question we are interested in. In the current thesis, the notion of representational efficiency confines to the definition which concerns with how well we can represent a given piece of information, given that it is constrained by the capacity limit of an information processing system. For an information-theoretic perspective, the major constraint of a processing system would be the communication capacity or information efficiency.

**Rate-distortion theory.** Barlow's efficient coding hypothesis derived many implications, one of which is that our representation of natural scenes should take into account the statistical properties we estimate from observing the physical world. Unlike

Figure 1.7A, which describes perception as two-stage processes, Barlow's efficient coding hypothesis focuses on the encoding of information (Figure 2.1B), it does not specify how the neural response  $y$  is decoded into a representation or how we act based on the inference of  $y$  with respect to input  $x$ . In other words, Barlow's efficient coding hypothesis does not make assumptions with respect to the accuracy in which the input is represented, given the communication constraints (Simoncelli & Olshausen, 2001). To measure the discrepancy between our representation and the actual physical input, the processing system needs to perform a reversed operation, where the neural response is used for reconstructing or decoding the input (Figure 2.1A). One can imagine that an organism is not only limited by how well the neural system efficiently communicates information regarding the external environment, but there should also be a cost for misrepresenting and wrongfully acting on the environment, similar to the notion of optimisation with respect to a loss function in BDT. The Rate-distortion theory allows a redescription of the redundancy reduction hypothesis into a constrained optimisation problem. It proposed that our perceptual system maximises the mutual information  $I(x, y)$  under the finite channel capacity  $C$  while also minimising the discrepancy between the representation reconstructed by the neural response and the actual input, aka 'distortion'  $L(x, \hat{x})$  (Figure 2.2). In other words, the theory provides a normative solution on how much distortion a system can afford, given the capacity constraints. The Rate-distortion theory has shown success in accounting for a number of perceptual and visual working memory limitations in humans, including perceptual generalisation (Sims, 2018), perceptual identification tasks (Sims, 2016), visual working memory capacity (Sims, 2016; Sims, Jacobs, & Knill, 2012) and even applications to behavioural economics (Caplin & Dean, 2015) and macroeconomics problems (Sims, 2003).

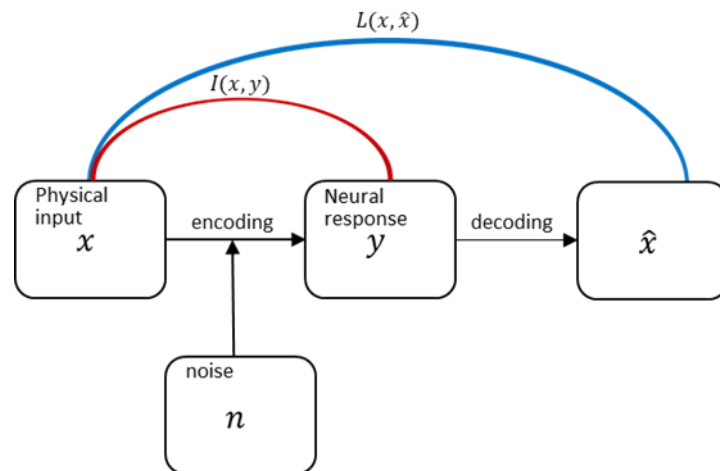
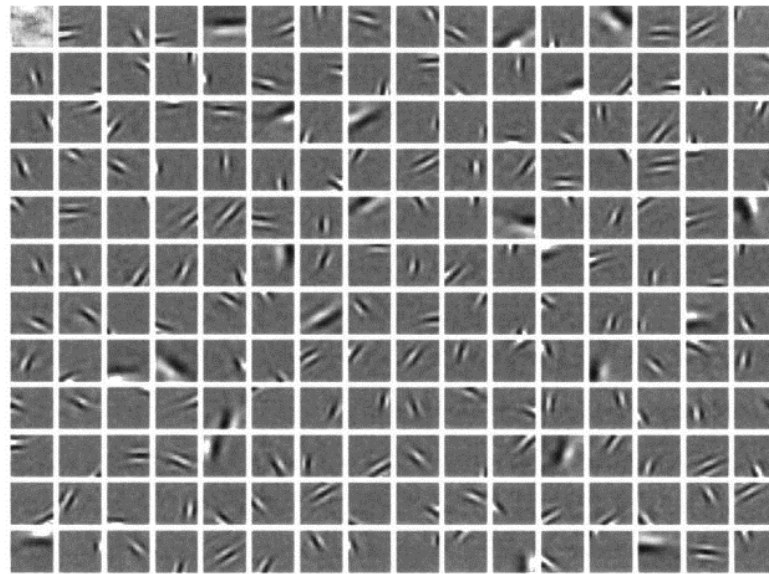


**Figure 2.2. Rate-distortion theory.**

The Rate-distortion theory builds upon the Information Theory in Figure 1.7B, and further suggests that there is a cost  $\mathcal{L}(x, \hat{x})$  for representing the input  $x$  inaccurately. A perceptual system first encodes the input  $x$  based on efficient coding principles, followed by a decoding stage where the neural response  $y$  is used for reconstructing the input into  $\hat{x}$ , for example, using Bayes' rule. According to the theory, the system measures and achieves the rate of distortion it can afford given its communication capacity.

**Representational Learning.** The alternative definition of representational efficiency is concerned with how the brain forms representations efficiently, i.e. *representational learning*. This is an idea machine learning scientists care about in order to build a machine that can perform image recognition without exhausting all of its computational resources, i.e., this definition of representational efficiency goes hand-in-hand with the notion of *computational efficiency*. Representation learning was first investigated in the 1990s when computational neuroscientists tried to understand how neural representations developed on the basis of input statistics. A number of theoretical models have been proposed. One way to estimate the input statistics is to make use of dimensionality reduction techniques (e.g., principal component analysis, factor analysis ;Dayan & Abbott, 2001), implying that the brain learns representations by extracting the most important features from a vast amount of high-dimensional neural signals. Others have proposed using a supervised learning method. For example, Olshausen and Field (1996, 1997, 2004) made use of sparse coding – a technique that assumed that information is represented by a small number of

active neurons from a large population to achieve energy efficiency. They showed that a neural network that is trained to reconstruct natural images while having to maximise



sparseness could learn a set of basis functions (receptive fields) that highly resemble simple cell properties in the primary visual cortex (Figure 2.3; (Olshausen & Field, 1996)). Generative models, which are motivated by Bayesian Inference, consider that the goal of the brain for representational learning is to learn the cause of the sensory input (i.e., the generative distribution) that allows economical description of the input but still allows the input to be reconstructed accurately (Dayan & Abbott, 2001; Friston, 2003; Hinton, Dayan, Frey, & Neal, 1995), a line of reasoning similar to Barlow's efficient coding.

**Figure 2.3. Learned basis functions from a sparse coding network.**

Figure copied from (Olshausen & Field, 1996). 192 basis functions from a neural network that maximises sparse coding and minimises reconstruction errors shares similarity with the receptive field filters in the early visual cortex.

More recent work from the field of machine learning showed that the ability of a machine to represent visual information or to perform a certain task depends on the architecture of the network. This gives rise to the following question: Can we construct a more efficient architecture for a neural network so that certain problems or functions can be approximated by fewer computational resources, fewer data and less use of human

knowledge to tailor the problem solver (Bengio, LeCun, & others, 2007)?. For example, a convolutional neural network (CNN), a class of artificial neural networks that can perform state-of-the-art pattern recognition nowadays, is a classic example of how hierarchical processing results in more efficient representations. The success of CNN is inspired by the discoveries in visual neuroscience. These discoveries include (i) the coding properties in visual areas discovered by (Hubel & Wiesel, 1962) and (ii) the visual hierarchy where early visual regions code for simple features (e.g. edges, orientation) and successive regions code for ensembles of features such as objects (LeCun, Bengio, & Hinton, 2015). Another study has shown that the third and the last layer of a trained CNN is highly predictive of the neural response in V4 and the inferior temporal cortex (Yamins et al., 2014). These areas are located in higher processing streams and display object-selectivity. These studies showcase how representational efficiency can be optimised by a machine with the structural characteristics of human visual processing.

As mentioned above, one key characteristic of our visual regions is that representations are formed in a progressively more complex and holistic way. The CNN was much inspired by this idea as it adopts a deep or hierarchical network architecture, i.e., a structure that involves multiple layers of units that are sequentially connected. A few studies suggest that there are benefits of having a “deep structure” network (a network that contains multiple layers). The benefits include: better generalization ability (LeCun et al., 2015), a need for fewer computational resources, and fewer training examples to train (Bengio et al., 2007). It becomes an interesting question of why hierarchically structured networks such as our brain or an artificial deep network is so powerful and efficient in solving different problems such as learning representation, planning, action control, etc. (Balaguer, Spiers, Hassabis, & Summerfield, 2016; Koechlin & Summerfield, 2007; Poggio, Mhaskar, Rosasco, Miranda, & Liao, 2017; Sanborn, Bourgin, Chang, & Griffiths, 2018). For instance, hierarchically structured semantic knowledge allows one to generalise information about members of the same category, and reduce the amount of information one is required to store.

### 2.1.3. Computational efficiency

---

*This section discusses a more general definition of efficiency—computational efficiency. I first describe divisive normalisation, an efficient neural computation that makes neural responses more statistically independent and thus reduce the metabolic costs. Apart from considering neural costs as type of computational cost, I discuss the idea of computational rationality, which focuses on solving a computation problem while taking into account the computational costs. I use the choice overload effect and “model-free and model-based learning” as examples to demonstrate how computational costs influence how we make decisions.*

---

All forms of efficiency can be put under the umbrella term “computational efficiency” once we take into account that operations involved in any form of efficiency are performed by an information processing system that has limited computational resources. Information efficiency started with the idea that neural coding is resource costly. As such, the primary motivation of the efficient coding hypothesis is that it will be computationally expensive if our visual system represents the physical input like a computer, where each pixel is independently processed. By extracting the input regularities, the sensory system can represent the same physical input with a lower metabolic cost. Representational efficiency focuses on the consequence of a computational mechanism that trades off the precision of stimulus encoding and the neural resource required for encoding. Instead of focusing on how the brain forms an optimal trade-off, the metabolic constraint of the brain has motivated a new proposal: the brain uses a set of canonical efficient computations for multiple problems. One efficient algorithm is called *divisive normalisation*, which has shown success in explaining the responses in sensory areas, and more recently it was proposed that reward processing areas in the brain also perform divisive normalisation. Thus it might be a “canonical neural computation” in the brain (Carandini & Heeger, 2012).

**Divisive normalisation.** Divisive normalisation assumes that responses of neurons are divided by the summed activity of a pool of neurons:

$$R_j = \gamma \frac{D_j}{\sigma + \sum_k^K D_k}$$

Where  $R_j$  refers to the normalised response of neuron  $j$  driven by the input  $D_j$ . The normalisation equation states that the driving input  $D_j$  is normalised by the total input from a pool of neurons  $K$ , scaled by some parameters  $\gamma, \sigma$  that can be fit to the data to control the degree of normalisation.

For example, surround suppression is an observation commonly found in the primary visual cortex. When an oriented Gabor is placed inside a placeholder that has a different orientation, the responses of the neurons are suppressed. Figure 2.4 illustrates the response of neurons in V1 under different stimulus conditions. Each stimulus is defined by a central oriented Gabor with different contrast levels, which is placed inside an annulus which is also associated with different contrast levels. The annulus itself elicited no neural response (left most signs). When the central stimulus is presented alone, it elicited high firing rates (yellow circles). However, when it is presented together with the annulus, the firing rate is suppressed, and it is the most suppressed when the annulus has a high contrast (purple triangles). Divisive normalisation can also explain feature-based attention. Treue and Trujillo (1999) presented a random dot motion pattern in the receptive field or outside the receptive field of the recorded neuron in area MT. The monkey was cued to attend to either the fixation point or the stimulus. They found that when the stimulus falls in the attentional modulation field, tuning curves of neurons become sharpened (

Figure 2.5). Simulated neural response using the divisive normalisation model showed that there is an increase in firing rate when attention is directed toward the neuron's preferred stimulus (Figure 2.5C; red line) compared to the case where attention is assumed to be equal for all motion directions (Figure 2.5C blue line). Apart from divisive normalisation, other efficient algorithms have been proposed. For example, range adaptation is another form of normalisation algorithms that have been proposed to explain relative value codes in the orbital frontal cortex (Padoa-Schioppa, 2009). The motivation for this normalisation scheme is based on the observation that the range of stimuli can serve as a contextual factor in influencing people's judgment (Parducci, 1974), and there is an upper bound on firing rates. In Chapter 5, I will provide a detailed description of

different forms of efficient computations and their role in context-dependent value-based decisions.

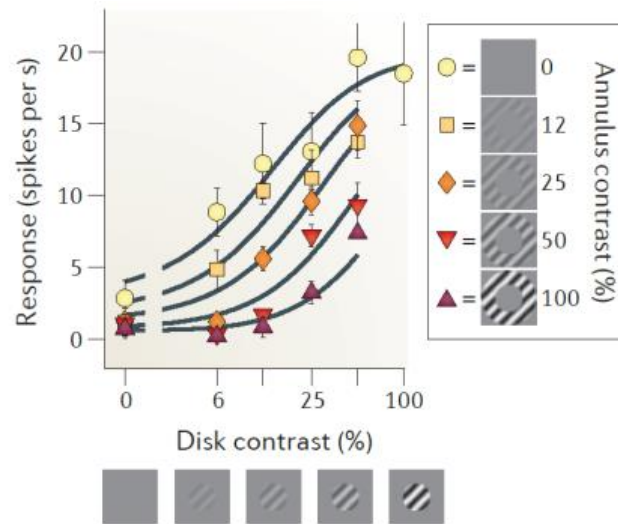


**Figure 2.4. Divisive normalisation and surround suppression**

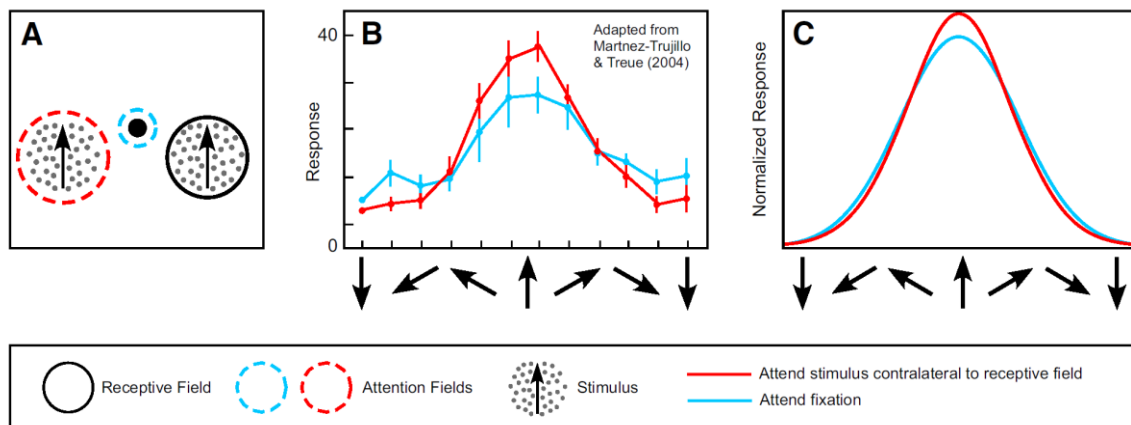


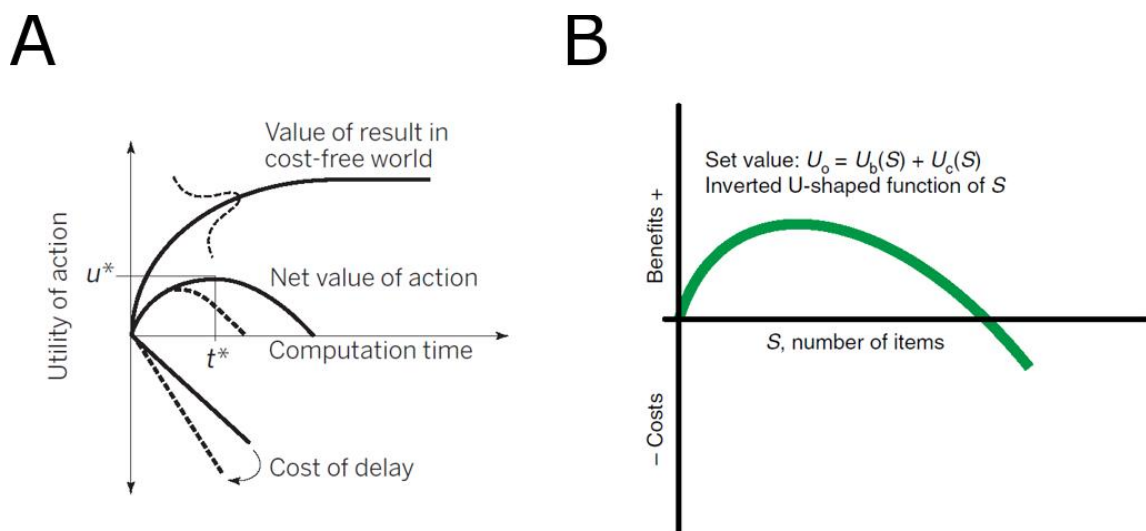
**Figure 2.5. Divisive normalisation and feature-based attention**

Figure copied from (Reynolds & Heeger, 2009) (A) Task. (B) empirical tuning curves of neurons in area MT. (C) divisive normalisation model simulated firing rate

**Computation time as a source of computational costs.** The idea of computational efficiency goes hand in hand with *computational rationality*, where the goal is to “identify decisions with the highest expected utility while taking into consideration the costs of computation in complex real-world problems in which most relevant calculations can only be approximated” (Figure 2.6; Gershman, Horvitz, & Tenenbaum, 2015). When trying to achieve a goal, an information system is faced with two kinds of computational costs: the cost of finding the optimal solution and the cost of executing the optimal solution (Lewis, Howes, & Singh, 2014). These costs can often be measured in terms of the time it takes to find the solution and the precision of the solution. For example, one might intuitively imagine having more available options in a choice set is valuable (e.g., most people might prefer to shop from a big supermarket than a local convenience store). However, Reutskaja, Lindner, Nagel, Andersen and Camerer (2018) showed that having too many options can result in a negative utility, which they referred to as a “choice overload” effect.

**Figure 2.6. The computational mechanism of bounded rationality**

(A) Figure copied from (Gershman et al., 2015). The idea of computational rationality does not



only focus on maximising the expected value of an action under constraints (i.e., bounded rationality) but also specifies how the constraints might have arisen from intrinsic limits within an information processing system or from an information processing mechanism. Rational or optimal actions, in this case, would maximise the expected value of the action given the reward and the computational cost (measured by the computational time in this case) required for solving a problem. (B) Figure copied from (Reutskaja et al., 2018). The idea of computational rationality is

consistent with the observation that more available options in a choice set (x-axis) do not necessarily provide extra utility benefits (y-axis) due to limited cognitive capacities. In the economics literature, this is termed as the “choice overload” effect.

Another well-known example of computational efficiency would be model-free versus model-based reinforcement learning in solving sequential problems (Daw, Niv, & Dayan, 2005; Gershman et al., 2015). A model-free reinforcement learning method is a model that learns values of an action or a state by experience (i.e., trial and error) and caches these values in a look-up table. In other words, a model-free method requires slow and inflexible learning, but finding a solution (i.e., picking the best action) is easy. Model-based reinforcement learning proposes that the problem solver forms a model of the environment they encountered, such as forming a model of how different states transit to each other, and solutions can be found by searching through possible branches of action. One advantage of this method is that learning is efficient and flexible since one can quickly adapt its model of the world when the environment changes without having to act and experience. The disadvantage of this model is that it is computationally expensive to find the optimal action and its consequence (e.g., finding the course of moves that allow one to win a chess game). Humans appear to exhibit both model-free and model-based learning, with different neural substrates underlying the two types of computation, suggesting that humans might trade-off both learning methods to achieve efficiency and flexibility (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw et al., 2005).

## 2.2. Efficiency-inspired models

---

*In this section, I will review different computational models that took the three types of efficiency as computational principles. I will first discuss the Bayesian Observer model, a model which considers that perception or subjective preference on economic goods involves both efficient information encoding (Information efficiency) and Bayesian decoding (related to Representational efficiency). Then, I will discuss models focus on the trade-off between neural representation of information and neural costs – i.e. binary encoding model and its biologically-plausible counterpart, and the resource rational model. Lastly I will focus on the adaptive gain model, a model that*

*is related to all three forms of efficiency, and discuss its existing applications on explaining perceptual decisions.*

---

In the following sections, I will review computational models of perceptual decision making that are principled on the three forms of efficiency. Some of the models were proposed to explain visual working memory representations. Nevertheless, I consider that these models are extremely relevant since the computational framework of these models fit well with the perceptual decision-making domain. This is the case because these models focus on how the brain might trade-off between the fidelity of neural representations of physical inputs and the resource required for representation. In Chapter 5, I will describe the background and behavioural predictions of a class of algorithms (based on the normalisation principle) for economic decisions. Therefore, in the current section, discussions of plausible applications for economic decisions will pertain to models which have a broad and general computational framework.

#### 2.2.1. Efficient (en)coding and Bayesian decoding.

Most of the theoretical and empirical investigations of how the brain forms efficient representations have focused on how the neural system encodes statistical regularities in the external environment, e.g. how the neural tuning properties could change as a function of the environmental statistics (see Bayes optimal perception, Information efficiency and Representational efficiency in previous sections). However, how the brain adapts its decoding or readout policy to form decisions is less understood (Gold & Stocker, 2017). Should neurons adapt to the environment by changing their coding properties rather than changing their decoding policy? What is the separate contribution of adaptation in the encoding and decoding processes in behaviour? These questions remained unanswered.

**Decoding neural signals into physical variables.** How does the brain adapt its decoding policy according to environmental needs? To answer this question, we first need to understand how the brain decodes neural signals. One of the decoding strategies

proposes a decoder that abides by Bayes' rule, which has been seen as the optimal decoding policy (Deneve, Latham, & Pouget, 1999; Pouget, Dayan, & Zemel, 2000; Zhang, Ginzburg, McNaughton, & Sejnowski, 1998). Activities of sensory neurons are inherently noisy, and thus the tuning curves for each neuron that specify the relationship between the input magnitude and the firing rate is not a delta function but a smooth function that peaks at the neuron's most preferred sensory input magnitude (e.g., orientations). A Bayesian decoder assumes perception to be a probabilistic inference process, and it considers each neuron's tuning curve as a likelihood function that represents the likelihood of a range of stimulus magnitudes (e.g., orientations) that give rise to the observed firing rate of that neuron. When applying Bayes' rule, i.e., combining the prior with the likelihood function, each neuron approximates a posterior distribution that describes the probability that a certain orientation is the true observed orientation, given its firing rate. The posterior distribution can then be used for creating a point estimate of the inferred orientation - given the neural signal - using either the maximum a posteriori (MAP, i.e., taking the estimate that has the highest posterior probability or the estimate at where the posterior distribution mode is) or the maximum likelihood estimation (MLE; i.e., taking the estimate that maximises the likelihood function) (Pouget et al., 2000). MLE and MAP give the same estimates when the prior of the stimulus is flat. MLE has been shown as the optimal decoding method provided that there is a large population of neural data (Deneve, Latham, & Pouget, 1999; Jazayeri & Movshon, 2006; Pouget, Dayan, & Zemel, 2000; Figure 2.7A). However, in the case where the prior is not uniform, both methods failed to explain the "non-optimal" percepts humans have. This is because MLE does not take into account the stimulus prior as it is only concerned with the joint likelihood of the neural response given the stimulus, and MAP estimates always predict an "optimal" percept, i.e., the perceptual estimate lies on the mode of the posterior distribution and the posterior distribution shifts in the direction where the prior peak lies. In other words, MAP would always predict an attractive bias whereby the neural tuning, as well as behavioural reports of the observed percept, are biased toward the adapted stimuli (assuming the adapted stimuli formed the prior expectation). However, both neural and behavioural reports in

animals and humans show that a percept can sometimes repel away from the adapted stimulus (Clifford, 2002; Dragoi, Sharma, & Sur, 2000; Kohn & Movshon, 2004).

**Bayesian Observer model.** As mentioned previously, perception can be seen as a two-stage process. Ideally, perception should contain an encoding operation where sensory representation is optimised with respect to the environmental statistics (efficient coding) and a decoding operation where the neural signal is translated into a physical variable faithfully (e.g., using a Bayesian decoder). Wei and Stocker devised a model that includes both components – the Bayesian Observer model to explain the “anti-Bayesian” percepts described above (Wei & Stocker, 2015, 2017). Interestingly, this model still assumes an optimal Bayesian decoder. Since it is a Bayesian decoder, the decoding policy adapts as a function of the prior, i.e., the environmental statistics.

Let’s consider perception as a probabilistic inference problem, a true stimulus value (e.g., a grating with orientation  $\theta$ ) evokes some noisy neural responses  $m$ , resulting in a likelihood function  $p(m|\theta)$  for each neuron. The goal of the observer is to decode  $p(\theta|m)$  into a percept  $\hat{\theta}$ . Deviation between the decoded percept and the true orientation is the perceptual bias (Figure 1.14B). In the Bayesian Observer model, the encoding of the stimulus value into neural response follows the efficient coding principle, i.e., each neuron’s likelihood function should depend on the stimulus distribution.

To achieve this, the authors made the following assumptions: 1) a separation of the stimulus space and sensory space, and 2) the fisher information (a measure of encoding accuracy) of the sensory representation in the stimulus space is proportional to the prior distribution and uniform in the sensory space. Regarding the first assumption, a certain value measured in the units of stimulus space can be represented as an equivalent value in the units of sensory space by a function  $F(\cdot)$  where  $\tilde{\theta} = F(\theta)$ . Conversely a value in the sensory space can be converted into a stimulus value by applying the inverse function  $F^{-1}(\cdot)$  such that  $\theta = F^{-1}(\tilde{\theta})$  (Figure 2.7C). The two spaces differ in the sense that the fisher information in the stimulus space can be non-uniform, but the fisher information in

the sensory space must be uniform. How does fisher information relate to efficient coding? According to Wei and Stocker, fisher information can serve as an approximation of the mutual information (Wei & Stocker, 2016). Take the simplest case where there is perfect learning of the stimulus distribution and thus our prior distribution  $p(\theta)$  matches the stimulus distribution perfectly; the prior should be proportional to the square-root of the fisher information (i.e., the extent to which noisy neural responses  $m$  help us to infer the true state  $\theta$ ) under the assumption of small Gaussian noise (Wei & Stocker, 2017). Therefore, the second assumption turns the encoding process into an efficient coding problem of maximising mutual information. The assumption about the uniformity of fisher information in the two spaces results in asymmetric likelihood functions in the stimulus space. These asymmetric likelihood functions are then used for Bayesian decoding. The posterior distribution  $p(\theta|m)$  can be computed by combining both the likelihood function  $p(m|\theta)$  and the prior  $p(\theta)$ . In other words, when the stimulus distribution changes, not only does the likelihood function change, the decoding policy also adapts to it. To get a point estimate (i.e. the percept), the observer can choose a loss function and aims to minimise the expected loss (see Section Bayesian Decision Theory). The authors tested the prediction of three different loss functions (binary error term -  $L_0$  loss; absolute error -  $L_1$  loss; squared-error-  $L_2$  loss). This is equivalent to using the mode (i.e., MAP), median, and the mean of the posterior distribution as the estimators respectively. They showed that both the  $L_1$  and  $L_2$  loss functions are able to qualitatively capture the perceptual biases and the discrimination threshold humans have for a variety of perceptual tasks (Wei & Stocker, 2015, 2017). One difference between the Bayesian account of efficient coding and models that consider efficient coding as the sole principle for neural codes is that the latter is a special case of the former where the loss function is defined as the posterior entropy (Park & Pillow, 2017). This is much more restricted than the broader idea that the choice of loss function and optimality might depend on the task at hand (e.g., see Wang, Stocker, & Lee (2016) for an investigation of the optimal  $L_p$  loss function for different tasks in different species).

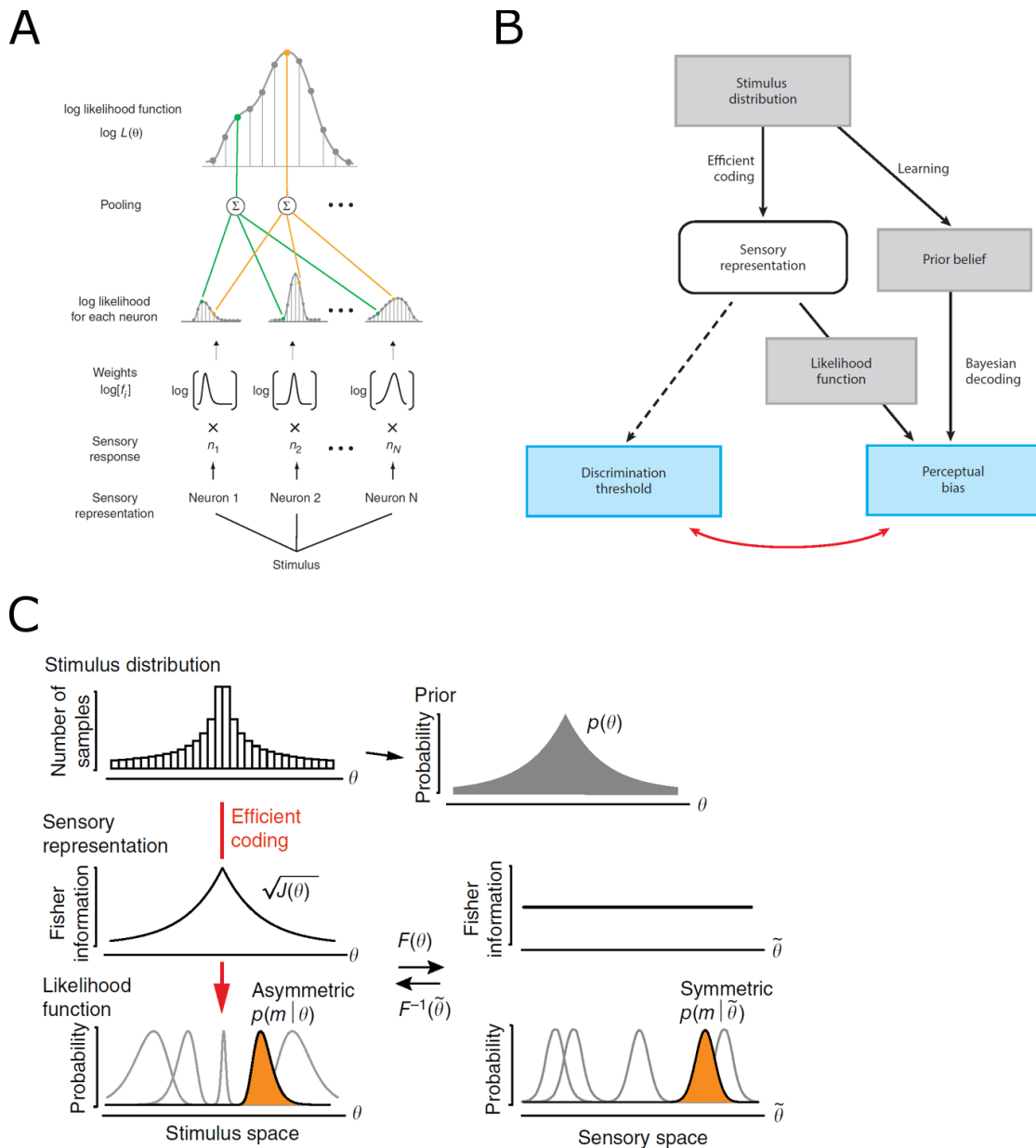


**Figure 2.7. Decoding perceptual variables from the neural activity using Bayesian methods**

(A) An example of optimal decoding using maximum likelihood estimation (MLE). Figure copied from (Jazayeri & Movshon, 2006). Jazayeri and Movshon proposed an optimal decoding strategy based on MLE where they approximate the joint likelihood function by pooling the log likelihoods from each neuron. The log likelihoods from each neuron in this decoding strategy is a composite of the neuron's logarithmic tuning curve and the number of spikes elicited by the stimulus. (B) The Bayesian observer model that adapts both the encoding and decoding strategies as a function of the environmental statistics. Figure copied from (Gold & Stocker, 2017). Perceptual bias can be simulated by comparing the model estimated percept and the true stimulus value. Based on previous work on the oblique effect, the discrimination threshold is assumed to vary inversely

with the prior probability densities. A stimulus that is more probable should have a lower discrimination threshold. (C) An illustration of how the stimulus distribution influences the likelihood functions in the stimulus space. Figure copied from (Wei & Stocker, 2015).

**Bayesian Observer model in value-based decisions.** In the value-based decision domain, humans often are assumed to be rational (Section 1.2). One common premise from these rational principles is that preferences are assumed to be consistent and noiseless, an assumption diametrically opposite from how scientists view human perception. Discoveries on “irrational” human value-based decisions are often attributed to the use of suboptimal heuristics or intuitive cognitive strategies in the economics literature (e.g., Bushong, Rabin, & Schwartzstein, 2015; Kőszegi & Szeidl, 2012; Tversky & Kahneman, 1974). More recently, studies have started to borrow the view that irrational value-based decisions might be based on preferences that are noisy in nature (Howes, Warren, Farmer, El-Deredy, & Lewis, 2016; Tsetsos, Moran, Chater, Usher, & Summerfield, 2016), similar to how SDT and Bayesian theorists view perceptions.

Polania, Woodford and Ruff (2019) brought forward the Bayesian framework on decision-making, where preferences are noisy and latent in nature. When an observer inspects a good, instead of assuming that this observer has an immediate and precise representation of the value and preference for this item, they assumed that the observer is agnostic about the value of the item ( $v$ ), but they can infer the value from some noisy sensory responses ( $r$ ) elicited by the item from the conditional probability distribution  $p(r|v)$  (Figure 2.8). This conditional probability distribution (i.e., the likelihood function) describes the probability density of some noisy sensory responses given the value of the item. The value of the item is a random variable sampled from a prior distribution  $p(v)$ , which is a probability distribution of values of the goods in the world. This prior distribution is possibly formed according to the observers’ past experience. As for the likelihood function, it is constrained by the prior based on the assumption by Wei and Stocker that there are two spaces functioning in parallel (one where fisher information follows the prior distribution, i.e., the ‘stimulus space’ and one where fisher information is

uniform over the space). Applying Bayes' rule, one can then form a subjective estimate about the value of the item ( $\hat{v}$ ) that maximises the joint likelihood of the subjective estimate and the true value ( $v_0$ ) of the item, given free parameters that describe the prior distribution (i.e., the mean and variance of a logistic function) and noise that is injected during the encoding or after the decoding of the value estimate. Lastly, to map this internal subjective estimate onto a bounded rating scale (since the experimenter can only assess the internal value of an option through the rating on that option), the subjective estimate is transformed into a rating using a sigmoid function.

They tested the model using four behavioural experiments that involved a rating task where participants rate their preference on each good twice to measure the rating consistency, a 2 alternative forced choice task between two goods to measure how choice pattern varies with the rating variability, and lastly a confidence task to measure how the first-order rating of a given good relates to the second-order judgement (i.e., confidence) on the rating they gave. Testing how participants rate 64 familiar food items, participants showed greater variability for items that they do not have a strong preference for, with ratings that are more consistent for more preferred items than less preferred items. Recovered priors from their model displayed a shift of the prior mean toward high-value items, resulting in an effect akin to a decreased perceptual discriminability threshold (increased acuity, and thus decreasing the choice variability) on items that are high in value. They further measured how bias from the choice task can be the result of limiting the presentation time of the goods. Assuming that shorter presentation time of the goods means increasing encoding difficulty (i.e., modelled by increasing encoding noise in their model), both humans and the model displayed a "preference-illusion"; items with values adjacent to the prior peak have a subjective estimate that is repelled away from its true estimate (i.e., a repulsive bias). Moreover, as the encoding noise increases, the prior belief becomes stronger, leading to a larger repulsion bias. Finally, assuming that the precision of the posterior distribution of the subjective estimate is the basis for a Bayesian observer to form confidence, they found that the model correctly predicts that confidence scales

monotonically with the rating consistency of items, with an overall higher confidence for high valued items (due to the shift of the recovered prior as mentioned above).

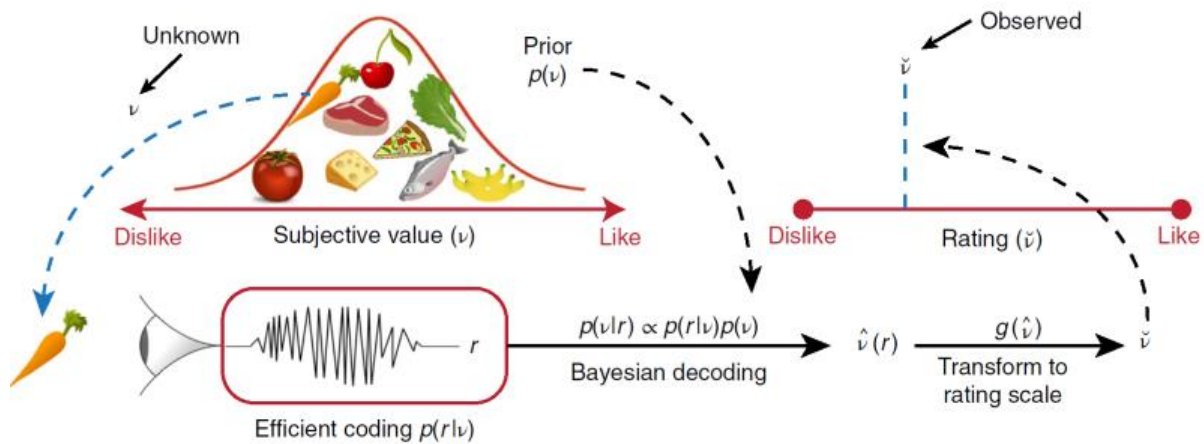


**Figure 2.8. Bayesian Observer model on preference-based decisions**

Figure copied from (Polania et al., 2019). In the model, observers have no access to the true stimulus value  $v_0$  of the item (carrot), but the momentary value of the item  $v$  can be inferred by Bayesian inference that is constrained by efficient coding. An observation of the good results in a noisy internal response  $r$  on that item (i.e., likelihood  $p(r|v)$ ). Importantly, the likelihood is constrained by the prior distribution  $p(v)$  (which is assumed to be a logistic function) since the value of the item must draw from the prior distribution. Assuming an optimal Bayesian decoder, a subjective estimate on the value  $\hat{v}$  with respect to the posterior distribution  $p(v|r)$  is determined by assuming a  $L_2$  loss (equivalent to using the mean of the posterior distribution as the estimator). Lastly, the reported rating of that given good is derived from a mapping between the subjective estimate to a bounded rating scale using a sigmoid function  $\tilde{v} = g(\hat{v})$ .

**Efficient coding with Decision by Sampling.** In Chapter 1 (Section 1.2.3. The uprising of bounded rationality), I have described the Decision by Sampling (DbS) account as an explanation on why humans are risk averse and loss averse. According to Decision by Sampling, many of the context-dependent effects are a consequence of a sampling process in order to construct a value estimate of an item, where the samples should reflect roughly the natural statistics of the world. Further theoretical work by Bhui and Gershman (2018) showed that in the unlikely case where infinite samples can be drawn to construct the value

of an option, DbS essentially can be understood as an implementation of efficient coding. More specifically, a maximisation of mutual information between drawn samples and the internal representation of the item. The idea is that if the value of an option is constructed by infinite samples, these drawn samples should follow a cumulative distribution function (cdf). Evaluating the value of this option with respect to the cdf corresponds to the probability that this option is better than all other options in our memory. In statistics, a cdf has the property that the probabilities drawn from it follow a uniform distribution ('probability integral transform'). An analogy for understanding this probability integral transform is to consider the cdf as a transformation of the original data into percentiles in the discrete variable case. Each data point has now a unique percentile that denotes the probability that other data points are smaller in value than this data point. In other words, the probability of getting a specific percentile from the data is as likely as getting another specific percentile from the data (e.g. what is the probability of having a 100<sup>th</sup> percentile data point vs. what is the probability of having a 40<sup>th</sup> percentile data point). The same logic applies when we ask what is the probability of drawing a particular probability from the cdf.

A new generalisation of DbS, to accommodate the fact that samples are likely to be finite, showed that smoothing these finite samples with a uniform kernel prior to ordinal comparisons naturally implements the range-frequency adaptation (Parducci, 1965, 1974) – a model that assumed that the value of an item depends on both the rank ("frequency") and its position relative to the range of the values from a set of items ("range"). In particular, the "range" component has been considered as a form of efficient computation performed in the brain so that the neural activity can efficiently represent the value of an option (see Chapter 5 for a detailed discussion of this form of computation). Depending on the number of samples being drawn, an optimal kernel bandwidth can be computed (where optimality is assumed as achieving maximal mutual information between the samples and the internal representation of the option). Together, it implies that human "irrationality" can be seen as an outcome of a system that maximises its processing

efficiency within limited capacities, e.g. bounded rationality. In the next chapter, I will explore what we mean by efficiency.

### 2.2.2. Representational efficiency in visual working memory.

**An Information-theoretic approach on visual working memory (VWM).** Human visual working memory – a form of memory that needs to be actively maintained or manipulated and is associated with a limited capacity (both in terms of storage and manipulation capacities), is a perfect example of representational efficiency; i.e. the brain has to optimise for representing memorised information with high fidelity under capacity constraints. Early work proposed that VWM has a capacity of  $7 \pm 2$  items (Miller, 1956a), other work proposed capacity of three to four items (Brady, Störmer, & Alvarez, 2016; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001). Subsequently, Luck and Vogels proposed that VWM capacity is defined as the number of independent representations one can store (slot model), which is three to four slots. More recent work showed that the VWM capacity is limited in the sense that memory precision decreases exponentially as the number of items that need to be stored increases, as if there is a central pool of resources to be allocated across the items rather than having a fixed number of resource slots (Bays & Husain, 2008; Ma, Husain, & Bays, 2014). The idea of having limited resources and thus resulting in limited capacities for item storage is in fact inspired by an information theoretic perspective proposed by one of the pioneer scientists on VWM: George Miller (Miller, 1956b). Take the case where we have to store four unique alphabets, e.g., [A, B, C, D] in information bits using binary code. In order for the system to distinguish the four letters, let's say we represent A as [00], B as [10], and C as [01] and D as [11]. To represent all 4 letters in a stream, that would require  $4 \text{ (letters)} \times 2 \text{ (bits)} = 8$  information bits. In other words, the more unique items we need to store, the longer the binary code we need to construct, the more bits the system needs to store.

**Binary encoding model.** What about the case where the items that we need to store are not discrete objects? Nassar and colleagues proposed a binary encoding model

with similar propositions as Miller has described. The key idea is that features that stem from a continuous physical space get represented and stored in a discretised way due to capacity limits (Nassar, Helmers, & Frank, 2018). Let's consider that we discretise the full-colour space into two bins. This means that colours with different hue within each bin get compressed as having the same representation. Thus, only 1-bit binary code is required in order to represent the two bins separately (Figure 2.9 left panel). By further discretising the space, each colour bin will require longer binary code lengths in order for the system to represent all bins distinctly (Figure 1.15 middle and right panel). In other words, the more precise we need to represent a certain colour hue, the more discretisation the system has to perform (assuming the space is discretized evenly). This results in a longer binary code that the system has to be able to store. Take the example of a colour delayed-estimation task where participants are asked to memorise six coloured shapes. After a delay, they need to recreate the colour in the cued location from a colour wheel (Ma et al., 2014). Let's assume that participants' memory capacity is limited to a binary code length of 15 bits. In this case, participants would either need to discard storing some items since they are able to store all items if they represent each colour bin in 3 bits ( $6 \text{ colours} \times 3 \text{ bits} = 18 \text{ bits} > 15 \text{ bits}$ ; Figure 1.15 right panel) or they can store all items in a lower precision by partitioning the colour space more sparsely ( $6 \text{ colours} \times 2 \text{ bits} = 12 \text{ bits} < 15 \text{ bits}$ ).

This formulation of the VWM capacity highlights a specific problem: what are the units of capacity? For example, in a letter span experiment, participants were able to recall longer letter spans if the experimenter introduced temporal pauses during the presentation that allowed participants to form meaningful or pronounceable phrases from some letters (e.g., I-B-M-pause-H-B-O; Bower & Springston, 1970). As Cowan (2001) concluded, VWM capacity can be hard to measure as it depends on whether participants can chunk or group information. In fact, chunking is perhaps an adaptive and efficient mnemonic strategy when one is facing storage limitations. Sims (2016) proposed that chunking can be incorporated into the Rate-distortion theory, such that our brain needs to arbitrate the rate of distortion to take into account the capacity limits – i.e. reaching

representational efficiency. Nassar and colleagues showed that there is an information-theoretic way to derive the best way to chunk information given a strict limit on information storage (Figure 2.9B). In their study, various set sizes (1 to 8 colour features) had to be stored given a fixed limit of 15 bits. The authors showed that a binary code model that partitions the colour space so that the stored colours are separated by the greatest distance from the boundary of partitioning (criterion-based partitioning) outperforms a model that does not chunk (i.e. full partitioning). The criterion-based model almost performs as good as the optimal partitioning model which partitions the colour space in a way that achieves the lowest possible number of errors.

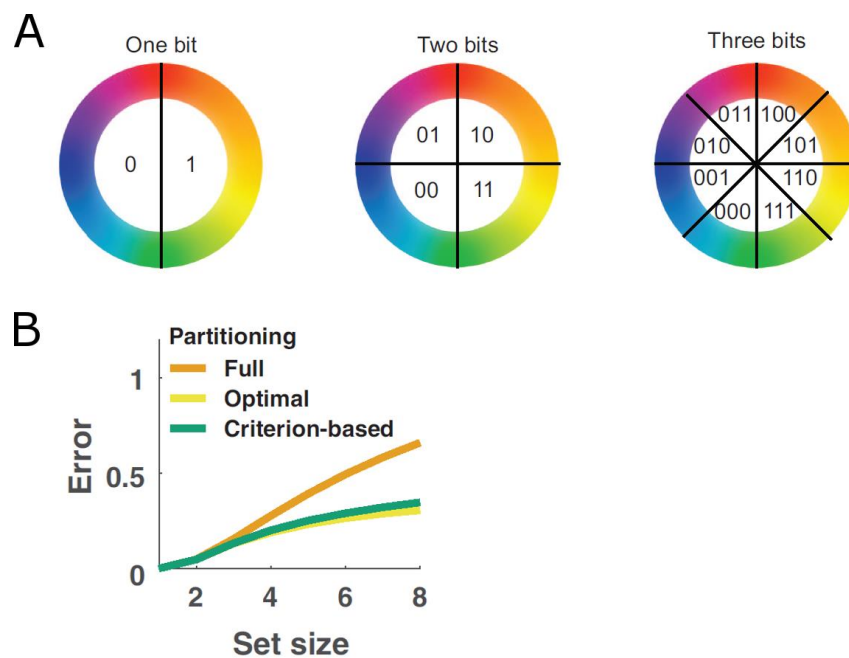


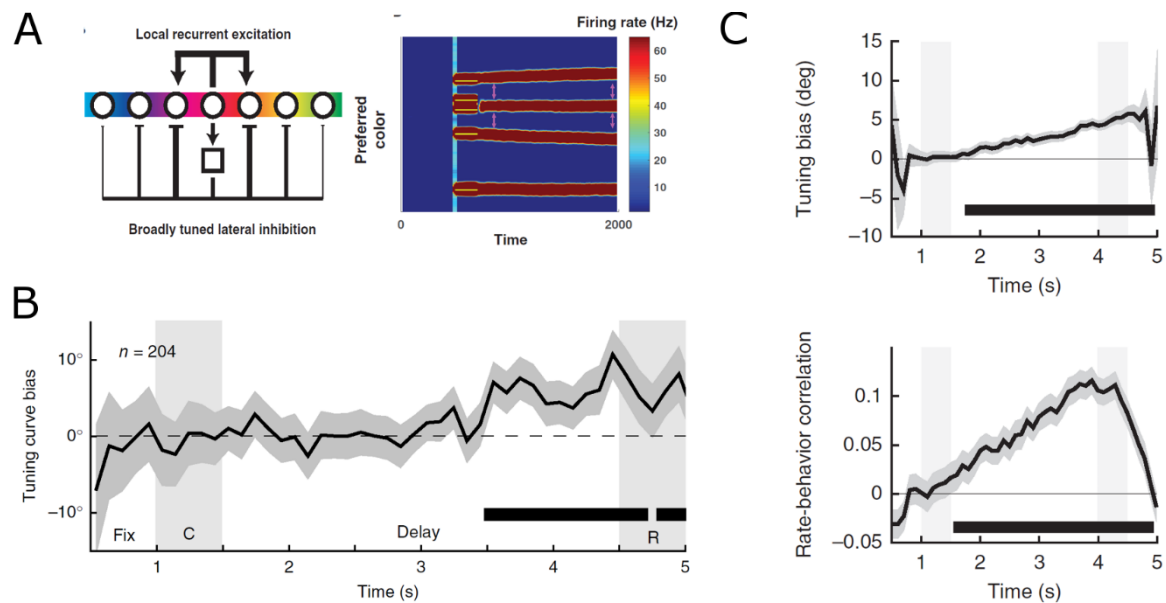
**Figure 2.9. Chunking as a rational strategy for visual working memory**

Figure copied from (Nassar et al., 2018). (A) Colours in a colour spectrum can be described using binary code. The more we partition the colour space, the more precise we are able to represent certain colours. In other words, the binary codes need more information bits in order to represent each unique colour ration. As a result, the binary codes take up more space in the storage system. (B) Forming chunks of representation results in better performance compared to not partitioning the representation space (Full partitioning).

**A biologically plausible model based on attractor dynamics.** Studies have created biologically circuit models to explain how VWM representations (including chunking) are formed. These studies provide an opportunity for us to understand how representational efficiency is achieved by the brain at the implementation level.

A biological plausible circuit model based on attractor dynamics and realistic biophysical properties of the human brain has displayed success in reproducing behavioural biases and errors in a VWM task made by human subjects (Nassar et al., 2018), as well as many neural signatures including neural tuning bias, neuronal variability and the neural-behaviour correspondence from non-human primates (Compte, Brunel, Goldman-Rakic, & Wang, 2000; Wimmer, Nykamp, Constantinidis, & Compte, 2014). The model architecture can be seen in Figure 2.10A left panel. The model circuit contains a Mexican-hat connectivity structure: strong local excitatory connections and a broad synaptic inhibition. This results in a net excitation profile that resembles a Mexican-hat shape (high net excitation at the feature that is being cued for retrieval in a VWM task and a moderate amount of net inhibition to the neighbouring features). The right panel shows the simulated firing rate for a pool of neurons during a delay period. The yellow bars correspond to the actual colour that is being cued, and red shaded strips show the firing rate of neurons neighbouring the cued colours. The circuit model predicts that during the early delay period, activity patterns of neurons that have similar cued colour preferences collide, resulting in a grouped representation of multiple colours. Later during the delay period, activity patterns from groups of colour become more distinct, implying the mnemonic strategy – chunking involves a two-stage process of first grouping and then partitioning. More convincing evidence from Wimmer and colleagues showed that this model was able to form neural predictions that match with empirical observations from a population of spatially selective neurons in the prefrontal cortex (PFC; (Wimmer et al., 2014)). In that study, monkeys were trained to perform a spatial working memory task where they need to memorise a cued location and make a saccade to the memorised location after a delay period. They measured the saccadic error (angular deviation from the

target location) and found that the directional bias in the tuning of spatially selective neurons in the PFC predicted the errors made by the monkey. More importantly, they



showed that a bump attractor model (a model with similar dynamics as described above) could replicate the neural tuning bias as well as display a neural-behaviour correspondence (Figure 2.10B-C).

**Figure 2.10. A bump attractor model for spatial working memory.**

(A) Figure copied from (Nassar et al., 2018). Left panel: A schematic of the model. Each circle corresponds to a neuron arranged according to their preferred colour. Lines correspond to connectivity with a flat hat referring to an inhibition connection and an arrow hat for excitatory connection. Right panel: Simulated temporal neural activity pattern. Red shaded strips correspond to a high firing rate in the neural population with the yellow bars indicating the colour input to the network model. (B). Tuning bias (the angular difference between the orientation where the tuning curve peaks and the actual cued orientation) computed from a population of spatially-selective PFC neurons from two rhesus monkeys. It shows significant positive tuning bias (more clockwise tunings for a cued clockwise location) during the delay period. (C). Top panel: Simulated tuning bias using a bump attractor model creates the neural tuning bias pattern. Bottom panel: Simulated behavioural change as a function of the tuning bias. Panel B and C copied from (Wimmer et al., 2014).

**Resource-rational model.** As mentioned previously, one key behavioural signature of VWM is that the encoding precision for each item decreases exponentially as the set size

increases. Most models assume a fixed amount of resource along with an ad hoc assumption by using a power-law to explain this set size effect. Van den Berg and Ma (2018) challenged these assumptions and suggested that a model which assumes that the observer can dynamically control the amount of resources required for the task can recover the power-law precision under different set sizes. In particular, a large body of work on perceptual decision making has focused on characterising a behavioural cost function for a corresponding task, i.e., what level of behavioural performance the observer tried to attain (e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff & McKoon, 2008). At the same time, theoretical and empirical neuroscience work has focused on the neural cost function i.e., the amount of spikes required to represent a stimulus, the ratio of mutual information, channel capacity etc. Van den Berg and Ma proposed a model in which the performance of a VWM task is governed by both the expected behaviour cost (the expected task error) and the expected neural cost (e.g., spike rate; although the model assumed that non-spiking components or other types of costs such as “mental effort” can also be part of the neural cost). Figure 2.11 illustrates the simulated set size effect. In this simulation, the authors assumed that the neural cost  $\bar{J}$  ranges between 0.01 to 20, and the behavioural cost is a product of two metrics – the predicted precision given the resource  $\bar{J}$  and a monotonic cost for estimating the target item badly.

First, the authors assumed that given neural resources  $\bar{J}$ , the predicted precision follows a von Mises probability density function that has a concentration parameter  $\kappa$  (proportional to  $\bar{J}$ ). This function peaks at the feature value of the target item. Evaluating the function at  $x$  indicates the probability of encoding the feature  $x$  from the feature space. The higher  $\bar{J}$  is, the higher the  $\kappa$  parameter, i.e., the sharper the von Mises function is. This implies that the representation error of the encoded item would not deviate much from the true feature value of the target item. On the contrary, under low  $\bar{J}$  i.e., a low  $\kappa$  parameter, the von Mises function is close to uniform, implying that there is as much chance that one would represent a different feature value compared to the target feature value of the encoded item. The second component of the behavioural cost assumes that

there is a monotonically increasing costs for having a mismatch between the encoded item and the true identity of the target item – simulated by a power-law:  $|x|^\beta$ , where  $x$  refers to the entire feature space, and the function should bottom at the feature value of the target item. Measuring the sum of total behavioural costs, the authors show that the sum decreases exponentially as extra units of neural resources become available. This results in a natural power-law set size effect (Figure 2.11). Their model showed success in capturing the set size effect as well as how item probing probability (when items are probed for retrieval with specific probabilities in a VWM task) influences behaviour across multiple studies. This study similarly suggests that humans are able to trade-off cognitive resource for representation fidelity like in the Nassar et al. study, but importantly they proposed that computational models do not have to assume a hard resource constraint, which is different from most conventional models. This study highlighted the importance of future studies to investigate how flexible the resource constraint is, and how the flexibility in resource constraint changes how we interpret suboptimalities.

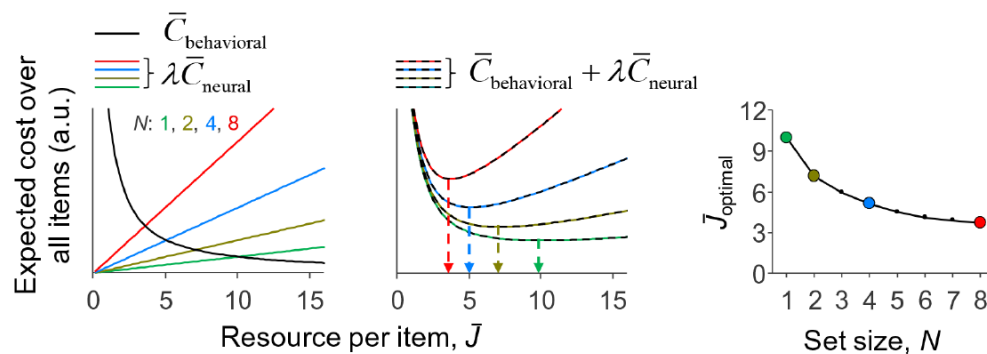


**Figure 2.11. Resource rational theory and power-law set size effect**

Figure copied from van den Berg & Ma, 2018. The Resource Rational theory provides an alternative normative account on the set-size effect in VWM. Unlike Nassar et al., this model assumed that observers could deliberately allocate both neural and behavioural resources and control the total resource required. Left panel: The simulated total behavioural cost and the neural cost ( $\lambda = 1$ ). The total behavioural cost falls exponentially as the number of resources for encoding each item increases linearly. Middle panel: combining the behavioural costs and neural costs results in a non-linear cost function. Right panel: Assuming that the observer is optimised for

the lowest total cost (y-axis), this simulation naturally recovers the power-law set size effect. In fact, the power-law set size effect in this model is not caused by introducing ad hoc parameters or functions. Any reasonable  $\beta$  and  $\bar{J}$  parameters (e.g.,  $\beta = 1$ , a linear estimation error cost) would still replicate the power-law set size effects.

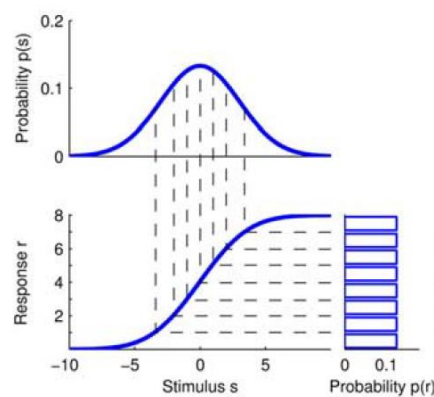
### 2.2.3. Adaptive Gain control

One major criticism of the efficient coding hypothesis is that it assumed that the neural system budgets its resources entirely based on the statistical likelihood of sensory information. However, Machens et al. (2005) showed that the sensory system prioritises behaviourally relevant stimuli in addition to optimising for the statistical regularities of the environment by studying grasshoppers' auditory neural responses. As illustrated in Figure 2.12, one form of efficient coding strategy by the neural system is to find an input-output function for each neuron that maps the stimulus distribution to a uniform response frequency profile. This can be done by taking the conditional cumulative probability of the neural responses  $r$  given the stimulus  $s$  as the mapping function. If auditory neurons are efficient, the stimulus distribution neurons were preconditioned on should give rise to an optimal mapping function. Conversely, the recovered stimulus distribution given the empirical response function should match with the actual stimulus distribution if neurons are truly efficient. This is what Machens and his colleagues tested. They presented grasshoppers sound snippets ensembles while controlling the mean and standard deviation of the ensembles. By reverse engineering the stimulus distributions neurons were preconditioned on, Machens and colleagues found that neural encoding does not follow the stimulus statistics of natural sounds. They further asked what sound snippets or ensembles could explain the neural encoding. By plotting out the sound segments that follow or do not follow a distribution that underlies the efficient neural response, they found that sound segments that mark sound transients were what the neural encoding is optimised for. These sound transients in fact have strong behavioural relevance for the

grasshoppers (communication purposes), showing that neural encoding is not only for prioritizing information prevalence, but also behaviourally relevant information.

**Figure 2.12. Efficient coding strategy.**

Figure copied from (Machens et al., 2005). Top panel: the stimulus probability in the environment  $p(s)$ . Bottom panel: An efficient coding neural system can adopt a coding strategy such that the neural responses have equal frequency (a uniform  $p(r)$ ). In this case, neurons which are coding for the most probable stimulus should become more sensitive. An example coding strategy would be to use a mapping function that is displayed in the bottom panel that is based on the conditional cumulative probability of the response weighted by the stimulus probability  $p(r|s)$ .



The adaptive gain control framework considers that a decision maker can have flexible control over the resource allocation policy in order to meet the contextual demands of an environment. The contextual demands often depend on the task. For simple perceptual tasks (e.g., deciding the orientation of the grating), the context can be determined by orientation frequencies in the world, covering the main prediction of the efficient coding hypothesis. In other tasks, the brain might consider other local contexts as the basis for gain control. For example, why would humans preferentially sample information that confirms their beliefs? Cheadle and colleagues showed that when humans are asked to integrate sequential sensory information, their decisions are more influenced by samples of information that are consistent with the previous sample (Cheadle et al., 2014). This can be explained by a simple model where the input-output mapping function

(transfer function that takes sensory information as input and gives out decision information) adapts such that its inflection point locates at the sample the observer just saw. This results in an increased gain for a subsequent sample that is close to the inflection point. This mechanism explains why consistent samples wield more influence on choice. This aligns well with the commonly observed heuristics where humans tend to select information that confirms their belief--“confirmation bias”. The implication for this goes beyond the case of passive sampling (where information reception is not controlled by the observer), Vandormael and colleagues showed that when subjects are placed in a task where they can freely sample information, they preferentially sample i) information that confirms their running hypothesis of the task, ii) information that is consistent with the previous sample (Vandormael, 2017), and iii) information that is statically most likely on a given trial (Vandormael, Castañón, Balaguer, Li, & Summerfield, 2017).

Moving beyond the case of sequential decisions, the adaptive gain model is able to explain different biases in summary statistics (representations of an ensemble of features by its average). For example, when humans are asked to make a summary judgment on multiple visual elements (e.g., is the average colour of multiple shapes more red or more blue), human observers display a bias in which inlying elements are relatively upweighted compared to outlying items where inlying items refer to items close to the mean feature value of the array--“Robust averaging” (de Gardelle & Summerfield, 2011). In Chapter 3, I will discuss in detail why robust averaging is an adaptive strategy in the presence of decision noise using the adaptive gain mechanism. In particular, I will show that robust averaging is caused by humans taking into account the entire stimulus distribution (global contexts) in a way predicted by the adaptive gain model.

In another situation, human summary judgement receives a performance benefit if the target array is preceded by a multi-element array (a prime) that shares statistical properties with the target array (e.g., if the variance of the prime is similar to the variance of the target array)--“variance priming” (Michael, de Gardelle, & Summerfield, 2014). Both effects can be explained by an adaptive gain control mechanism where the multiplicative

gain is allocated to features that are consistent with the local context. In the robust averaging case, the local context is the distribution from which the feature values are sampled, which is a normal distribution with a pre-specified mean and variance. Therefore, allocating gain (resulting in higher decision weights) to inlying items is optimal for the task. In the variance priming case, the local context is set by the distribution of the prime array. A model which assumed that the statistical distribution of the prime array sets the gain field for processing individual features successfully explained the behavioural performance observed in humans (Michael et al., In prep.). In Chapter 4, I will show how an adaptive gain mechanism that allocates gain to local context can explain why irrelevant information influences human decisions across the perceptual, cognitive, and economic domains. Lastly, in Chapter 5, I am going to compare the adaptive gain model against other efficient algorithms – neural computations that have shown success in explaining context-dependent decisions in the literature.

## Chapter 3 - Robust averaging protects decisions from noise in neural computations

### 3.1. Introduction

Decisions about the visual world often require observers to integrate information from multiple sources. An ideal observer will give each source a weight that is proportional to its reliability. Thus, where all sources are equally trustworthy, the best policy is simply to average the available features or decision information. For example, a decision about which fruit to buy at the supermarket might involve averaging the estimated size and colour of the produce, or a wager about which football team will win might be made after averaging the speed and skill of all the players on a team (Wald & Wolfowitz, 1949).

Previous studies have investigated how humans average perceptual information by presenting participants with array composed of multiple visual elements and asking them to report the mean size, colour or shape of the items displayed (Ariely, 2001; Chong & Treisman, 2003, 2005; De Fockert & Marchant, 2008; Solomon, Morgan, & Chubb, 2011). Interestingly, recent reports suggest that human averaging judgments do not resemble those of an ideal observer (de Gardelle & Summerfield, 2011; Haberman & Whitney, 2010; Michael, de Gardelle, Nevado-Holgado, & Summerfield, 2013; Michael et al., 2014). Rather, when averaging, humans tend to downweight or discount visual features that are unusual or outlying with respect to the distribution of features occurring over recent trials (“robust averaging”). Haberman and Whitney first showed that observers discount emotional deviants when averaging the expression in human faces (Haberman & Whitney, 2010). Subsequently, de Gardelle, Summerfield and colleagues provided evidence that observers discount outlying colour or shape values during averaging of features in a multi-element array (de Gardelle & Summerfield, 2011; Michael et al., 2013). Control analyses ruled out the possibility that the observed effect was an artefact of hardwired nonlinearities in feature space. Together, these studies suggest that humans are “robust averagers”, overweighting inliers relative to outliers

rather than giving equal weight to all elements (although see (Gorea, Belkoura, & Solomon, 2014) for a failure to replicate this finding using a 2-alternative forced choice averaging task).

According to a widely-accepted framework with its roots in Bayesian decision theory (Gold & Shadlen, 2001; Wald & Wolfowitz, 1949), robust averaging is suboptimal. Intuitively, robust averaging discards information about the stimulus array, and should thus reduce performance relative to a policy that integrates the stimulus feature values evenly. Why, then, do humans give more weight to inliers than outliers during integration of decision information? Here, we tackled this question using psychophysical testing of human observers and computational simulation. We asked participants to average the orientation (tilt) in a circular array of gratings, relative to a central reference grating that either *(i)* remained the same or *(ii)* varied in a trial-wise fashion over a block of trials. This latter manipulation allowed us to test whether robust averaging is still observed even when the distribution of sensory information is uniform around the circle and varies randomly from trial. Using this approach, we show that human robust averaging can be conceived of as a policy that rapidly allocates limited resources (gain; see Equation 3-2 below) to items that are closest to the category boundary (or indifference point). Although this policy is suboptimal in the absence of noise, it has a surprising protective effect on decisions that are corrupted by “late” noise arising during or beyond information integration.

We begin by describing the behaviour of a cohort of human observers performing the orientation averaging task. Next, we describe a simple psychophysical model in which feature values (tilt, relative to a reference value) are transformed nonlinearly before being averaged to form a decision variable. This variable is corrupted with “late” (post-averaging) noise and then used to determine model choices. This model accounts better for human behaviour (including observed robust averaging) than a rival account, based on an ideal observer, that replaces the initial nonlinear step with a purely linear multiplicative transformation. Next, we use simulations to explore the properties of this model. We show that as we increase late noise, a model that engages in robust averaging comes to outperform the linear model, i.e., achieves higher simulated choice accuracy. Finally, we return to the human data, and show

that for both model and humans, the use of a robust averaging strategy is a positive predictor of decision accuracy, in particular under high estimated late noise.

## 3.2. Methods

### 3.2.1. Design and behavioural analysis

**Participants.** 24 healthy human observers (9 males, 15 females; age  $23.4 \pm 4.7$ ) participated in two testing sessions that occurred one week apart. The order of testing sessions was counterbalanced across participants. The task was performed whilst seated comfortably in front of a computer monitor in a darkened room. Participants received £25 in compensation.

**Task.** All stimuli were created using the Raphaël JavaScript library and presented with the web browser – Chrome Version 49.0.2623.87 on desktop PC computers. The monitor screen refresh rate was 60Hz. Each session consisted of 8 blocks of 128 trials each. On each trial, following a fixation cross of 1000ms duration, participants viewed an array of 8 square-wave gratings with random phase (2.33 cycles/degree, 0.33 RMS contrast, 1.72 degrees visual angle per grating) arranged in a ring 7.82 degrees from the centre of the screen (Figure 3.1). The array was presented for a fixed duration against a grey background in each block (250ms, 500ms, 750ms or 1000ms; this manipulation had little impact on accuracy, and we collapsed across it for all analyses; Figure 3.2). A single Gabor patch was presented in the centre of the ring contiguous with the array elements (3.49 cycles/degree, 0.33 RMS contrast, 1.15 degrees visual angle). Participants were asked to judge as rapidly and accurately as possible whether the mean orientation of the array of 8 peripheral gratings fell clockwise (CW) or counterclockwise (CCW) of the orientation of the central grating. Feedback was provided immediately following each response: the fixation cross turned green on correct trials for 500ms, and red on incorrect trials for 2500ms. Participants received instructions and completed a training block of 32 trials prior to commencing each session. During the training

block, the central grating patch and the array of grating patches remained on the screen for 1 minute or until participants made a response.

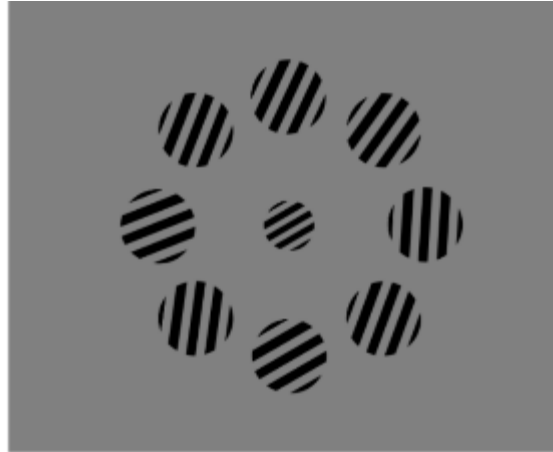


**Figure 3.1. Schematic demonstration of the stimulus array**

The task was to report whether the average orientation of the outer ring of gratings fell clockwise or counterclockwise of the orientation of the central (reference) grating.

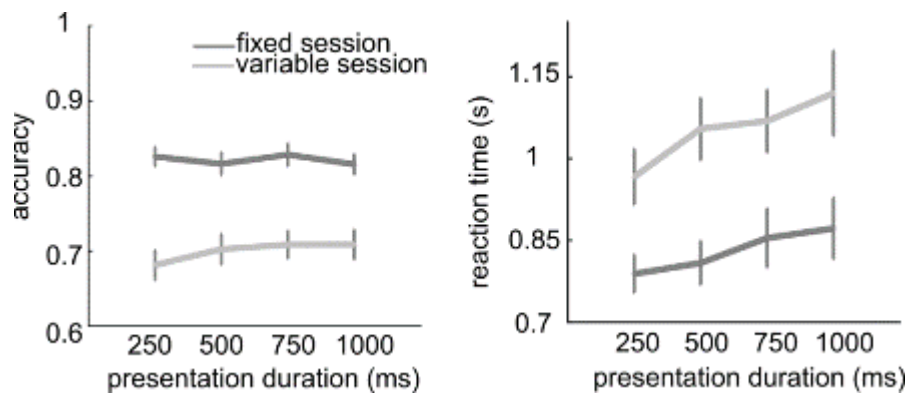


**Figure 3.2. Performance under different presentation duration conditions**

The mean and standard error of the mean for  $|\mu|$  on accuracy (left panel) and reaction times (right panel) under different presentation durations (x-axis) in fixed (dark grey line) and variable reference session (light grey line).

**Design.** Orientations were sampled from Gaussian distributions with means of  $R \pm \mu$  where  $R$  is the reference grating orientation, and variances of  $\sigma^2$  on each trial. We crossed  $\mu$  and  $\sigma$  as orthogonal factors in the design, drawing the orientation mean (in degrees) from  $\mu \in -20^\circ, -10^\circ, 10^\circ, 20^\circ$  and orientation standard deviation  $\sigma \in 8, 16$ . Levels of  $\mu$  and  $\sigma$  are

counterbalanced and the order of presentation is randomised across trials in every block. To ensure that the sampled orientations matched the expected distribution with the given  $\mu$  and  $\sigma$  resampling of orientation values occurred until the mean and standard deviation of orientation values fell within  $1^\circ$  tolerance of the desired  $\mu$  and  $\sigma$ . We refer to each of the 8 gratings in the array as a “sample” of feature values. Reference orientations were drawn randomly and uniformly from around the circle. There was a total of 8 blocks per session, leading to a total of 1024 trials per session. In the fixed-reference session, the reference orientation remained fixed over each block of 128 trials. In the variable-reference session, the reference orientation changed from trial to trial. Our experiment thus had a 2 (fixed vs. variable reference)  $\times$  2 ( $\mu = 10, \mu = 20$ )  $\times$  2 ( $\sigma = 8, \sigma = 16$ ) factorial design.

**Analysis.** 3 subjects were excluded from all analyses due to lower than 60% accuracy performance in either of the reference condition. Data were analysed using ANOVA and regressions at the between-subjects (group) level. A threshold of  $p < 0.05$  was imposed for all analyses, and we used a Greenhouse-Geisser correction for sphericity where appropriate, so that some degrees of freedom (d.f.) are no longer integers. We first compared accuracy and reaction times for different levels of  $\mu$  and  $\sigma$  in each session. Next, we used probit regression to estimate the weight with which each sample influenced choices, as a function of its position relative to the reference angle in both fixed and variable reference session. For all analyses, we excluded 13% of trials (‘wraparound’ trials) that contained one or more orientations that were  $>0.79^{rad}$  or  $<0.79^{rad}$  (equivalent to  $>45^\circ$  or  $<-45^\circ$ ) relative to the reference, thereby ensuring that we were working within a space in which feature values  $X$  were approximately linearly related to angle of orientation. In the current task, a feature value of  $90^\circ$  relative to the reference orientation carries zero decision information, meaning that in a discrimination task where orientation is judged relative to vertical ( $0^\circ$ ), the decision information of a grating orientation of  $+90^\circ$  and a grating orientation of  $-90^\circ$  would be identical. The decision information of the current task grows up to  $45^\circ$  and then begins to shrink, when it reaches  $\pm 90^\circ$ , the signal is completely ambiguous about whether the orientation is clockwise or counterclockwise to the reference, so we chose  $45^\circ$  as the cut-off

criteria for trial exclusion. A further 0.2% of trials on which no response was registered were also excluded.

For each sample  $i$  trial  $t$ , we assumed that orientations in the sensory space were being recoded as orientations relative to reference in the decision space, and thus refer to the feature values  $X$  as the orientation relative to the reference. After excluding ‘wraparound’ orientations, all orientations fell within the range of  $-0.79^{rad}$  to  $0.79^{rad}$  (equivalent to  $\pm 45^\circ$ ). To compute weighting functions, we created for each participant a predictor matrix by tallying values of  $X$  within each of 8 equally spaced bins (in feature space) with centres between  $-0.75^{rad}$  and  $0.75^{rad}$  on a trial-by-trial basis. Values from each bin were entered competitive regressors to regressed against participants’ choices using probit regression.

### 3.2.2. Modelling

**Power model.** Each element  $i$  was characterised by a feature value  $X_i$  in radians (in the range  $-0.79^{rad}$  to  $0.79^{rad}$ ) that was proportional to its orientation relative to the reference. Our model assumes that the decision value ( $DV$ ) that determined choice on each trial was computed by transforming orientations relative to reference using a power-law transducer parameterised by an exponent  $k$ .

$$DV = \sum_{i=1}^8 \text{sign}(X_i) \cdot |X_i|^k$$

Equation 3-1

The functions that map feature value  $X$  onto decision values  $DV$  for low and high values of  $k$ . For the special case  $k = 1$ , the  $DV$  is equivalent to the simple sum of  $X_i$ ; this is the rule used by the experimenter to determine feedback. Next, we calculated choice probabilities by passing the  $DV$  through a sigmoidal choice function (see choice probability function and Equation 3-11) with the inverse-slope  $s$ . Higher values of  $s$  imply shallower slopes and thus greater “late” noise. The sign of sum of  $X_i$  always reflect the sign of the mean of the distribution in which  $X_i$  was being drawn from, which we used for providing feedback.

**The associated equivalent gain factor for each exponent  $k$ .** Different levels of the exponent  $k$  vary the convexity or the concavity of the transfer functions. By considering the integral of the absolute of these functions, it is easy to see that  $k$  in turn varies the overall scaling of any hypothetically occurring feature values onto  $DV$ . We assume that this overall scaling corresponds to the amount of “resources” participants have. When  $k < 1$ , average (absolute) values of  $DV$  are inflated, and thus pushed away from the category boundary, increasing simulated performance. We wished to ensure that model comparisons cannot be trivially explained by this unequal scaling of feature values to decision variable under different levels of  $k$  – that is, any performance benefits or losses we observed in humans is being attributed to a nonlinear weighting strategy rather than a change in scaling or in the amount of resources. To correct for the scaling, we thus computed the equivalent gain factor ( $g$ ) that quantifies the average increase in absolute  $DV$  under different levels of  $k$ :

$$g = \frac{2}{1 + k}$$

Equation 3-2

The quantity  $g$  is equal to  $\frac{\sum F^k}{\sum F}$  where  $F$  is a hypothetical space of features (here, positive only for convenience) that could occur in the experiment. This is equivalent to assuming an agent with no existing prior regarding the distribution of the features. Figure 3.3 illustrates the mapping between  $k$  and  $g$ . Multiplying equivalent linear models (the next section) by  $g$  thus corrects for the inflation that would occur under differing values of  $k$ . This is equating the amount of inflation or resources under both linear and nonlinear weighting schemes, so that comparisons on the model would only restrict to the nature of the weight scheme. We implemented this correction when comparing equivalent linear and nonlinear models with parameter  $k$ , either by multiplying the input features of the linear model by  $g$ , or equivalently, by dividing the output of the nonlinear model by  $g$ . Importantly, this correction was applied over the features that could occur, not the features that did occur under our mixture of Gaussian-distributed categories. It is for this reason that the nonlinear model leads to improved predicted performance in the experiment we conducted, but not in

a simulated experiment in which features were uniformly drawn from across feature space (see results).

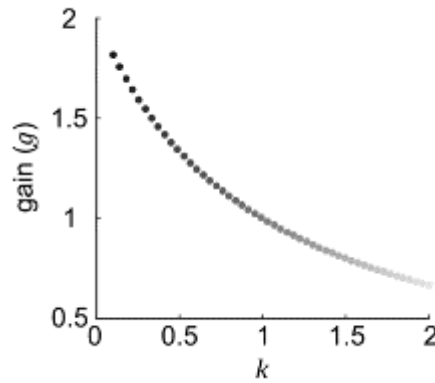


**Figure 3.3. exponent  $k$  and gain ( $g$ )**

Lower values  $k$  (darker dots) have higher multiplicative gain, therefore the corresponding  $g$  is higher for low value of  $k$

**Equivalent gain linear model.** For each nonlinear model variant  $k$  in the power model, we compute  $DV$  using a linear model with equivalent gain factor, i.e., a model with the following form:

$$DV_{linear} = \sum_{i=1}^8 X_i \cdot g$$

Equation 3-3

Where  $DV$  linear refers to the cumulative decision value of all feature value  $X_i$  after applied with equivalent gain –  $g$ . This ensures that each nonlinear power model is compared to a linear model with an equivalent total input-to-output scaling of decision values. Using this approach, we could thus compare the benefits of allocating gain preferentially to inliers ( $k < 1$ ) or outliers ( $k > 1$ ) to allocating gain evenly across feature space ( $k = 1$ ), under the assumption that neural resources were limited to a fixed value defined by  $g$ , for example, the total number of spikes across population of neurons sensitive to orientations. Mind here that in this model, late noise  $s$  trade-off with  $g$ . The reason we include both parameters in this model is because the two parameters serve completely different theoretical functions, one

corrupts the post averaging estimate of decision values, the other serves as a gain normaliser to ensure input-output scaling under different variants of  $k$  remained constant. Therefore, even seemingly  $s$  and  $g$  are trading off each other, it is essential to keep the two parameters separate so that it allows fair comparisons between the linear and non-linear model under each given level of late noise ( $s$ ).

The model comparison of power model against the equivalent gain model is mathematically identical to comparing model performance for  $k < 1$  or  $k > 1$  against  $k = 1$  of a power model which is normalised by  $g$  in this form:

$$DV_{constant} = \frac{DV}{g}$$

Equation 3-4

Where  $DV_{constant}$  refers to the decision variable with constant gain across different levels of  $k$ . Under a  $k < 1$  case, inlying items will be allocated with more resources at the expense of depriving resources from outlying items, while under a  $k > 1$  case, outlying items will be allocated with more resources at the expense of inlying items. Any difference in simulated model performance of nonlinear transformation of feature values across different values of  $k$  are not due to differential resources in a linear model.

**Weighted sum power model.** A mathematically equivalent way of devising the power model is to formulate Equation 3-1 as the weighted sum of feature values:

$$DV = \sum_{i=1}^8 X_i \cdot |X_i|^{k-1}$$

Equation 3-5

Where  $|X_i|^{k-1}$  can be seen as the weights correspond to each  $X_i$  as a function of parameter  $k$ . This formulation would yield an equivalent non-linear transformation of feature values to decision values as the power model. As mentioned previously regarding the unequal total input-output scaling under different exponent  $k$ , we also implement normalisation on

this model. We first compute the weight associated with each feature value  $\omega_F = F^{k-1}$ , where  $F = \{0,1\}$ , possible unsigned features in the feature space. Similar to the computation of the equivalent gain factor, the normalisation of the weights was conducted over the features that could occur in this experiment (i.e., it does not take into account the Gaussianity of the feature distribution). Next, we normalise  $\omega$  so that the weights for all possible feature values are summed to 1:

$$\omega = \frac{\omega_F}{\sum_{i=1}^N \omega_F^i}$$

Equation 3-6

This means that when  $k < 1$ , having greater weight (or more resources) for inliers necessarily means outliers will now have lesser weight (deprived of resources) due to normalisation. Conversely, when for the  $k > 1$  case, having more resources for the outliers means inliers will now have lesser weight. Figure 3.4 shows the normalised weights ( $\omega$ ) and the resulting  $DV$  as a function of different  $k$  with reddish coloured lines corresponding to low  $k$  values and bluish coloured lines corresponding to high  $k$  values. This form of normalisation, although is mathematically different to normalising through the Equivalent gain factor described in Equation 3-2, works similarly to that which we have already described. The quantity  $\sum_{i=1}^N \omega_f^i$ , scales monotonically with  $g$ .

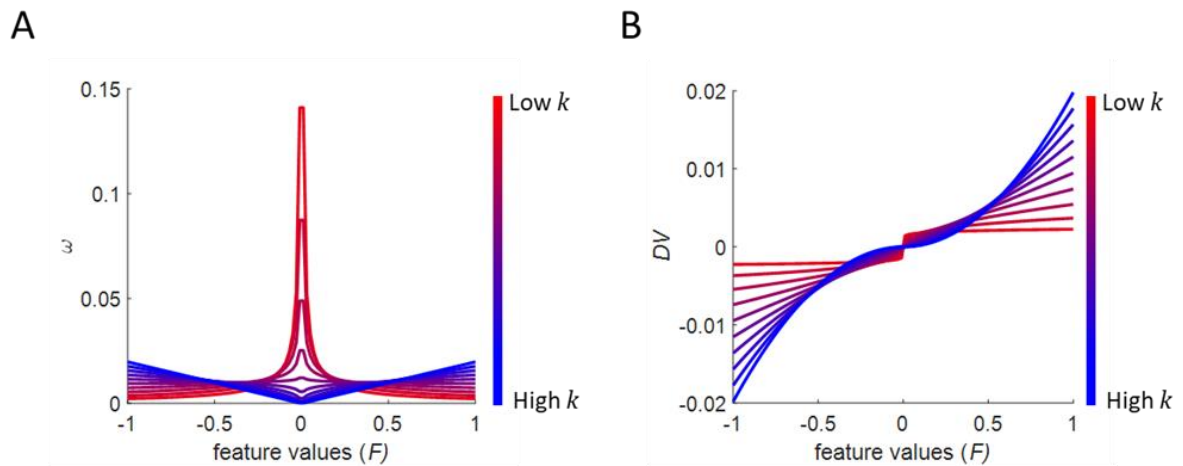


**Figure 3.4. Weighted sum power model**

(A) Normalised weights under different  $k$  (B) Transfer functions of the weighted sum power model. Decision values ( $DV$ ) were computed by taking the product of the associated normalised weights with feature value  $F$ . It shows that under  $k < 1$ , features close to the category boundary gets relatively upweighted, but at the expense of relatively downweighting the outlying features. Conversely, under  $k > 1$ , outlying features are upweighted but inlying features are downweighted.

**Early noise only model.** To test our assumption that early sensory noise (noise arise prior to averaging) alone cannot explain subjects' choice behaviour, we created a model where each feature value  $X_i$  was corrupted by  $\epsilon$ , a sample of noise drawn independently from  $N(0, \xi)$ , a Gaussian distribution zero mean and standard deviation  $\xi$  :

$$x_i = X_i + \epsilon$$

Equation 3-7

After transforming  $x$  with exponent  $k$  using Equation 3-1, we converted the summed of  $x$  values into a choice probability of 0 or 1 depending of its sign (i.e., via a step function) on a trial-by-trial basis. We fit this model to psychometric functions, by computing the conditional probability of a clockwise response  $p(CW)$  given the presence of a feature  $X_i$  (sorted in to 8 equally spaced bins between  $-0.75^{rad}$  to  $0.75^{rad}$ ). We did this separately for the fixed reference session and variable reference session in humans. Using a grid search method, we identified best-fitting for  $\xi$  among 20 linearly spaced values from 0 to 3 for each subject and reference condition (fixed, variable) by minimising the MSE between the predicted and observed psychometric functions. This model has an equivalent number of free parameters as the power model which has a late noise parameter ( $s$ ) on top of the exponent  $k$ . Having identified the best-fitting parameters, we used these to predict accuracy for each level of mean and variance, and the weighting function in the fixed and variable reference conditions.

**Population coding power model.** As with the power model, we assume that feature values were recoded from presented orientations relative to the reference into a linear space

spanning between  $-3^{rad}$  and  $3^{rad}$  where 0 is the value of the reference (roughly spanning the entire circular space). We assumed theoretical neurons coding for a wider range of orientation values to avoid the truncation of tuning curves of neurons for outlying orientations presented in the experiment. Any truncation of tuning curves would distort the transformed feature values in a non-linear way that is independent of the values of  $k$ . We assumed a population of 600 neurons ( $M = 600$ ) whose tuning curves are linearly spaced across the feature space. The tuning curve for any neuron,  $j$ , is defined as a Gaussian probability density function centred at the neuron's preferred feature value,  $f_j$ , and with a tuning width fixed across the population,  $\epsilon$ , specified by an additional free parameter. The amplitude of each neuron's tuning curve (i.e., its maximum firing rate) was controlled by a gain factor which is a function of the neuron's preferred feature value,  $f_j$ , and the power law:

$$G_j = |f_j|^{k-1}$$

Equation 3-8

Where  $G_j$  represents the gain,  $G$ , applied to neuron,  $j$ , whose preferred feature value is  $f_j$ , and a free parameter,  $k$ , controls the gain applied across the feature space in the neural population. The firing rate,  $R_{ji}$ , for each neuron  $j$  given a particular stimulus,  $X_i$ , is computed as:

$$R_{ji} = N(X_i, f_j, \epsilon) \cdot G_j \cdot \frac{\rho}{M}$$

Equation 3-9

Where  $N(X_i, f_j, \epsilon)$  correspond to the probability density of a Gaussian with mean,  $f_j$ , and variance,  $\epsilon$ , evaluated at point,  $X_i$ . To adjust for the scaling of output values, the product of the Gaussian density function and gain function is additionally scaled by  $\frac{\rho}{M}$ , which is the ratio of range of the linear space in radians ( $\rho$ ;  $\rho = 6$  for a space between  $-3^{rad}$  to  $3^{rad}$ ) to the number of neurons ( $M$ ). This ensures that the output of the population activity  $R$  will

remained invariant to these factors of no interest in our model. Lastly, the model's estimate of a stimulus,  $X_i$ , is a scalar computed using a linear decoder:

$$\theta_i = \sum_{j=1}^{600} R_{ji} f_j$$

Equation 3-10

Where  $R$  is the population activity vector for  $X_i$ . Firing rate ( $R_{ji}$ ) of each neuron  $j$  is weighted by the corresponding neuron's preferred feature value ( $f_j$ ) before summation to get the model estimate for stimulus ( $\theta_i$ ). This is then used for computing the cumulative decision values (summation of model estimated angles) on a trial by trial basis for computing choice probability using Equation 3-12 and negative log-likelihood for model fitting.

**Choice probability function.** A choice function with a noise-term  $s$  was used to transform  $DV$  of each model into choice probabilities. These choice probabilities are then used for maximum likelihood estimation. We used a sigmoid choice function of the following form:

$$CP = \frac{1}{1 + e^{\frac{-DV}{s}}}$$

Equation 3-11

We ensured via visual inspection that the resulting fits were convex over this search space. We then used parametric tests to assess whether the resulting best-fitting parameters differed positively (indicating upweighting of outliers) or negatively (indicating downweighting of outliers) from 1. For each participant, we searched exhaustively over values of  $k$  (in the range 0.02 to 2) and  $s$  (in the range 0.05 to 10) that minimised the negative log likelihood of the model (Equation 3-12).

**Maximum likelihood estimation.** Model fitting on the choice data (both human choice and synthetic choice) was performed using the maximum likelihood estimation. We first computed the probability that each human choice comes from a set of model parameters

$(k, s)$  given a model of interest (termed as a "model" for brevity in this section) (Equation 3-12). This is equivalent to the model predicted choice probability computed using the sigmoidal choice function described above, where  $Y = \{y_1, y_2, \dots, y_n\}$  corresponds to the trial-by-trial model prediction on  $P(CW)$ . Next we compute the likelihood of the human choice data  $Y$  given the model i.e.,  $P(Y|model)$  by taking the product of model predicted choice probability from all trials  $n$ . For the mathematical convenience, we convert the likelihood into log likelihood (Equation 3-12). The model (under a specific parameterisation) with the highest likelihood (i.e., summed log likelihood) or the lowest negative sum of log likelihood is the best model i.e., the best fitting parameters given the choice data.

$$\begin{aligned}
 P(Y|model) &= P(y_1, y_2, \dots, y_n|model); \\
 P(Y|model) &= P(y_1|model) \cdot P(y_2|model) \cdot \dots \cdot P(y_n|model) = \prod_{i=1}^n P(y_i|model) \\
 \log P(Y|model) &= \log \prod_{i=1}^n P(y_i|model) = \sum_{i=1}^n \log P(y_i|model)
 \end{aligned}$$

Equation 3-12

**Parameter recovery.** To test the ability of the fitting procedure to accurately identify the parameters of the best-fitting power model. We sampled 20 equally-spaced values of  $k$  (in the range of 0.02 to 2) and  $s$  (in the range of 0.05 to 10). For each  $k$  and  $s$  combination, we transformed a set of orientations presented to subjects in the experiment using the given  $k$  and computed the choice probability of the  $DV$  with the given  $s$ . Then we compared the trial-to-trial estimated choice probability against a random probability drawn from a uniform distribution with a range of 0 to 1 to generate model choices. We then used these artificial choices to recover best-fitting values of  $k$  and  $s$  via maximum likelihood estimation.

**Model performance simulation.** We simulated model performance (decision accuracy) under different  $k$  in a range of 0.02 to 2 and  $s$  in a range of 0.05 to 5 for the power model. For each combination of  $k$  and  $s$ , trial-to-trial estimate of  $DV$  was computed and transformed into choice probability using Equation 3-11. Model choices were created by

comparing the choice probability against a probability drawn randomly from a uniform distribution. Model accuracy was computed as the proportion of model choices that were the same as the pre-defined correct choice, which is simply determined by the sign of the sum of  $X$ .

### 3.3. Results

Human participants ( $N = 24$ ) took part in two psychophysical testing sessions separated by approximately one week. On each of 2048 trials, they viewed an array of 8 high-contrast gratings presented in a ring around a single central (reference) grating (Figure 3.1). The grating orientations were drawn from a single Gaussian distribution with mean  $\mu \in -20^\circ, -10^\circ, 10^\circ, 20^\circ$  and standard deviation  $\sigma \in 8, 16$  relative to the reference. Their task was to report whether the average orientation in the array was clockwise (CW) or counterclockwise (CCW) of the central grating. The reference grating was drawn uniformly and randomly from around the circle, and varied on either a trial-by-trial (variable reference) or block-by-block (fixed reference) fashion. Fixed and variable reference conditions occurred in different sessions whose order was counterbalanced over participants. Fully informative feedback was administered on every trial.

#### 3.3.1. Human behaviour

Mean accuracy and standard errors of mean (S.E.M.) for the human participants (lines) are shown in Figure 3.5. Participants responded more slowly when the orientation mean approached the reference (main effect of  $\sigma$ :  $F_{1,20} = 47.14$ ,  $p < 0.0001$ ) and when the orientation variance increased (main effect of  $\sigma$ :  $F_{1,20} = 6.84$ ,  $p = 0.017$ ). They also made more errors for lower values of  $|\mu|$  ( $F_{1,20} = 397.1$ ,  $p < 0.0001$ ) and higher values of  $\sigma$  ( $F_{1,20} = 116.1$ ,  $p < 0.0001$ ). Directly comparing the low  $|\mu|$  low  $\sigma$  condition ('low-low') to the high  $|\mu|$  high  $\sigma$  condition ('high-high'), participants made more errors and are slower under high-high condition (accuracy:  $F_{1,20} = 48.53$ ,  $p < 0.001$ ; RT:  $F_{1,20} = 20.67$ ,  $p < 0.001$ ) even though the

$|\mu|$  to  $\sigma$  ratio is identical in these two conditions. This result replicates previous findings (de Gardelle & Summerfield, 2011).

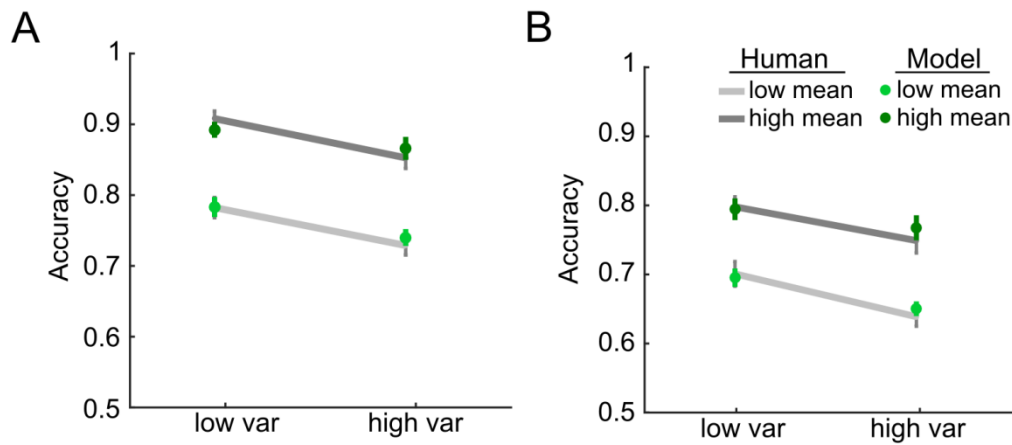


**Figure 3.5. Model and human data**

Mean accuracy and the standard error of mean of human (grey lines) and model (green dots) for high and low variance conditions, with low mean (i.e., orientation close to the reference; light grey lines) and high mean (dark grey lines). Panel A shows performance in the fixed reference session, and the panel B shows the variable reference condition.

As expected, participants were overall faster ( $F_{1,20} = 64.4, p < 0.0001$ ) and more accurate ( $F_{1,20} = 89.95, p < 0.0001$ ) in the fixed reference than variable reference condition. An interaction between mean and session was observed for both RT ( $F_{1,20} = 9.63, p < 0.001$ ) and accuracy ( $F_{1,20} = 5.83, p = 0.025$ ) indicated that the cost incurred by lower values of  $\mu$  was greater under the fixed than variable reference condition. No interactions between session and feature variance were observed. There was a significant interaction for both accuracy ( $F_{1,20} = 4.18, p = 0.041$ ) and RT ( $F_{1,20} = 8.06, p = 0.01$ ) with sessions for the low-low and the high-high condition, showing that the relative performance cost for the high-high condition was lower under the variable reference condition. These findings indicate that our manipulation of fixed vs. variable reference successfully influenced human categorisation performance, and that  $\mu$  and  $\sigma$  have comparable impact on accuracy and RT to that described in previous studies. The same results were obtained when this analysis was carried out on  $d'$  rather than % correct values (see Figure 3.6, and Appendix Table 3.1).

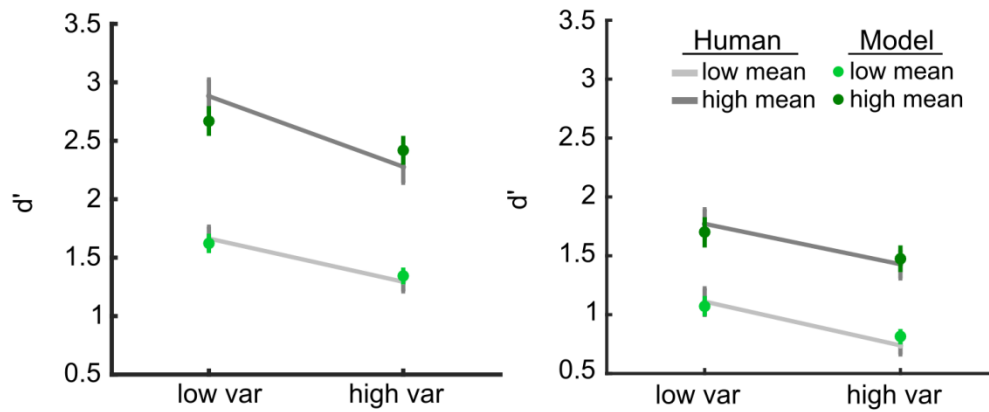


**Figure 3.6.  $d'$  analysis**

$d'$  for each level of  $|\mu|$  (mean) and  $|\sigma|$  (variance) conditions were computed separately for fixed reference (Left panel) and variable reference session (Right panel). The grey lines correspond to human's average  $d'$  for low mean (light grey) and high mean conditions (dark grey). The green dots correspond to the model fits for each condition (low mean in light green dots and high mean in dark green dots).

Next, to probe for robust averaging, we measured the influence that each feature carried on the decision, as a function of its angle relative to the reference (see methods). Figure 3.7A shows the average regression coefficient (weight) associated with each of 8 bins of the feature values (i.e., orientations relative to reference) for the session with fixed reference (red line) and the session with variable reference (green line). The shaded area shows the standard error of the mean across observers. We first compared the coefficients with a factorial ANOVA, crossing the factors of session (fixed vs. variable reference) and bin. Consistent with the accuracy data above, this yielded a main effect of session ( $F_{1,20} = 59.54$ ,  $p < 0.001$ ). However, there was also a main effect of bin ( $F_{2.02,40.37} = 6.23$ ,  $p = 0.004$ ) with no interaction between these factors ( $p = 0.31$ ). Next, for each session, we directly compared the weights associated with (i) the four inlying bins (bin 3, 4, 5, 6] and (ii) the four outlying bins (bin 1, 2, 7, 8]. In both sessions, participants gave more weight to those samples falling in inlying than outlying bins (fixed reference:  $t_{20} = 7.8$ ,  $p < 0.0001$ ; variable reference:  $t_{20} =$

6.3,  $p < 0.0001$ ). In other words, under both fixed and variable reference, participants displayed a pattern of behaviour consistent with a “robust averaging” policy for orientation.

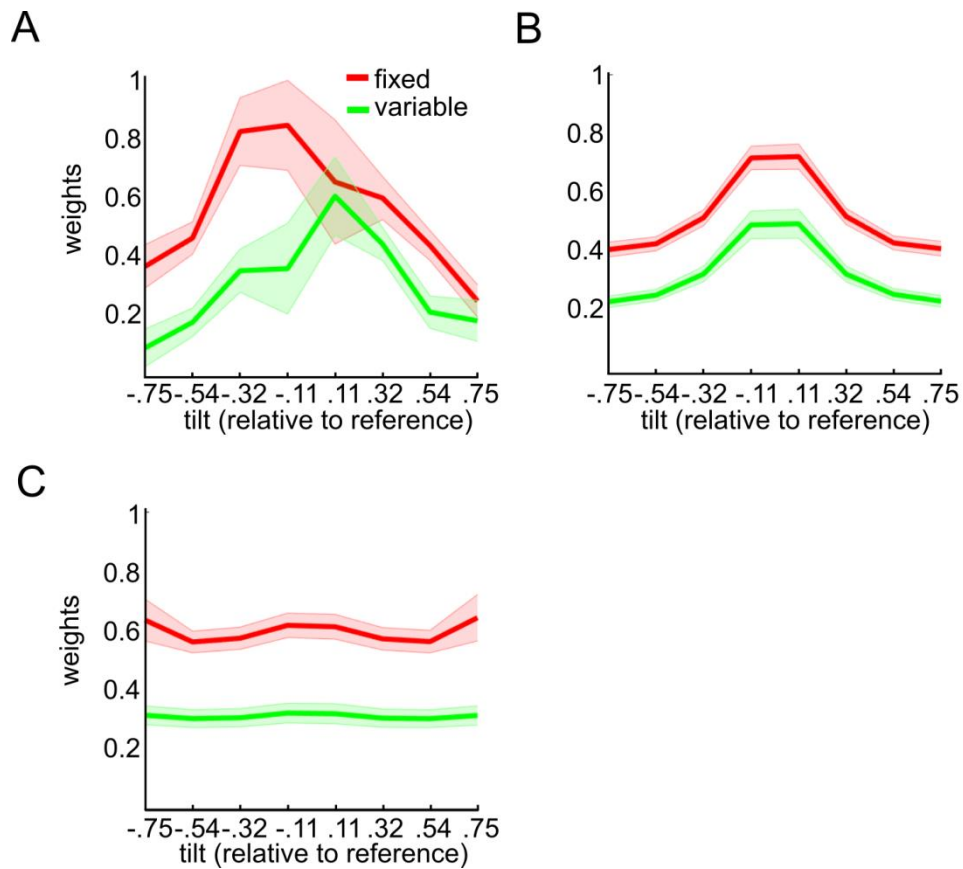


**Figure 3.7. Parameter estimates of orientation of each grating relative to the reference**

The y-axis shows parameter estimates for a probit regression in which the angles of orientation of each grating (relative to the reference) were used to predict choice. Angles were tallied into 8 bins, from most negative to most positive relative to the reference, so that each parameter estimate shows the relative weight given to a particular portion of feature space. The x-axis shows the bin centre of each bin. The inverted-U shape of the curve is a signature of robust averaging. Shaded areas are the standard error of mean. (A) Weighting functions estimated using human choices (B) Weighting functions for recreated model choices using the best fitting parameters from the power model using the best fitting parameters from human data. (C) Weighting functions for simulated model choice under a case in which angles are linearly mapped onto  $d\nu$ .

### 3.3.2. Model fitting.

We fit our data with a simple psychophysical model (power model; see methods). Each array element  $i$  was characterised by a feature value  $X_i$  that was proportional to its

orientation, recoded to be relative to the reference (in radians, i.e., in the range  $-0.79^{rad}$  to  $0.79^{rad}$  corresponding to  $-45^\circ$  to  $+45^\circ$ ). The model computes a decision value ( $DV$ ) by transforming  $X$  with a nonlinear function parameterised by an exponent  $k$ , and summing the resulting values:

$$DV = \sum_{i=1}^8 \text{sign}(X_i) \cdot |X_i|^k$$

The functions mapping  $X$  onto  $DV$  under different levels of  $k$  (red to blue lines respectively) are shown in Figure 3.8A. For the special case  $k = 1$ , the transfer function is linear, and  $DV$  is equivalent to the simple sum of  $X_i$ ; this is the rule used by the experimenter to determine feedback.

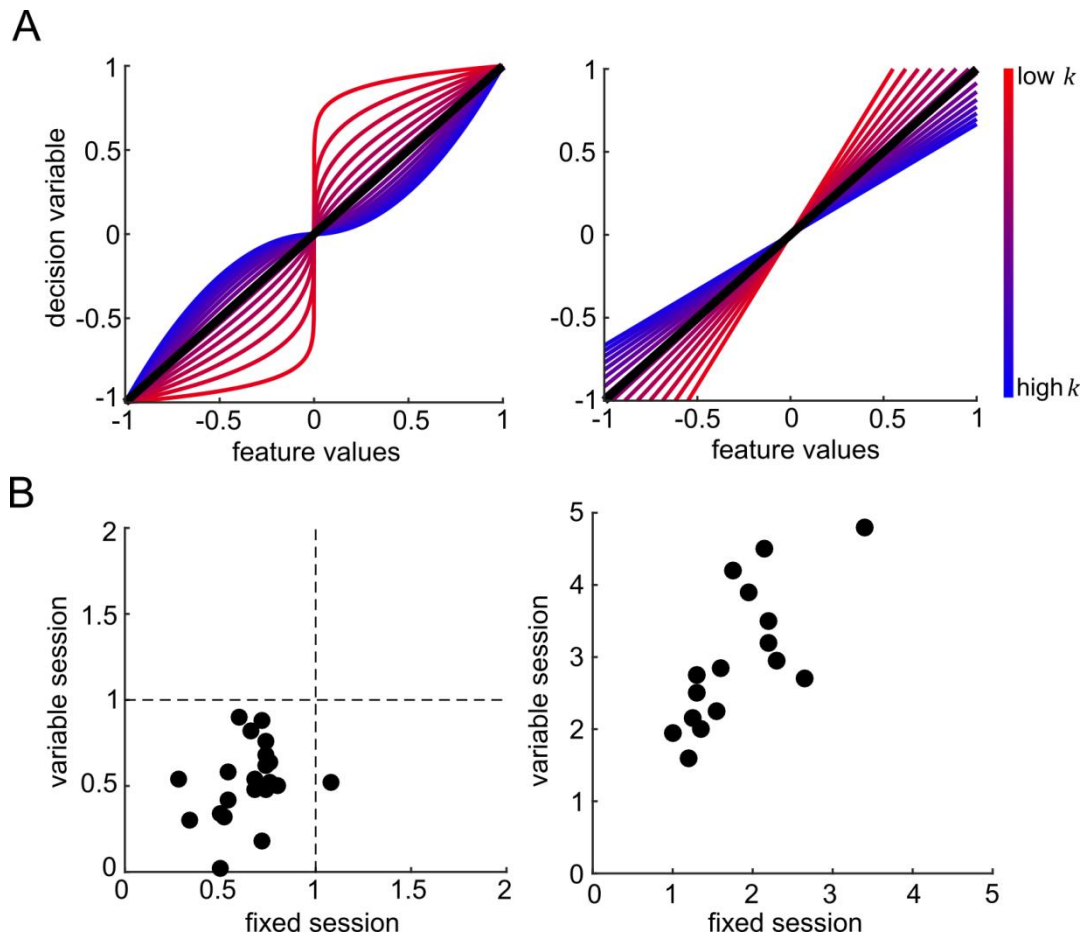


**Figure 3.8. Mapping sensory inputs to decision values**

(A) Left panel: the different functions that map feature values (angles relative to the reference in radians) to decision values for the power model. Coloured lines represent functions for different values of  $k$  from 0.1 to 2, with low values represented by reddish lines and high values represented by bluish lines. Right panel: the equivalent functions for the equivalent gain linear model. In the left

and right panels, models with equivalent gain are represented with lines of equivalent colour. (B) The best fitting  $k$  values (left panel) and  $s$  values (right panel) in human for fixed reference (x-axis) and variable reference session (y-axis).

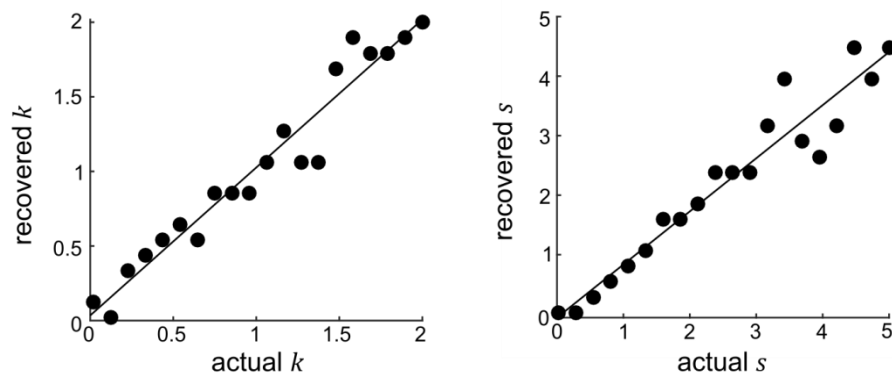
Next, we calculated choice probabilities by passing the  $DV$  through a sigmoidal choice function with the inverse-slope ( $s$ ; Equation 3-11). Varying the inverse-slope of the choice function is approximately equivalent to assuming that decision values are corrupted with varying levels of zero-mean Gaussian noise at a post-averaging stage (e.g., “late” noise), with high values of  $k$  (shallower slope) implying more late noise and thus lower sensitivity. This model allowed us to obtain best-fitting values of  $k$  and  $s$  for each participant in both fixed and variable reference conditions, using maximum likelihood estimation. Values of  $k$  and  $s$  for each participant are plotted for each reference session are displayed in Figure 3.8B.

We observed that values for the inverse-slope of the choice function  $s$  were steeper in the fixed than variable reference condition ( $t_{20} = 4.27$ ,  $p < 0.001$ ), consistent with lower performance in the variable reference condition. This is likely to reflect the additional processing cost for recoding raw orientations relative to the reference when the latter changed from trial to trial. Values of  $k$  did not differ between the fixed and variable reference conditions ( $p = 0.93$ ), but for both conditions, best-fitting values of  $k$  were lower than 1 (fixed:  $t_{20} = 9.41$ ,  $p < 0.0001$ ; variable:  $t_{20} = 3.15$ ,  $p = 0.005$ ). This is consistent with a compression of those array elements that were outlying relative to the reference, i.e., a robust averaging policy. To confirm that the model was showing robust averaging, we then created model choices under the best-fitting parameterisation, by randomly simulating binary choices from the estimates of choice probability using the best-fitting model. Using this approach, we were able to recreate the pattern of accuracy (Figure 3.5 dots) and weighting profile (Figure 3.7B) displayed by human participants. In other words, the model displayed comparable costs to humans in each condition, and exhibited the same tendency to engage in robust averaging.

In the model, robust averaging occurs because of the nonlinear form of the function that maps  $X$ , the feature values, onto  $DV$ , the decision values, which is steeper in the centre (near 0) and shallower at the edges (far from 0). As a control, we tested the weighting profile observed when  $X$  is linearly mapped onto  $DV$ . This confirmed that a linear transformation of feature values did not give rise to robust averaging (Figure 3.7C). Parameter recovery simulation confirmed that  $k$  and  $s$  were fully identifiable for the power model (shown by Figure 3.9 that actual parameters and recovered parameters fall close to the identity line).

**Figure 3.9. Parameter recovery**

Recovered parameters (y-axis) plotted against the actual parameters (x-axis) for  $k$  (left panel) and  $s$



(right panel). Black line is the identity line.

As thus described, our model assumes no noise in the encoding of each individual grating. This assumption follows from the fact that in the experiment, each individual array element (grating) was presented with full contrast and thus the orientation should have been relatively easy to perceive. For example, using a similar stimulus array, one report finds estimates of equivalent encoding noise in the range of 2-6° when contrast values exceed about 0.3 (Solomon, May, & Tyler, 2016). Moreover, although we additionally randomised the latency with which arrays were presented at 4 levels (250, 500, 750 or 1000 ms). Long presentation latencies led to longer RT on correct choices ( $F_{2,47,56.73} = 8.65$ ,  $p < 0.001$ ), but this factor had no influence on accuracy ( $p = 0.42$ ; Figure 3.2). Nevertheless, to test this explicitly, we fit a variant of the model in which feature values  $X_i$  were corrupted by “early” noise alone – a source of variance that arises before any nonlinearity and averaging, that

corrupts each tilt independently relative to the reference (see methods). This model failed to capture the robust averaging effect because the introduction of early noise with power transformation would lead to a more stochastic choice pattern. The same feature value that are corrupted by random early noise would sometimes drive the decision to one choice and sometimes to the other choice. We formally compared this “Early noise only” model to our “Late noise only” model, i.e., to that with  $k$  and  $s$  described above, finding that it fits the condition-wise accuracy worse in both the fixed reference session ( $t_{20} = 8.06, p < 0.0001$ ) and the variable reference session ( $t_{20} = 7.97, p < 0.0001$ ; Figure 3.10).

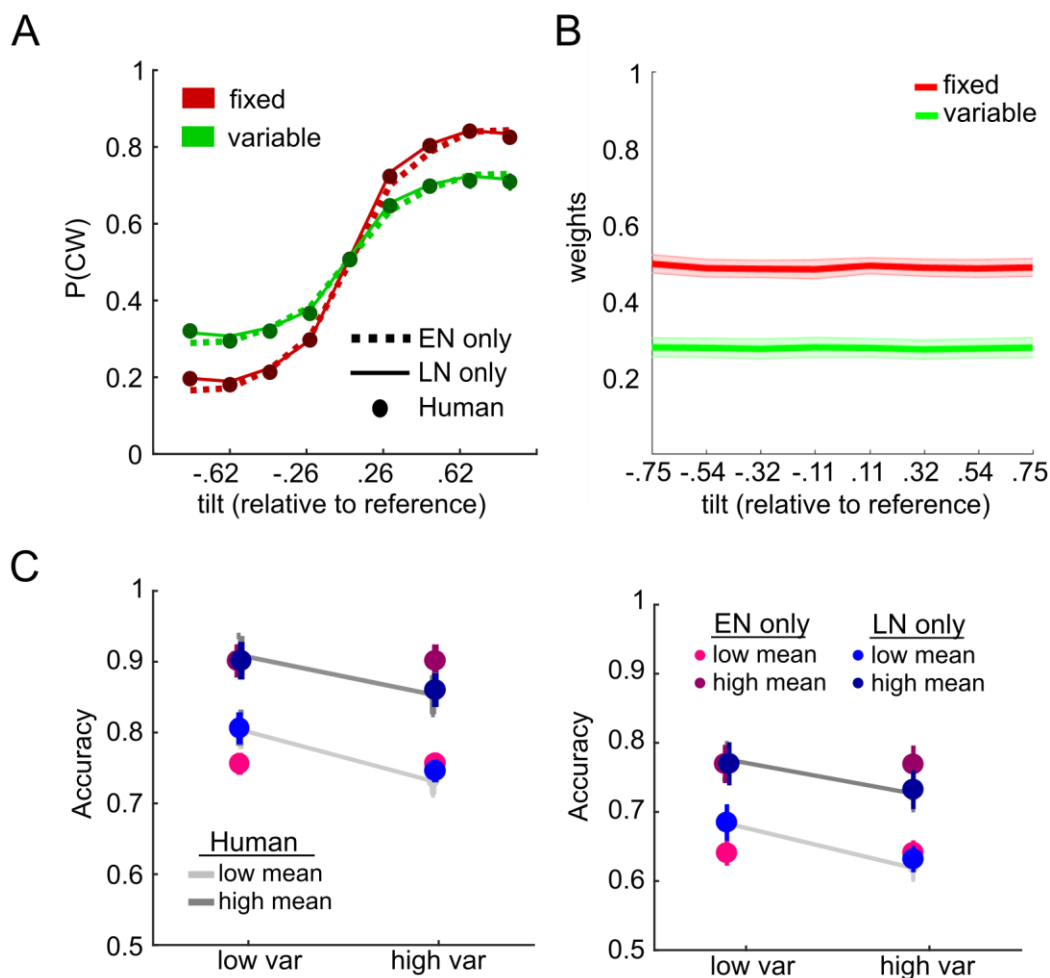


**Figure 3.10. Model comparison of Early noise only model and Late noise only model**

(A) Model psychometric functions (dotted line for “EN only” model and thin solid line for “LN only” model) were plotted against humans (darker coloured dots). Both models successfully capture human psychometric functions of the fixed reference and the variable reference sessions (red vs.

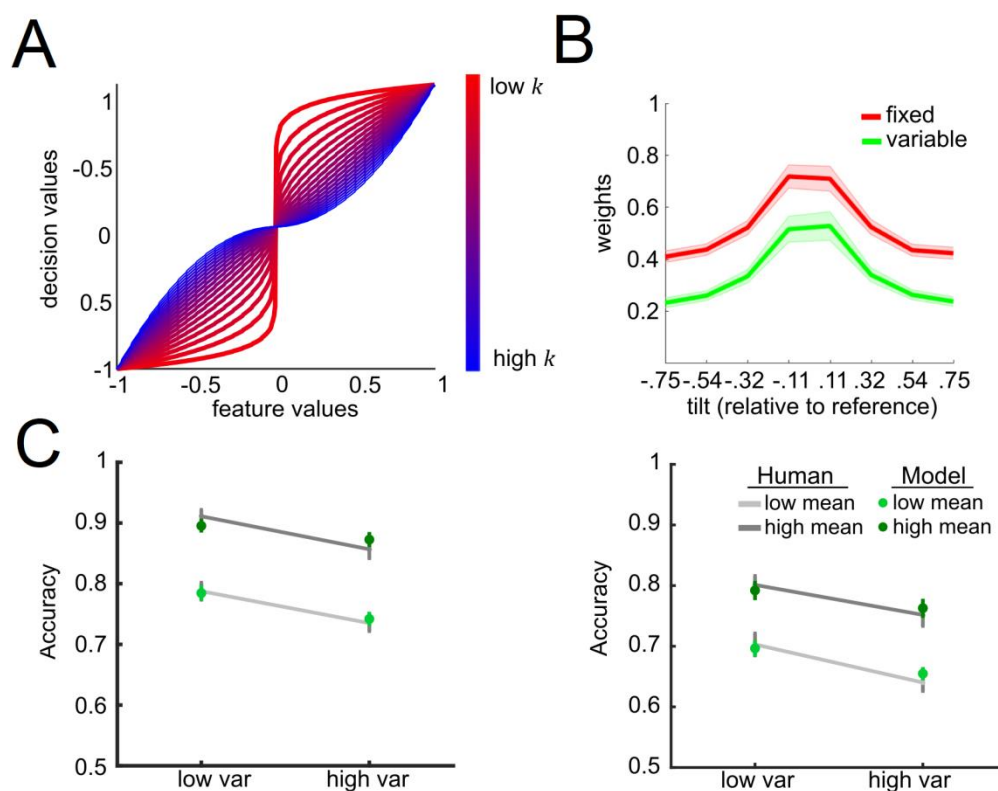
green). (B) Recreation of the weighting function under simulated choices from the best fitting parameterisation of the early noise model. This model failed to replicate human robust averaging as shown in Figure 3.7A. (C) Condition-wise mean accuracy and standard error of mean of the “EN only” model (pinkish dots) and the “LN only” model (bluish dots) superimposed on human accuracies (grey lines). Left panel shows the performance in the fixed reference session, and the right panel shows that of the variable reference condition.

Our model describes the computations that underlie human choices in a simplified fashion, using power-law transducers. However, these functions are intended to describe the output of computations that occur at individual neurons. To demonstrate how transfer functions of this form might arise, we additionally simulated decisions with a population coding model, in which features are processed by a bank of simulated neurons with tuning functions of variable amplitude (see methods). By assuming the height of tuning functions for neurons coding inliers or outliers can vary, we showed in Figure 3.11A that we can recreate the family of transfer functions shown in Figure 3.8A. Given that we could recreate the power-law transducer functions using this model, it is unsurprising that the population coding model was also able to recreate the pattern of accuracy (Figure 3.11B) and the weighting profile (Figure 3.10C) displayed by human participants. However, we chose to model our data with the simpler, psychophysical variant of the model, because it does not require additional

assumptions that are not germane to our main points (e.g., the range of tuning widths for the neuronal population).

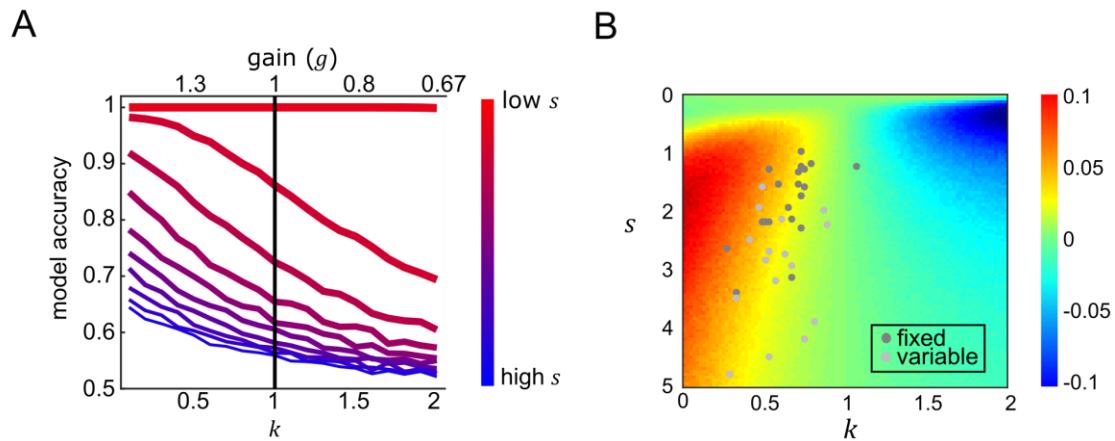
**Figure 3.11. Population coding power model**

(A) Feature values and decision values generated by a population coding power model. Transfer functions that showed feature values were being transformed into decision values in nonlinear ways under different values of  $k$  (coloured lines, in a range of 0.02 to 2), similar to transfer functions shown in Figure 3.8A, which were generated by a simple power model. Tuning width of neurons ( $\epsilon$ ) was assumed to be 0.5 in this illustration. (B) Simulated accuracy under best-fitting parameterisation of population coding. Similar figure shown in Figure 3.5, this figure is showing the mean (and standard error of mean) accuracy of human (grey lines). Green dots represent the simulated mean accuracy (and standard error of mean) using best-fitting parameters yield from humans with the population coding power model. (C) Recreation of parameter estimates using the population coding model. This figure is the same as Figure 3.7B, but instead of using the simple power model, model choices were simulated using the population coding power model under best-fitting parameterisation of 3 parameters ( $\epsilon, k, s$ ).



**Understanding drivers of model performance.** Next, turning to our main point, we used simulation to understand how model performance varied under different levels of late

noise and degree of robust averaging by exploring different values of  $s$  and  $k$ . Model performance (simulated decision accuracy) for the power model under different values of  $k$



and  $s$  is shown in Figure 3.12A (left panel). As expected, performance worsens with increasing late noise (bluish lines). However, performance also depends on  $k$ . When late noise  $s$  is higher, the model performs better with lower values of  $k$  (i.e., those that yield robust averaging). Notably, performance is best with values of  $k$  that are lower than 1, i.e., under a policy that distorts feature information rather than encoding the feature values linearly.

**Figure 3.12. Model accuracy**

(A) Simulated model accuracy for the power model under different values of exponent  $k$  (bottom x-axis, corresponding  $g$  is plotted on the top x-axis) and late noise ( $s$ ; in a range of 0.05 to 5) in coloured lines with reddish (bluish) lines show simulations with lowest (highest) late noise. The black line is the accuracy of the model when items were allocated with equivalent gain and equally integrated ( $k = 1$ ) (B) After simulating model accuracy of the equivalent gain linear model, performance difference between the power model and the linear model is shown in the coloured surface. Positive values (yellow-red) show parameters where the nonlinear model performance is higher than equivalent linear variants, and negative values (cyan-blue) show the converse. Best fitting  $k$  and  $s$  for each subject of the fixed (dark grey dots) and variable reference session (light grey dots) were displayed to show the performance gain relative to using linear weighting scheme.

One trivial reason why model performance might grow as  $k$  is reduced relates to the scaling of the decision values  $DV$  that are produced when  $X_i$  is transformed. After passage through the sigmoidal choice function, larger values of  $DV$  will yield choice probabilities that

are closer to 0 or 1 and thus increase model performance. To adjust for this, we first calculated the scaling of the decision values that resulted from each transfer function parameterised by a different value of  $k$  with Equation 3-2.

This gain normalisation term is proportional to the integral of the absolute value of the curves in Figure 3.3A. This normalisation thus adjusts for the expected gain (i.e., proportional increase or decrease in  $DV$ ) that would be incurred by the nonlinear transducer (in the theoretical case in which there is a flat distribution of features). The normalization thus allowed us to compare nonlinear and linear models with equivalent gain. Figure 3.3 shows the resulting value of  $g$  for each corresponding  $k$ . We then compared the performance of the model under each transfer function with an equivalent linear model, in which decision values were computed under  $k = 1$  (no compression) but rescaled by  $g$ . This is equivalent to assuming that decisions are limited by a fixed resource (or gain), for example an upper limit on the aggregate firing rates produced by a population of neurons.

Creating this family of yoked linear and nonlinear models allowed us to directly assess the costs and benefits to performance of different values of  $k$  in a way that controlled for the level of gain. This can be seen in Figure 3.12B, where we plotted the difference in accuracy between the linear model and a power model that is matched for gain. The red areas in lower left show that when late noise is higher, performance benefits when the model engages more strongly in robust averaging ( $k < 1$ ). In other words, a policy of allocating gain to inliers rather than outliers protects decisions against late noise.

At first glance, this effect might seem counter-intuitive. Why should allocating gain preferentially to one portion of feature space prior to averaging benefit performance, if overall gain is equated? One way of thinking about the difference between a power model (with parameter  $k$ ) and a linear model with equivalent gain  $g$  is that whereas linear model allocates gain evenly across feature space (i.e., equivalently to inliers and outliers), the power model with  $k < 1$  focuses gain on those items that are closest to the category boundary, where the transfer function is steepest. Because the overall distribution of features across

the experiment is Gaussian with a mode close to the boundary, this means that the power model allocates gain more efficiently, i.e., toward those features that are most likely to occur. We have previously described such “adaptive gain” phenomena in other settings (Cheadle et al., 2014).

To verify this contention, we repeated our simulation with a new simulated set of input values  $X$  that were drawn from a uniform random distribution with respect to the reference, rather than using the Gaussian distributions of tilt values that were viewed by human observers. This simulation revealed no performance advantage for robust averaging. Rather, under uniformly distributed features the best policy was to avoid the nonlinear step and simply average the feature values, as predicted by the ideal observer framework. This is shown in Figure 3.13, where best performance under the lowest late noise case occurs when feature values are equally integrated. Under high late noise, values of  $k < 1$  lead to relatively better performance than when all features are equally integrated. However, there is no performance gain for robust averaging compared to the equivalent gain linear model, meaning that unlike in Figure 3.12, the performance gain shown in Figure 3.13 is purely due to a larger scaling of input to output values under  $k < 1$ . We did the same simulation on model predicted accuracy with the weighted sum power model. This model predicts a slight difference that under uniformly distributed stimuli, the model predicts that there is a stronger performance cost under robust averaging ( $k < 1$ ) and performance benefit under “inverse” robust averaging ( $k > 1$ ), see Figure 3.14. This is unsurprising because the log likelihood ratio for features that are far from the category boundary is higher (i.e., “off-channel” tuning, cf.: (Navalpakkam & Itti, 2005; Scolari & Serences, 2009)). When all items are equally likely, under a fixed resource regime, the best policy would be to give more resources to the most diagnostic items i.e., outliers and thus opposite robust averaging is beneficial to performance. This is in fact confirmed by a separate sequential number integration experiment with a different class of stimulus - symbolic numbers. The study showed that that the optimal  $k$

values under high late noise is greater than 1 since the stimulus were drawn from a uniform distribution .

**Figure 3.13. Model accuracy under uniform distributions**

Same figure as Figure 3.12B and Figure 3.12B (A) Stimuli are drawn from a Gaussian distribution (B) Stimuli are drawn from a uniform distribution

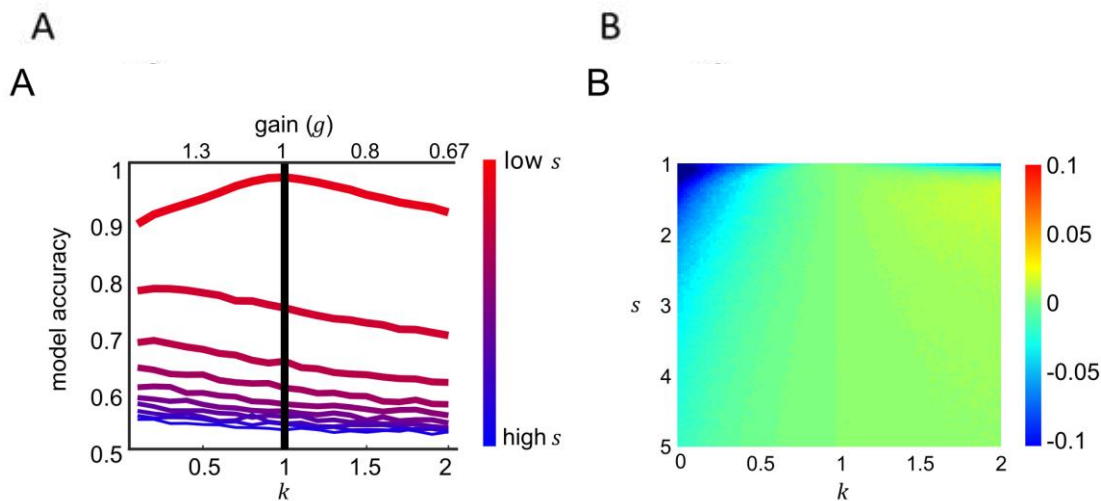


**Figure 3.14. Model-predicted accuracy of the weighted sum power model**

Panels A and B are equivalent to panel A and B for Figure 3.12. However, here the simulations are performed by drawing feature values from uniform random distributions, rather than those used in the human experiment.

### 3.3.3. Linking decision policy to performance.

These explorations allow us to make a new and counterintuitive prediction for the human data. If late noise is high, then rather than hurting decision performance, robust averaging should help. We tested this contention using an analysis approach based on multiple regression. For each participant, we split trials into two groups (even and odd). We first obtained the best-fitting  $k$  and  $s$  parameters for each participant using even trials. Then, using multiple regression, we estimated multiplicative coefficients that best describe the relationship between the best-fitting parameters for each subject and performance on (left out) odd trials, separately for the fixed and variable reference sessions:

$$cor = \beta_0 + \beta_1 k + \beta_2 s + \beta_3 s \cdot k$$

Where *cor* is a vector of mean accuracies (one accuracy for each subject per session), and *k* and *s* are vectors of corresponding best-fitting parameters. In the variable reference condition, both *k* and *s* were significant negative predictors of performance (*k*:  $\beta_1 = -0.14$ ,  $t_{17} = -2.51$ ,  $p = 0.022$ , 95% CI [-0.032 -0.26]; *s*:  $\beta_2 = -0.041$ ,  $t_{17} = -7.15$ ,  $p < 0.001$ , 95% CI [-0.03 -0.052]). In other words, in the variable reference condition, where late noise is intrinsically higher, low values of *k* led to enhanced performance across the human cohort. In the fixed reference session, neither *k* nor *s* was significant predictors of performance ( $p = 0.56$  and  $p = 0.16$  respectively), but their interaction was significant ( $\beta_3 = -0.13$ ,  $t_{17} = -2.88$ ,  $p = 0.01$ , 95% CI [-0.04 -0.21]). In other words, in the fixed reference condition, predicted performance was higher under lower *k* only for those participants with higher estimated late noise *s*. These findings confirm that in our experiment, robust averaging conferred a benefit on performance under high late noise.

### 3.4. Discussion

Human observers have previously been shown to be “robust averagers” of low-level visual features such as shape and colour (de Gardelle & Summerfield, 2011; Michael et al., 2013), and even of high-dimensional stimuli such as faces (Haberman & Whitney, 2010). Here, we add to these earlier findings, describing robust averaging of the tilt of a circular array of gratings. However, the focus of the current experiment was to use computational simulations to understand why humans engage in robust averaging. We describe a simple psychophysical model in which features values are transformed nonlinearly prior to averaging. This model assumes the decisions are limited by a fixed resource, and that gain is allocated differentially across feature space, giving priority to inliers – those features that fall close to the category boundary. Through simulations, we find that in our experiment, this relative discounting of outliers gives a boost to performance when decisions are additionally corrupted by “late” noise, i.e., noise arising during, or beyond, the integration of information.

Previously, robust averaging has been considered a suboptimal policy that incurs an unnecessary loss by discarding relevant decision information (Van den Berg & Ma, 2012). The current work offers a new perspective, suggesting that robust averaging is a form of bounded rationality. If we consider an observer whose neural computations are not corrupted by late noise, it is true that robust averaging incurs a cost relative to perfect averaging. However, here we consider decisions as being constrained not just by sources of noise that are external to the observer, or that arise during sensory transduction, but also capacity limits in human information processing. Processing capacity allows a multiplicative gain to be applied to feature values, with higher gain ensuring that feature values are converted to cumulative decision values that fall further from the category boundary (here, the reference orientation). When decision values are further from the category boundary, they are more resilient to “late” noise, which might otherwise drive them to the incorrect side of the category boundary, thereby forcing an error. However, when gain is limited, it must be allocated judiciously. Our simulations show that allocating gain to stimuli that are most likely to occur confers a benefit on performance, and suggest that humans may adopt a robust averaging policy in order to maximise their accuracy on the task.

One longstanding hypothesis states that neural systems will maximise the efficiency of information encoding by allocating the highest resources (e.g., neurons) to those features that are most likely to occur (Barlow, 1961). For example, enhanced human sensitivity to cardinal angles of orientation (those close to  $0^\circ$  and  $90^\circ$ ) may reflect the prevalence of contours with this angle in natural scenes (Girshick et al., 2011). Indeed, neural systems learning via unsupervised methods will naturally learn to represent features in proportion to the frequency with which they occur. Here, we make a related argument for neural gain control. The efficiency of gain control allocation depends on the distribution of features that occurs in the local environment. Allocating gain to features that are rare or unexpected, even when they are more diagnostic of the category, is inefficient, as resources are “wasted” in feature values that are highly unlikely to occur; whereas allocating gain to those features that occur most frequently will confer the greatest benefit. This benefit, however, is only

observable when decisions are corrupted by “late” noise, i.e., that arising beyond information averaging. This finding has important implications for our understanding of what may be the “optimal” policy for performing a categorisation task. The ideal observer framework allows us to write down a decision policy that will maximise accuracy for an observer that is limited not by capacity but by noise arising in the external environment. Here, we show an example where the policy that is optimal for an unbiased, noiseless observer is not the one that maximises accuracy for healthy humans.

The current study adds to an emerging body of work that the human brain may have evolved perceptual processing steps that squash, compress or discretise feature information in order to make decisions robust to noise (Summerfield & Tsetsos, 2015). In another recent line of work, participants were asked to compare the average height of two simultaneously-occurring streams of bars (Tsetsos et al., 2016) or average value of two streams of numbers (Tsetsos, Chater, & Usher, 2012). Human choices were best described by a model which discarded information about the locally weaker item, but this “selective integration” policy paradoxically increased simulated performance under higher late noise. As described here, participants seemed to adjust their decision policy to account for their own internal late noise: participants with higher estimated late noise were more likely to engage in robust averaging. Like selective integration, thus, robust averaging is a decision policy that discards decision information but paradoxically confers a benefit on choice.

Additionally, the design of our study allows us to draw conclusions about the timescale over which gain allocation occurs. In previous work, robust averaging was found to vary with the overall distribution of features present in a block of trials. For example, when averaging Gaussian-distributed features in a red-to-purple colour space, purple features were relatively downweighted, but when averaging in a red-to-blue colour space, purple features were relatively upweighted (de Gardelle & Summerfield, 2011). In other words, the allocation of gain to features depended on the overall distribution of features in the block of trials, with the most frequently-occurring (i.e., expected) items enjoying preferential processing. Here, we saw no difference in robust averaging between a fixed reference condition (in which the

Gaussian distribution of orientations remained stable over a prolonged block of trials) and a variable reference condition (in which the Gaussian distribution of orientations changed from trials to trial, and was uniform over the entire session). In other words, any adaptive gain control was set by the reference, and thus occurred very rapidly, i.e., within the timescale of a single trial. Evidence for remarkably rapid adaptive gain control has been described before. Indeed, short-lag repetition priming may be considered a form of gain control (Grill-Spector, Henson, & Martin, 2006), in which the prime dictates which features should be processed preferentially. During sequential averaging, the behavioural weight and neural gain applied to a feature depend on its distance from the cumulative average information viewed thus far, as if features pass through an adaptive filter with nonlinear form (Cheadle et al., 2014). These observations are consistent with the theoretical framework that we propose here.

Finally, we discuss some limitations of our approach. Firstly, our model uses a simple power function to describe the nonlinear transformation of inputs prior to averaging. We chose this function for mathematical convenience – it provides a simple means of parameterising the mapping function feature to decision information in a way that privileges inliers ( $k < 1$ ) or outliers ( $k > 1$ ). However, other forms of nonlinear transformation that are not tested here may also account for the data. Secondly, our best-fitting model assumes zero sensory encoding noise (or ‘early’ noise). Adding early noise to the model did not change qualitatively the benefit of robust averaging under higher late noise, unless it becomes performance-limiting in itself. However, in other settings, early noise will be an important limiting factor on performance. Although we found that our “late noise only” model fit better than an “early noise only” model, we do not wish to claim that there is no early noise in our task. Since the current experiment was not designed to estimate the level of early noise, it may be of interest to directly manipulate both early and late noise in future experiments.

## Chapter 4 - Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making

### 4.1. Introduction

Decisions about sensory signals, cognitive propositions, or economic prospects are often made in the context of competing or distracting information. Consider the following everyday situations: you are judging whether a painting hangs straight on the wall, but the nearby pictures are hung askew; you are waiting at a red stop signal, but the car in front decides to jump the light; you are contemplating the purchase of a new watch, but it is displayed next to a range of more elegant but unaffordable models. In each of these situations, the best decisions will be made by ignoring the distracting sensory signals (the competing picture frames, vehicles, or watches) and focusing exclusively on the choice-relevant information. This normative contention can be formalised in a variety of ways, for example via the notion that rational choices should be independent of irrelevant alternatives (Luce, 1959; von Neumann & Morgenstern, 1944) or that sensory signals should be weighted lawfully by their reliability and relevance to the choice at hand (Kording, 2007; Ma, Beck, Latham, & Pouget, 2006; Wald & Wolfowitz, 1949).

Nevertheless, empirical observations suggest that human decisions are unduly influenced by distracting information. Consider a generic problem in which a target stimulus  $X^i$  and distracters  $X^j$  occur at fixed spatial locations  $i$  and  $j$ . In this general formulation, decision values  $X$  may be perceptual features (such as the tilt of a grating) or economic attributes (such as the quality of a consumer product) that are to be evaluated or categorised. Humans show systematic biases that reflect the influence of the distracters on decisions about the target. For example, vision scientists have long studied the “tilt illusion”, in which the reported orientation of  $X^i$  (e.g., a central grating) is repulsed away from the mean tilt of  $X^j$  (surrounding gratings with similar but nonidentical tilt; Blakemore, Carpenter, &

Georgeson, 1970; Figure 4.1). In cognitive psychology, the influence of distracter items is usually studied with a view to understanding the attentional or control mechanisms that allow information to be selected in the face of conflict. For example, in the classic Eriksen flanker task, observers classify a target stimulus (e.g., a central arrow) that is flanked by distracters (e.g., arrows pointing in compatible or incompatible directions; Eriksen & Eriksen, 1974). It is ubiquitously observed that incompatible flankers incur a cost, and compatible flankers confer a benefit, relative to a neutral condition, as measured in response times (RTs) and accuracy (Figure 4.1B). Finally, behavioural and neural economists have charted the irrational influence that a decoy alternative of value  $Z$  has on choices between two choice-relevant prospects  $X$  and  $Y$ , where  $X > Y$  (Chau, Kolling, Hunt, Walton, & Rushworth, 2014; Louie, Khaw, & Glimcher, 2013; Tsetsos, Usher, & Chater, 2010). A common finding is that rational choices (i.e., for  $X > Y$ ) initially decline as  $Z$  increases in value, but then increase sharply as  $Z$  comes to approximately match the other two items in value (Figure 4.1C).

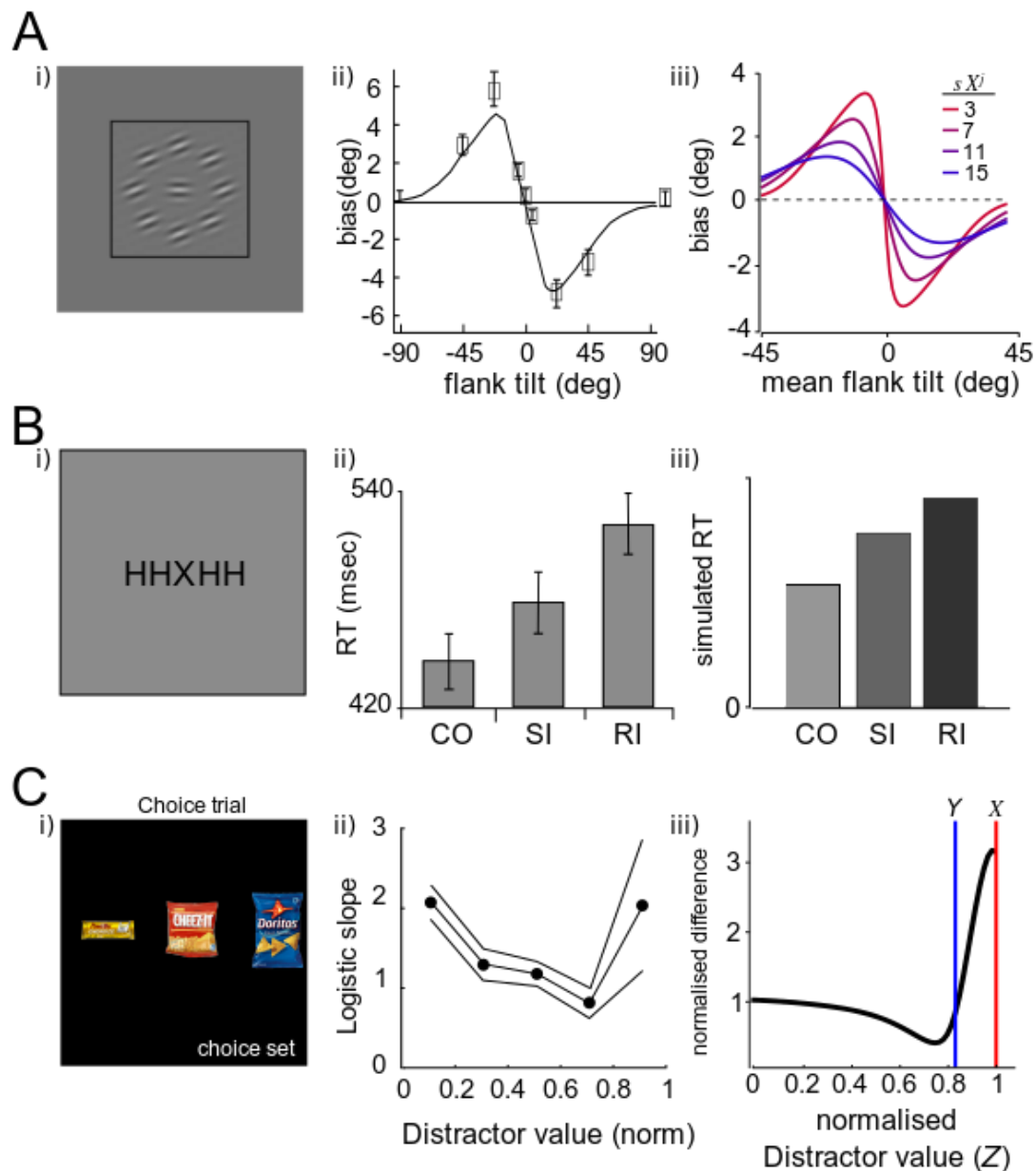


**Figure 4.1. The effect of distraction across perceptual, cognitive, and economic domains.**

(A) i) subjects were asked to discriminate the tilt (relative to horizontal) of a central Gabor surrounded by tilted distractors. ii) participants are biased to report the target as more clockwise when the flankers were counterclockwise and vice versa. Panels i and ii republished with permission of Royal Society, from (Solomon & Morgan, 2006); permission conveyed through Copyright Clearance Centre, Inc. iii) simulation of the adaptive gain model replicates the qualitative tilt illusion pattern as (Solomon & Morgan, 2006) and replicates the current human data both qualitatively and quantitatively. It further predicts that the magnitude of the bias is modulated by flanker variance; coloured lines reflect flanker variabilities from low (red) to high (blue). (B) i) In the Eriksen flanker task, participants respond with a key press to a central letter while ignoring the flankers. ii) response

times are the fastest on CO ('congruent') trials, then the SI ('Stimulus Incongruent') trials, and slowest on RI ('Response incongruent') trials; Panel ii republished with permission of MIT Press - Journals, from (Van Veen & Carter, 2002); permission conveyed through Copyright Clearance Centre, Inc. See Methods for details of how CO, RI and SI trials were defined. iii) the adaptive gain model predicts the same pattern of reaction time across the three conditions. (C) i) participants chose the most preferred of three food items. ii) increasing the value of the least-preferred item reduces the choice efficiency (i.e., probability of choosing the highest-valued target) as the normalised distractor value increases, shown by logistic slope from fitting logistic choice functions on humans choice; Panel ii republished with permission from (Louie et al., 2013).iii) Using the adaptive gain model, we simulated the normalised subjective difference between two options X and Y (blue and red line) as a function of a third decoy Z (x-axis). The subjective difference is first reduced and then increased in a qualitatively similar fashion.

In the fields of psychology, economics, and neuroscience, diverse theoretical proposals have been offered to explain the cost that distracters incur during decision-making. These include models that describe how control systems detect and resolve conflict among inputs (Botvinick, Braver, Barch, Carter, & Cohen, 2001), accounts that emphasise inhibitory interactions among competing sensory neurons, or favour a normalisation of stimulus values by a local average or range (Heeger, 1992; Louie et al., 2013; Schwartz et al., 2009; Soltani, De Martino, & Camerer, 2012), and Bayesian accounts that model spatial uncertainty among targets and distracters (Dayan & Solomon, 2010; Yu, Dayan, & Cohen, 2009), or that assume a non-uniform prior on the compatibility of decision information (Yu et al., 2009). These accounts disagree about the computational mechanisms involved, the neural processing stages at which the cost of distraction arises, and about the brain structures that are recruited to protect decisions against irrelevant information. For example, divisive normalisation mechanisms may occur in sensory neurons in visual cortex (Heeger, 1992), or amongst value representations in the orbitofrontal cortex (Padoa-Schioppa, 2009), whereas the control systems that detect and resolve conflict have been attributed to medial and lateral prefrontal structures (Botvinick et al., 2001). As such, the field currently lacks a single, unified theory that can account for the effect of distraction on human decisions, or an integrated neural account of its implementation across perceptual, cognitive and economic domains.

The goal of the current paper is to offer such an account. We begin with a simple computational model that is motivated by past work, and shows that contextual signals determine the gain of processing of consistent (or “expected”) features during decision-making tasks (Cheadle et al., 2014). Our model, which is described here at the level of neural population codes, proposes that contextual signals sharpen the tuning curves of neurons with a compatible preference for decision-relevant features, and is motivated by a large literature emphasising the need for adaptive gain control in the service of efficient coding (Rangel & Clithero, 2012; Simoncelli & Olshausen, 2001). Using computational simulations, we first show that the model can recreate qualitatively two classic phenomena in very different domains – perceptual choice (the tilt illusion) and economic choice (simple distractor effects). Next, we turn our attention to a task that has been a mainstay of cognitive studies of distraction – the Eriksen Flanker task. We built novel variants of the task in which the statistics of the flankers, and the difference between target and the decision bound can vary across conditions. Our simulations show that the model predicts a range of striking, counterintuitive behavioural findings, including “reverse” compatibility effects (where fully-visible, compatible flankers actually hinder, rather than help, behavioural performance). Over 4 behavioural experiments involving human participants, we validate these predictions, using visual stimuli defined by both tilt and colour. Finally, we use functional brain imaging to show that the modulatory influence on decision signals predicted by the model correlates with BOLD signals in the dorsal anterior cingulate cortex (dACC) and interconnected structures, where neural signals have variously been implicated in the context-sensitive encoding of action values (Kolling, Behrens, Mars, & Rushworth, 2012), and the expected value of cognitive control (Shenhav, Cohen, & Botvinick, 2016). We show how our framework, while not wholly inconsistent with either account, can bring together diverse views concerning the function of this controversial brain region (Heilbronner & Hayden, 2016).

## 4.2. Methods

### 4.2.1. Subjects.

For behavioural studies Exp.1 and 2, human participants were recruited via the online testing platform provided by Amazon Mechanical Turk (Exp. 1a:  $N = 37$ ; Exp.1b:  $N = 36$ ; Exp. 2a:  $N = 28$ ; Exp.2b:  $N = 30$ ). For Exp.3, 20 healthy volunteers with normal or corrected-to-normal vision and no history of neurological disorders were recruited to participate from the University of Granada, Spain. All participants gave informed consent to participate in the study and were compensated at a rate of 6 per hour for Exp. 1 and 2 and €10 per hour for the fMRI scanning session. All experiments were approved and conducted in accordance with the Medical Sciences Inter-divisional Research Ethics Committee (MS IRDREC) ethical guidelines.

### 4.2.2. Design and behavioural analysis.

**Experiment 1a, 1b and 2a.** Each trial began with a fixation dot 700ms, followed by the simultaneous presentation of target and flankers for 1000 ms. Target and flankers were full-contrast square-waved gratings with a  $180^\circ$  phase (5 cycles, and 75 pixels in diameter). The flankers are presented within a circular aperture of the target grating with 1.5 pixels separation with adjacent flankers. Participants were asked to sit at a comfortable distance from the screen, where the stimuli were presented in a small central area of the screen. The orientation of each individual grating in the flanker surround was chosen from a Gaussian distribution with a particular mean ( $\overline{X^j}$ ) and standard deviations  $sX^j$ ; see main text for precise values). To ensure that the sampled orientations matched the expected distribution with the given  $\overline{X^j}$  and  $sX^j$ , resampling of orientation values occurred until the mean and standard deviation of orientations fell within  $1^\circ$  tolerance of the desired  $\overline{X^j}$  and  $sX^j$ . Within the stimulus presentation duration, participants had to indicate whether the central grating was clockwise or counterclockwise to the vertical axis by pressing the right and left keyboard arrows respectively. They were instructed to ignore the flanking gratings. Visual feedback was

given immediately after their response. Participants completed 6 blocks of 96 trials for Exp. 1a, 10 blocks of 96 trials for Exp. 1b, and 10 blocks of 108 trials for Exp. 2a and 2b).

**Experiment 2b.** While keeping all other aspects of the experiment (e.g., timing) the same, in Exp.2b we substituted the grating stimuli with coloured stimuli from a continuous colour space. Target and flankers were circles (100 pixels diameter per circle; 19 pixels separation with adjacent flanking circles) with hue ranging from blue to red. Participants were asked to indicate whether the central colour circle was more red (right arrow) or blue (Left arrow) while ignoring the surrounding coloured circles. Individual colour values ( $c$ ) were defined by RGB values using expression:  $RGB = [c, 0, 1 - c]$  with the category boundary being defined as  $c = 0.5$  or  $RGB = [0.5, 0, 0.5]$ . The colour value for the central target stimulus and the flanker mean colour value were drawn from the set 0.35, 0.4, 0.45, 0.55, 0.6, 0.65. To facilitate comparison between Exp.2a and Exp.2b, we recentred the category boundary value to 0, which means the central target colour values ( $X^i$ ) can be redescribed as  $\pm 0.05$ , 0.1, 0.15 with respect to a category boundary of zero. In this space individual flanker values were drawn from Gaussian distributions with means ( $\overline{X^j}$ ) of  $\pm 0.05$ , 0.1, 0.15 and standard deviations ( $sX^j$ ) of 0, 0.1, 0.2. Resampling of flanker feature values occurred until  $\overline{X^j}$  and  $sX^j$  fell within 0.01 from the desired values. Participants completed 10 blocks of 96 trials for this experiment.

**Experiment 3 (fMRI study).** All stimuli were presented on a grey background. Participants first saw a fixation dot for 500ms, followed by the presentation of stimuli (subtended  $3.09^\circ$  visual angle) for 500ms, and were asked to respond within 1500 ms of stimulus onset. Stimuli were tilted gratings as for Exp.1 and Exp.2a. At the end of response window, visual feedback in the form of a green or red dot indicated whether the response was correct or not for 300ms. There was a jittered interval of between 2-6s (on average 4s) interposed between trials. Target tilt and flanker tilt statistics were varied independently on a trial-by-trial basis. On each trial, tilts of the target grating and flanker mean orientation were drawn independently from uniform distributions with a limit of  $-45^\circ$  and  $+45^\circ$ . Similarly, flanker standard deviation could be valued between 0 to  $30^\circ$  from a uniform distribution.

Resampling was used to ensure that the number of congruent and incongruent trials was matched across the entire experiment. Participants completed 4 ( $N = 10$ ) or 5 ( $N = 10$ ) blocks of the task in the scanner, with 130 trials on each block.

For experiments 1 and 2, the design orthogonalised the manipulation of target feature value ( $X^i$ ), mean of flankers ( $\overline{X^j}$ ) and variability of flankers ( $sX^j$ ). We can further designate trials as “congruent” when  $X^i$  has the same sign as  $\overline{X^j}$  or “incongruent” when  $X^i$  has the opposite sign as  $\overline{X^j}$ . In experiment 1, we thus have a  $3 \times 2$  (flanker variability  $\times$  congruency) within-participant factorial design. In experiment 2, we introduced three levels of  $|X^i|$  (equivalent to six levels of signed  $X^i$ , three levels for each category), three levels of ( $|\overline{X^j}|$ ) and three levels of  $sX^j$ , resulting in  $3 \times 3 \times 3 \times 2$  ( $|X^i| \times |\overline{X^j}| \times sX^j \times$  congruency) within-participant factorial design with 54 conditions. Full list of  $|X^i|$ ,  $|\overline{X^j}|$ , and  $sX^j$  levels were displayed in Table 4-1. For Exp. 1 and 2, ANOVAS with Greenhouse-Geisser correction for sphericity were carried out at group-level analyses. A threshold of  $p < 0.05$  was adopted for all behavioural analyses. Effect size - partial eta-squared ( $\eta_p^2$ ) was computed for all significant effects. We only analysed RT from correct trials, and additionally excluded trials where RT was faster than the 1% percentile or slower than the 99% percentile of the RT distribution. We used the same exclusion criteria across experiments. These two exclusion criteria led to the following mean percentage (standard deviation; SD) of trials exclusion across subjects for each experiment: Exp. 1a – 6.85% (4.21); Exp. 1b – 6.73% (3.52); Exp. 2a – 5.62% (3.76); Exp. 2b – 10% (5.44). We have also verified that all the reported effects remained significant when we replaced mean RT with median RT or log-transformed RT.

**Table 4-1. Summary of experiments. \***

| Experiment | Target decision variable ( $X^i$ ) | Flanker mean decision variable ( $\overline{X^j}$ ) | Flanker standard deviation ( $sX^j$ ) |
|------------|------------------------------------|-----------------------------------------------------|---------------------------------------|
| 1a         | $\pm 45^\circ$                     | $\pm 45^\circ$                                      | 0, 15, 30                             |
| 1b         | $\pm 20^\circ$                     | $\pm 20^\circ$                                      | 5, 10, 15                             |
| 2a         | $\pm 15, 30, 45^\circ$             | $\pm 15, 30, 45^\circ$                              | 0, 15, 30                             |
| 2b         | $\pm 0.05, 0.1, 0.15$              | $\pm 0.05, 0.1, 0.15$                               | 0, 0.1, 0.2                           |
| 3          | U(-45,+45)                         | U(-45,+45)                                          | U(0,30)                               |

\*Target decision variable, flanker mean decision variable and flanker variance were orthogonally manipulated in each experiment. **Exp. 1** focused on the influence of flanker variance and congruence on RT. **Exp. 2** introduces additional factors of target orientation and flanker mean in addition to flanker variance. In **Exp. 3**, the fMRI task, all variables were both parametric and orthogonal to each other. Values of each variable were drawn randomly from a uniform distribution  $U(min, max)$  with the range shown. See methods for more details.

**fMRI acquisition and preprocessing.** Images were acquired with 3-Tesla Siemens TrioTim with a 32-channel head coil using a standard echo-planar imaging sequence. Images were 64 x 64 x 32 volumes with voxel size 3.5 x 3.5 x 3.5 mm; acquired with a 2s repetition time and echo time of 30 ms. We acquired fMRI data in 4 or 5 runs with 425 volumes per run. All preprocessing and fMRI analyses were carried out using SPM12. Preprocessing of the imaging data includes realignment of function images, co-registration of anatomical scan to the mean functional image, followed by segmentation and spatial normalisation to the standard template brain of the Montreal Neurological Institute (MNI brain). Lastly function images were spatially smooth with a 6-mm full width half maximum Gaussian kernel. A 128-s temporal high-pass filter was applied to exclude low-frequency artefacts.

#### 4.2.3. Adaptive gain model.

The adaptive gain model is based on a framework that was previously developed to understand how humans performing spatial and temporal averaging tasks adapt to the context provided by proximal decision information (Cheadle et al., 2014; Michael et al., 2014). Inputs arrive at a population of  $n$  decision neurons each characterised by a Gaussian tuning curve centred on its preferred feature value  $\theta_k$ . Each neuron  $k$  responds to the target stimulus  $X^i$  with rate  $R_k = f(X^i|\theta_k, \sigma_k)$ , where  $f(X|\theta, \sigma)$  denotes the probability density function of the normal distribution with mean  $\theta$  and variance  $\sigma^2$  evaluated at value  $X$ .

The estimated output of the neural population is then linearly decoded into a subjective percept on the decision value of the target  $\widehat{X}^i$  by weighting the population activity  $R$  by the corresponding feature values  $\theta$ .

$$\widehat{X}^i = \sum_{k=1}^n R_k \theta_k$$

Equation 4-1

When the gain is uniformly spread across the feature space (i.e., the tuning widths  $\sigma_k$  for all neurons are equal), this approach decodes each input to its original feature value (subjected a linear scaling depending the range of feature space and the number of neurons, see Chapter 2 and 4 for a correction of the scaling). However, our model proposes that the context provided by the distracters modulates the sharpening of neuronal tuning (Gilbert & Wiesel, 1990), with a tuning width envelope that matches the inverse distribution of contextual features  $X^j$  with mean  $\overline{X^j} = \sum_{j=1}^6 X^j / d$  and standard deviation (SD)  $sX^j$  where  $d$  is the number of distractors.

$$\sigma_k = \sigma^{max} - f(\theta_k | \overline{X^j}, sX^j + \epsilon) n$$

Equation 4-2

In other words, neurons with a preferred orientation that matches ( $\overline{X^j}$ ) have the sharpest tuning curves, and these tuning curves are even sharper if the flanker variance ( $sX^j$ ) is low (see Figure 4.4A). In Equation 4-2,  $\sigma^{max}$  denotes the maximum tuning width in the population,  $\epsilon$  is a constant parameter added to  $sX^j$  ensure that the tuning widths are not zero, and  $n$  is a scaling parameter equivalent to the number of decision neurons.

To fit model outputs to human response time data (i.e., on a common scale in ms), for each parameterisation we regressed inverse decision values against each individual participants' response times:

$$RT_{gain} = \beta^0 + \beta^1 \cdot |\widehat{X}_i|$$

## Equation 4-3

This calculation of RT is equivalent to modelling the data with ballistic (noiseless) diffusion process, with the two additional parameters  $\beta^0$  and  $\beta^1$  encoding the non-decision time and the drift rate respectively (fixed across conditions). We used a ballistic accumulation process for simplicity, but note here that errors could be modelled by adding a noise term to the accumulation process. Searching exhaustively across values of  $\sigma^{max}$  and  $\epsilon$  from Equation 4-2, we identified the parameters that minimised mean squared error (MSE) between the human and model-predicted average RTs for each condition.

#### 4.2.4. Conflict models.

For the conflict model, we use a formulation described previously, whereby conflict  $C$  depends on the weighted product of competing inputs for the two actions  $A^{CW}$  and  $A^{CCW}$

$$\begin{aligned} C &= [A^{CW} \cdot A^{CCW}]; \\ A^{CW} &= g(\overline{X^j}) \cdot (1 - w) + g(X^i) \cdot w; \\ A^{CCW} &= h(\overline{X^j}) \cdot (1 - w) + h(X^i) \cdot w; \end{aligned} \quad \text{Equation 4-4}$$

where  $g(X)$  and  $h(X)$  respectively denote positive and negative linear rectification functions. In conflict model 1, activation for the two actions  $A^{CW}$  and  $A^{CCW}$  are proportional to the tilt of the target and flanker mean. Alternatively, they can also be defined according to the tilt of the target and each of the  $d$  individual flankers (conflict model 2).

$$\begin{aligned} C_i &= [A_i^{CW} \cdot A_i^{CCW}]; \\ A^{CW} &= \sum_{j=1}^N g(X^j) \cdot \frac{1-w}{d} + g(X^i) \cdot w; \\ A^{CCW} &= \sum_{j=1}^N h(X^j) \cdot \frac{1-w}{d} + h(X^i) \cdot w; \end{aligned} \quad \text{Equation 4-5}$$

Where  $j$  is the number of flankers that are either congruent or incongruent, and  $d$  is the total number of flankers on a given trial. In other words, these models made different assumptions about conflict: that it was computed at the level of individual flankers (conflict model 2) or at the level of summary statistics (conflict model 1).

Finally, we compute RTs in a similar fashion as for the gain model:

$$RT_{conflict} = \beta^0 + \beta^1 \cdot C$$

Equation 4-6

To reduce the risk of overfitting we used cross-fitting, estimating model parameters on half the trials and computing mean squared error (MSE) for the other half. These MSE were then fed into Bayesian Model Selection (BMS) to compute exceedance probabilities.

#### 4.2.5. Simulations of perceptual and economic decisions.

When fitting to human data, the adaptive gain model contained 2 free parameters: maximum tuning width ( $\sigma^{max}$ ) and an constant parameter ( $\epsilon$ ). Simulations of the model aimed at qualitatively recreating effects from the past literature (e.g., for Figure 4.1) assumed a fixed  $\sigma^{max}$  ( $\sigma^{max}=10$ ) and a fixed  $\epsilon$  ( $\epsilon = 5$ ) unless noted otherwise. We have also imposed a floor value for  $\sigma_k$  (Equation 4-2). Any values that are below 0 will set as 0.1 so that the tuning width of any neurons will not be in negative values.

**Tilt illusion.** In this simulation, we plot the difference between the true target angle (here, zero) and the gain modulated decision value ( $\widehat{X}^i$ ) as a function of flanker mean decision value  $\overline{X}^j = \{-45, -44, \dots, +45^\circ\}$  and flanker standard deviation  $sX^j = \{3, 7, 11, 15\}$ . For each variant of flanker mean decision values and flanker standard deviation,  $\widehat{X}^i$  is computed using Equation 4-1 and Equation 4-2. We then plot ( $\widehat{X}^i$ ) against levels of flanker mean decision value and flanker standard deviation in Figure 4.1A.

**Conflict effects.** We computed a proxy of RT ( $1/|\widehat{X}^i|$ ) for the 3 conditions: ‘CO’: Congruent, where the target shares the same response association with the flankers; ‘SI’: Stimulus Incongruent, where the target is perceptually different to the flankers but the response associations of the two are still the same; ‘RI’: Response Incongruent, where the target has a different response association as the flankers (Figure 4.1B). In the simulation, flanker standard deviation  $sX^j$  is set to be 0 in both ‘CO’ and ‘RI’ condition (i.e., we assumed

flanker standard deviation as  $\epsilon$ , or  $5^\circ$ ). In 'CO',  $X^i$  is equal to  $|\widehat{X^i}|$  (both are  $+45^\circ$ ). In 'RI',  $|\overline{X^i}|$  has the opposite sign to the  $X^i$ . Lastly, we simulated 'SI' condition by assuming  $sX^j$  is higher than 0 i.e.,  $sX^j=5$ ; individual flankers are variable but  $|\overline{X^i}|$  remained the same as  $X^i$ . We assumed a higher maximum tuning width ( $\sigma^{max}=15$ ) in this simulation.

**Decoy effect (single attribute).** We simulated the difference between the model estimated decision values from the two targets ( $X=20$  &  $Y=10$ ) as a function of a third distractors' decision value  $Z = \{-45, -44, \dots, +20\}$ .  $\epsilon$  was assumed to be 10 in this simulation. In Figure 4.3, we computed the model output associated with the highest-valued choice-relevant alternative ( $\hat{X}$ ) and the next best alternative ( $\hat{Y}$ ), assuming that the mode of the gain field determined by the lowest-valued (i.e., irrelevant) alternative  $Z$ . We then plotted the normalised subjective estimates difference between the best option and the next best option ( $\frac{\hat{X}-\hat{Y}}{\hat{Y}}$ ), a quantity proportional to choice probability in Figure 4.1c.

**Value of chosen vs. unchosen option.** We simulated the model output as a function of the value of a theoretical chosen and unchosen option with a  $\epsilon = 10$ . On each trial, decision values for two stimuli ( $X^1$  and  $X^2$ ) were drawn independently from two zero-mean Gaussian distribution with a standard deviation of 10. On every trial, we allowed simultaneous evaluation of each stimulus in the context of the other, i.e., we passed each stimulus through the model as a target with the alternative as a distracter. We then assumed that participants chose according to the relative subjective (i.e., model output) value of the maximum and minimum resulting values, using a value of 5 for the slope of the choice function:

$$D = \hat{X}^{max} - \hat{X}^{min}$$

$$CP = \frac{1}{1 + e^{-\frac{D}{5}}}$$

Equation 4-7

This allowed us to plot the relationship between  $X$  and model output  $D$  separately for the chosen and unchosen options.

**Choice bias and tilt illusion.** We used an approach that is related to a probit regression approach that described previously to understand the weighting of information in perceptual averaging judgments. Here, we assess the weight (or influence) that distracters wielded over choices, as a function of whether the tilt of individual flankers was similar or dissimilar to the target tilt (de Gardelle & Summerfield, 2011; Li, Castañón, Solomon, Vandormael, & Summerfield, 2017). We first fitted the adaptive gain model to subjects' RT using the methods described in the main text. Choice probability is computed by passing the model output ( $\widehat{X^i}$ ) under the best parameterization into a choice function in Equation 4-5. We attempted to predict choices on each trial as follows:

$$p(CW) = \Phi(\beta^0 + \beta^1 \cdot X^i + \beta_q^2 \cdot [X^j - X^i]^q)$$

$$RT = \beta^0 + \beta^1 \cdot X^i + \beta_q^2 \cdot [X^j - X^i]^q$$

Equation 4-8

Where  $\Phi$  refers to a probit link function. We first tallied the signed angular difference between the individual flankers and the target  $[X^j - X^i]$  into bins ( $q = 11$ ). Negative/positive values mean that the flankers are more counter-clockwise/clockwise than the target.  $\beta^1$  is the regression coefficient associated with the target orientation, and  $\beta_q^2$  is the coefficient associated with each bin  $q$  of the angular difference between each flanker and the target. We plotted the coefficients  $\beta_q^2$  associated with each bin  $q$  for human choices and RT in Figure 4.7. Same analysis was completed on model choices and RT.

#### 4.2.6. fMRI data analyses

We analysed our data using Statistic Parametric Mapping (SPM12) with the general linear model (GLM) framework and in-house scripts running in Matlab. For all analyses, we ensured that sequential orthogonalisation of predictors in SPM was disabled. All GLMs also included regressors encoding the estimated movement parameters from preprocessing as a nuisance covariate. We modelled trials by convolving regressors coding the onsets and

durations of events with the canonical hemodynamic response function, and regressed them against the BOLD signal. Error trials, and those for which RT fell within the most extreme percentiles ( $<1\%$  or  $>99\%$ ) were modelled separately as a nuisance regressor in all GLMs (trial exclusion:  $6.68\% \pm 3.44SD$ ). We first constructed GLM1 with a single predictor encoding the parametric target decision values  $|X^i|$  of the stimulus, time-locked to the onset of the stimulus. We identified voxels that correlated negatively with this regressor to define regions of interest (ROI) in the dACC, AIC and SPL. Activations in these regions survived FDR correction for multiple comparisons at  $p < 0.05$ . To avoid double-dipping, each region was identified in a leave-one-out fashion, with each participant in turn being omitted from a group-level analysis, which was used to define an ROI with a threshold of uncorrected  $p < 0.0001$ , from which beta values were extracted from the left out participant. For GLM2, we discretise each parametric modulator (i) target decision values  $|X^i|$  and (ii) flanker mean feature value  $|\widehat{X^j}|$  into 4 quartiles bins and included a total of  $4 \times 4 = 16$  regressors (corresponding to each quartile bins of feature values) in the GLM. BOLD betas from each subject were again extracted using each ROI masks defined by the leave-one-out analysis. In GLM3, we included empirically observed response time as a competing regressor to  $\frac{1}{X^i}$  (Full list of active voxels for  $\frac{1}{X^i}$  were displayed in Appendix Table 3.3). In GLM4, we included the following predictors as parametric modulators of the stimulus at the time of the decision (i) target decision values  $|X^i|$ , (ii) flanker mean feature value  $\overline{X^i}$ , (iii) flanker precision, i.e., inverse flanker variability  $1/sX^j$ , (iv) absolute distance between target and flanker mean orientations  $|X^i - \overline{X^j}|$ , (v) the interaction of the distance between target and flanker mean with target decision values  $|X^i - \overline{X^j}| \times |X^i|$ , and lastly (vi) the interaction of the distance between target and flanker mean, target decision values and flanker precision ( $|X^i - \overline{X^j}| \times |X^i| \times 1/sX^j$ ). Full details of active voxels associated with these regressor were displayed in Appendix Table 3.4. Subsequent statistical tests on the betas extracted from each ROI were conducted using two-tailed one-sample *t*-test at a significance threshold of  $p < 0.05$ .

#### 4.2.7. Model Comparisons.

**Behavioural model comparison.** We implemented Bayesian model selection at the group level to obtain posterior likelihoods and exceedance probabilities as described by Stephan and colleagues (Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

**Neural Bayesian model comparison.** Neural Bayesian model comparison. We estimated the following GLMs using Bayesian statistics and conducted model comparisons to quantify and compare the predictions of different GLMs: (i) no-gain modulated decision value  $|X^i|$  i.e., GLM1 described in the main text methods, (ii) gain modulated decision value ( $|\widehat{X}^i|$ ; GLM5), 3 GLMs that contained the no-gain modulated decision value and conflict defined in three ways (iii) GLM6: a binary congruency regressor: ('Cong') (iv) GLM7: conflict computed using Equation 4-4 (4) (v) GLM8: conflict computed using Equation 4-5 ('Ci'). The fits of each GLM were assessed by Random Effects Bayesian Model Selection from the VBA toolbox (Daunizeau, Adam, & Rigoux, 2014). This allows us to estimate the exceedance probabilities and expected frequencies for each model by feeding the log evidence for the aforementioned GLMs within each masks from the leave-one-out analysis. The best model is characterised with the highest exceedance probability and an expected frequency that is above chance (i.e.,  $1/\text{number of models in comparison} = 20\%$ ).

### 4.3. Results

#### 4.3.1. Adaptive gain simulations.

We first show how the model explains both tilt illusion in perceptual choice tasks and decoy effects in economic choice tasks. In Figure 4.1A, we plot the tilt bias over different values of  $\overline{X}^j = \{-45, -44 \dots 45\}$  predicted by the model as a difference of subjective and objective estimates of the target:  $\widehat{X}^i - X^i$  for different levels of  $sX^j = \{3, 7, 11, 15\}$ . The model predicts that subjective estimates are repulsed away from the mean flanker value, as described in numerous previous studies (Blakemore et al., 1970; Gibson, 1937; Solomon & Morgan, 2006), and additionally that the strongest repulsion effect occurs when the flankers

are homogenous, i.e., they are drawn from a distribution with low dispersion. This repulsion of the subjectively decoded values from their objective counterparts occurs when the target feature is close to, but is not identical to, the mean of the distracters (i.e., the location of sharpest tuning), because the variable tuning profile induces a skew in the population activity over features  $\theta_k$  (see Figure 4.2 for a more detailed explanation).

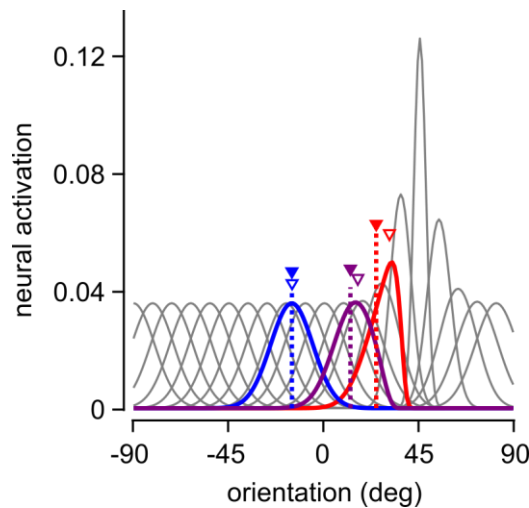


**Figure 4.2. Tilt illusion prediction from the adaptive gain model.**

The illustration displayed the population activity vector  $\theta_k$  for 3 hypothetical targets ( $X^1 = -15^\circ$ , blue solid line;  $X^2 = 15^\circ$ , purple solid line;  $X^3 = 30^\circ$ , red solid line) from a population neurons that their tuning width envelope is based on a ( $X^j = 45^\circ$  and  $sX^j = 5$ ). The population activity for  $X^3$  is more skewed than  $X^1$  and  $X^2$ , the hypothetical target which has the greatest distance to ( $X^j$ ). This target is now perceived as closer to the category boundary at  $0^\circ$ . The strength of repulsion away from ( $\overline{X^j}$ ) is also the greatest with this target. This can be seen clearly by comparing solid triangles and dotted line (decoded orientation from the model) against open triangles (the true orientation). The difference between decoded orientation and objective orientation is the greatest for  $X^3$  (red triangles) and smallest for  $X^1$  (blue triangles).

To model economic decoy effects, we envisage a choice between two prospects of value  $X$  and  $Y$  (where  $X = 20$  and  $Y = 10$  in arbitrary units, such as dollars) that is made in the context of distracters with a value  $Z$ . We plot the difference in their corresponding subjective estimates  $\hat{X} - \hat{Y}$  as a function of  $Z$ , observing a pattern with a striking qualitative resemblance to that reported previously (Figure 4.1c). Again, the model's ability to predict this counterintuitive pattern comes from the repulsive effect induced by differential tuning

across feature space. The model predicts that as the value of the decoy  $Z$  increases, repulsion is first strongest toward  $Y$  (leading to a reduced preference for the best option  $X$ ) but then, as the decoy approaches the two items in the choice set, repulsion is maximal for  $X$ , reversing this effect. In further simulations, we systematically varied both the distance between  $X$  and  $Y$ , and  $Z$ , we were able to capture the two-dimensional pattern of multialternative choice data described in a different study involving abstract shapes associated with different economic values (Chau et al., 2014) (Figure 4.3).

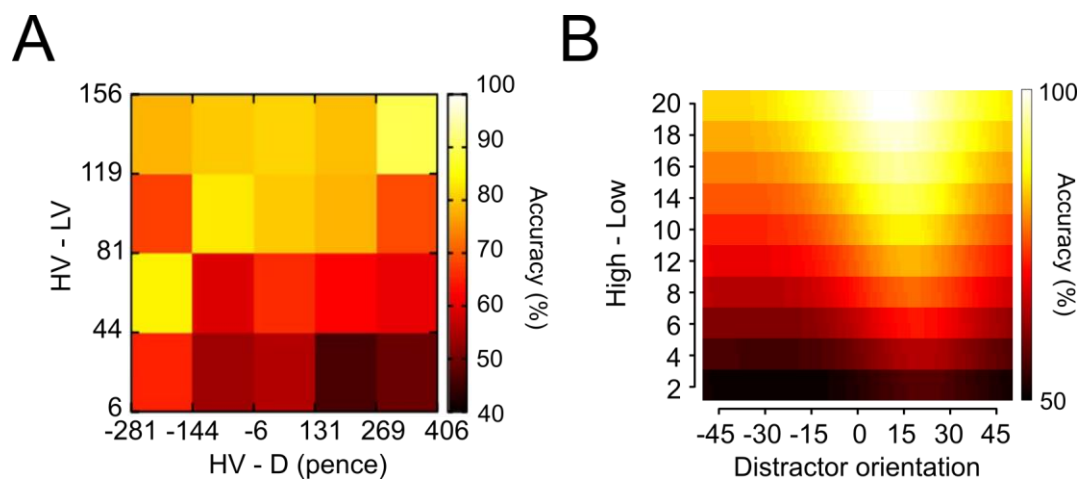


**Figure 4.3. Simulated accuracy for a multialternative choice task.**

(A) Average choice accuracy in humans indicated by colour as a function of both the difference in value between the high valued and low valued options ( $HV - LV$ ) and the difference in value between high valued options and the distractor ( $HV - D$ ). Choices were less accurate when the  $HV - LV$  difference was small. For a given difference between the value of two-valued options, accuracy was lower when the difference between high valued options and the distractor is low than when it is high. Reprinted by permission from (Chau et al., 2014) Springer Nature Nature Neuroscience, copyright (2014). (B) The surface plot illustrates the accuracy (colour bar) for choosing the tilt that is further from the category boundary ('High') among tilts that are closer to the category boundary ('Low') as a function of a third distractor's tilt. When the two tilts are dissimilar, the overall accuracy is higher than when the two tilts are similar. Accuracy is further modulated by the tilt of a third distractor that the accuracy is lower for a given dissimilarity between the two target tilts if the distractors are further away from the category boundary. This replicates the effect from Chau and colleagues that the participants' accuracy over choosing two valued option decreases when the third distractors' values are high.

### 4.3.2. The effect of distractor variance.

Next, we used our model to simulate performance on a new variant of the flanker task that involves categorising a central grating  $X^i$  tilted at  $-45^\circ$  from vertical, in the face of flanking gratings that are on average tilted in a compatible ( $\overline{X^j} = -45^\circ$ ) or incompatible ( $\overline{X^j} = +45^\circ$ ) fashion. As can be seen, the model predicts a compatibility effect: faster RTs for trials where the target and distracters were of congruent sign. However, it also makes a new, testable prediction: that as  $sX^j$  (flanker variance) decreases, response times should be reduced on compatible trials but remain the same on incompatible trials (Figure 4.4c). In this setting, the model predicts slower RTs for incongruent trials, or for congruent but physically dissimilar flankers, as ubiquitously observed. In this setting, the model predicts slower RTs for incongruent trials, or for congruent but physically dissimilar flankers, as ubiquitously observed (Figure 4.1B). Figure 4.4c illustrates predictions from the adaptive gain model under two orthogonally varying factors: compatibility and flanker variability. Because flanker effects for fully-visible stimuli are strongest for response times, we plot the inverse model output  $1/\widehat{X^i}$  as a proxy for RT (i.e., we assume a ballistic evidence accumulation process with slope proportional to  $|\widehat{X^i}|$ ; see Equation 4-3 and Methods). This occurs because on compatible trials (flankers at  $-45^\circ$ ), more gain is allocated to the target feature when the variance of the distribution of flanker orientations is lower. However, the gain allocated to incongruent targets (flankers at  $+45^\circ$ ) is negligibly different across different flanker variance levels since the neural gain they received are similar at the tail of the gain distribution, and so the model predicts that flanker variance should not affect performance on incongruent trials (Figure 1.1Figure 4.4a). By contrast, classic models propose that response conflict varies with the amount of crosstalk interference among responses (Botvinick et al., 2001). These models predict that heightened flanker variance should have equal impact on compatible and incompatible trials (see Methods for model details).

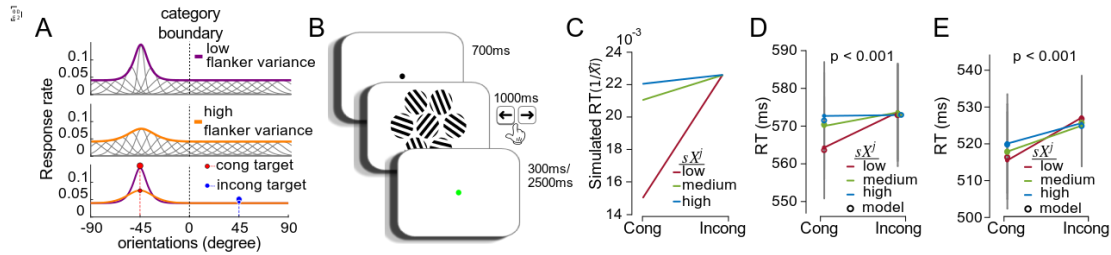


**Figure 4.4. Adaptive gain model of the Eriksen flanker task.**

(A) Illustration of the adaptive gain model as applied to Exp.1. The model assumes that the population width tuning envelope is governed by the trial flanker mean ( $\overline{X^j}$ ) and trial flanker variability ( $sX^j$ ). Congruent targets (red circles) receive higher neural gain than incongruent targets (blue circles), leading to faster RTs. The model further predicts an interaction between congruency and flanker variability. Congruent targets receive high neural gain under low flanker variability trials (large circles) than high flanker variability trials (small circles) while incongruent targets under the two flanker variability trials received the same, low level of neural gain, meaning that flanker variability has no influence on RT for incongruent targets. (B) Task schematics. Participant first saw a fixation dot, followed by an array that contained a central target and 6 surrounding flankers. They responded whether the central target was clockwise or counterclockwise to the vertical axis, receiving feedback after each response. (C) Simulation of  $RT \propto 1/|X^i|$  using the adaptive gain model for 3 levels of flanker variance ( $sX^j$ - in coloured lines) and congruency (x-axis). (D - E) Human mean reaction time pattern (lines) under three levels of flanker variance and congruency condition in Exp. 1a (Panel D) and Exp.1b (Panel E). Both interaction (congruency x flanker variance) are significant at  $p < 0.001$ . Coloured circles represent the fitted mean reaction time from the adaptive gain model. Error bars are showing the standard errors of mean (SEM).

We tested this prediction in Exp. 1 by asking healthy human participants to judge, relative to vertical, the tilt of a single target grating surrounded by 6 flanking distracter gratings (Figure 4.4B; see task details in Methods). The orientation of the target grating  $X^i$  and the mean of the flankers  $\overline{X^j}$  were set to  $\pm 45^\circ$  and the standard deviation of the flankers  $sX^j$  was varied at three levels:  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$  in Exp. 1a ( $n = 37$ ; % error =  $5 \pm 4.3$  SD) and  $5^\circ$ ,  $10^\circ$ ,  $15^\circ$  in Exp. 1b ( $n = 36$ ; % error =  $4.9 \pm 3.7$  SD). Specifically, the zero flanker variance condition in Exp. 1a mimics the classic Eriksen flanker task, where flankers are identically tilted. In both experiments, we observed that flanker variance modulated RTs on compatible trials (Exp. 1a:  $F_{1,93,69.45} = 10.11$ ,  $p < 0.001$ ; Exp. 1b:  $F_{2,67.91} = 9.66$ ,  $p < 0.001$ ) but not on incompatible

trials (both  $p$ -values  $\geq 0.25$ ). This finding was qualified by an interaction between compatibility and flanker variance (Exp. 1a:  $F_{1.76,63.39} = 9.72$ ,  $p < 0.001$ , Figure 4.4D; Exp. 1b:  $F_{1.96,66.79} = 9.67$ ,  $p < 0.001$ , Figure 4.4D). Because previous work has shown that the ratio of compatible to incompatible flankers can modulate performance (Wu, Dufford, Mackie, Egan, & Fan, 2016), we repeated this analysis limited to those trials where all flankers fell on the compatible/incompatible side of the boundary, finding a similar interaction for Exp. 1a ( $F_{1.72,61.9} = 10.61$ ,  $p < 0.001$ ) and Exp. 1b ( $F_{1.85,62.97} = 9.21$ ,  $p < 0.001$ ). A full list on the ANOVA statistics and effect size on RT and accuracy for all experiments is reported in Appendix Table 4.1 and 4.2 respectively.

We fitted our adaptive gain model to the data, and compared its predictions to those of a model proposing that RT depends on response conflict alone. The fits for the gain model (coloured circles) are shown superimposed upon the human data in Figure 4.4D and E. We compared the models head-to-head by computing mean-squared error (MSE) in RT across conditions on half of the data (even trials), after estimating parameters from an independent dataset (odd trials). Bayesian model selection showed that the adaptive gain model fits the human data more closely than the conflict model, with exceedance probabilities for the adaptive gain model of 0.9 in Exp. 1a and 0.58 in Exp. 1b (see Figure 4.5A for the fits of the conflict models). We also compared a version of the gain model in which the contextual modulation was driven by both target and distracters; this model yielded both qualitatively and quantitatively similar results to the original gain model ( $p > 0.3$  for both Exp. 1a and 1b), meaning that it is possible that contextual modulation arises from the entire array, rather than the flankers alone.

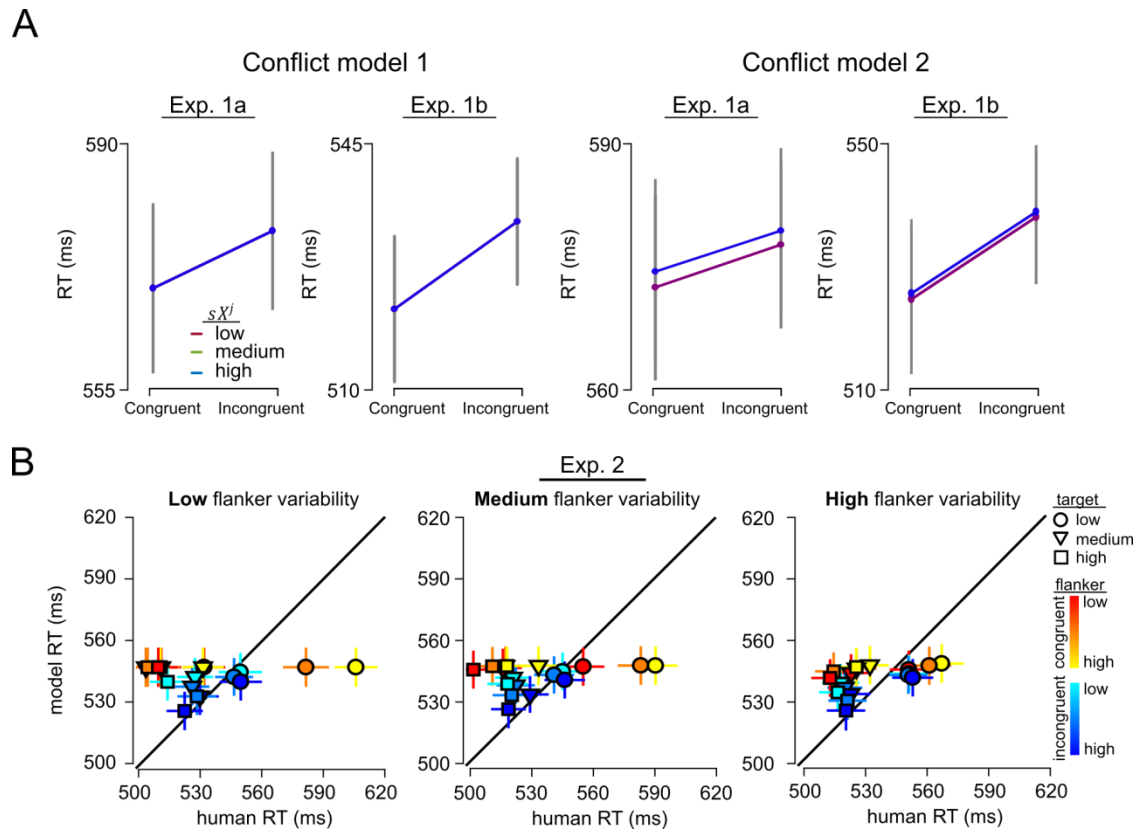


**Figure 4.5. Model fits of the conflict monitoring account using cross-fitting on human response time**

(A) Predicted RT of conflict model 1 (two left-most panels) and conflict model 2 (two right-most panels) for Exp. 1a and Exp. 1b (see Methods for the definition of the conflict models). Conflict model 1 made no further assumptions on the role of flanker variability, thus it failed to predict the impact of flanker variability on performance (different flanker variabilities in three coloured lines were superimposed on each other). Conflict model 2 computed conflict as a function of individual flankers, thus it predicts an increased conflict for high flanker variability trials. However, this is solely because there were some individual flankers that crossed the boundary and were incongruent to the target under high flanker variability condition. (B) Predicted RT of conflict model 2 on Exp. 2. The model failed to capture the two key behavioural signatures we found in Exp. 2. The predicted RT shows a reversed congruency effect; this is because the conflict model would always predict no conflict for congruent trials or a positive conflict for incongruent trials, meaning that it would fail to capture the disproportionate cost for congruent trials rather than incongruent trials in humans. However, for consistency of model fitting practice across the manuscript, we did not limit the range of drift rate ( $\beta^1$ ; Equation 4-6), resulting in most of the subjects having a negative best fitting drift rate in order to capture the RT data.

#### 4.3.1. Repulsion effects in a conflict task.

Next, we moved beyond the simple case in which  $X^i$  and  $X^j$  fell equidistant to the category boundary, using instead a more complex design where they could vary independently around vertical at  $\{15^\circ, 30^\circ, 45^\circ\}$ , and  $sX^j$  could once again vary at 3 levels  $\{0^\circ, 15^\circ, 30^\circ\}$ . In this case, the model makes several predictions, some of them highly counterintuitive. Firstly, it predicts that there should be no main effect of congruence on RTs. In other words, the predicted inverse decision values  $1/|X^i|$  (or equivalently, unscaled RTs) are indistinguishable when the target ( $X^i$ ) and flanker mean ( $\overline{X^j}$ ) are of the same and different sign. Secondly, the model predicts the existence of strong “reverse compatibility” effects under specific circumstances: there will be a disproportionate cost on congruent trials when the target  $X^i$  is closer to the category boundary than the mean of the flankers  $\overline{X^j}$ , e.g., when  $X^i = 15$  and  $\overline{X^j} = 30$  or  $\overline{X^j} = 45$  (Figure 4.6; A upper left corner of each plot), and this effect should diminish with increasing flanker variance (panels). Finally, although the model indicates that RTs will be dominated by the distance between  $X^i$  and the category boundary, it predicts that in the specific case where  $X^i$  and ( $\overline{X^j}$ ) are both close to vertical, this cost will be strongly attenuated.

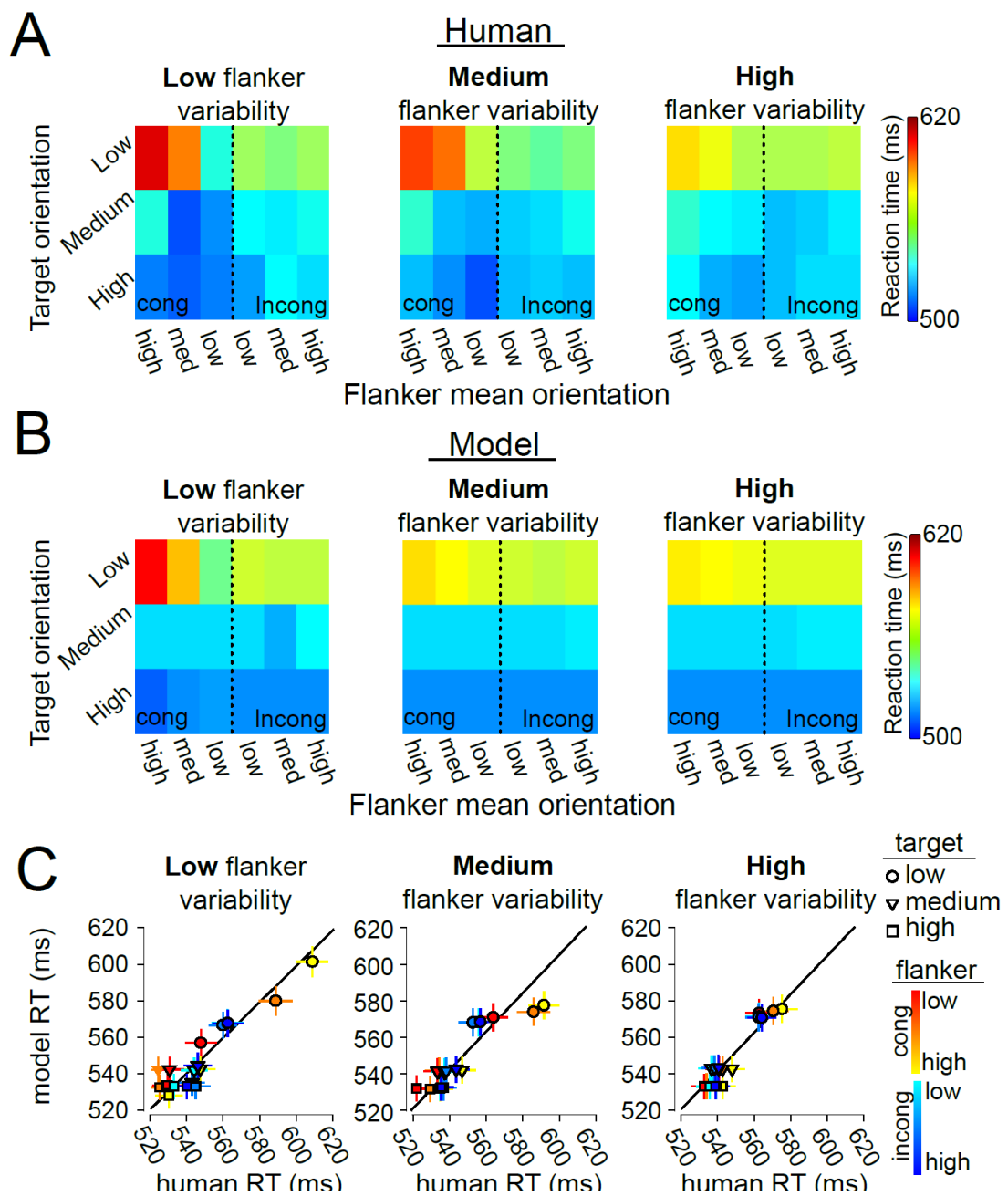


**Figure 4.6. Model predictions and human data for Exps. 2a and 2b.**

(A) Surface plots showing the mean RT pattern in humans under different conditions (3 levels of target orientation  $\times$  3 levels of flanker mean orientation  $\times$  congruency  $\times$  3 levels of flanker variability). Warmer colours correspond to longer RTs. There is an overall cost when the target is close to the category boundary (top row of each surface plot). There was an additional cost when these targets were flanked by congruent flankers that are further from the boundary (top left corners of each subplot). The introduction of higher flanker variability reduces these additional costs in those conditions (overall faster RTs across surface plots). (B) Fitted mean RT pattern from the adaptive gain model. (C) Mean ( $\pm$  SEM) RT in humans were cross-plotted against the fitted mean ( $\pm$  SEM) model

RT for each condition. Warm colours (Red to yellow) correspond to levels of mean orientation from congruent flankers. Cold colours (Blue to cyan) correspond to levels mean orientation from incongruent flankers  $\bar{X}^j$ s  $|\bar{X}^j|$ . Different shapes correspond to 3 levels of target decision variable  $|X^i|$ . This plot shows that the adaptive gain model captures not only the mean RT pattern in panel A but also the source of RT variability in humans.

We tested these predictions using the flanker paradigm on two new cohorts of participants, one of which (Exp. 2a,  $N = 28$ ; % error =  $3.76 \pm 3.9$  SD) performed the tilt categorisation task described above, except with the full 6 (target mean) x 6 (flanker mean) x 3 (flanker variance) design. Another (Exp. 2b,  $N = 30$ ; % error =  $8.47 \pm 5.58$  SD) performed a task with the same design that involved judging the colour of a central circle (red vs. blue) surrounded by distracting flankers that varied continuously in colour from red to blue. Results from the two experiments were qualitatively very similar and so after normalising the feature values (tilt, colour) to an equivalent scale in the range  $[-180, 180]$ , we collapsed over them for display purposes. All three model predictions were strongly present in the human RT data (Figure 4.6A-C; see Figure 4.5B for the fits of the conflict model). Firstly, splitting trials naively into compatible and incompatible on this novel variant of the Flanker task, there was no overall significant difference in RT, either in the combined cohort ( $p = 0.26$ ) or separately for each experiment (Exp. 2a:  $p = 0.79$ ; Exp. 2b:  $p = 0.17$ ). Secondly, we see a strong cost on congruent trials when the flanker mean is further from the boundary than the target mean, as demonstrated by an  $X^i \times X^j$  interaction (Exp. 2a:  $F_{3.56, 92.54} = 17.38$ ,  $p < 0.001$ ; Exp. 2b:  $F_{3.64, 105.65} = 7.67$ ,  $p < 0.001$ ), but not on incongruent trials (Exp. 2a:  $p = 0.26$ ; Exp. 2b:  $p = 0.23$ ). This difference was qualified by a reliable three-way  $|X^i| \times |\bar{X}^j| \times congruence$  interaction on human RTs (Exp. 2a:  $F_{3.42, 88.88} = 10.18$ ,  $p < 0.001$ ; Exp. 2b:  $F_{3.64, 105.64} = 6.48$ ,  $p < 0.001$ ). The analyses described thus far pertain to RT data. In a final analysis of Exp. 2 we examined choices, using a previously-described approach based on probit regression to assess the weight (or influence) that distracters wielded over choices, as a function of whether individual flankers' tilt was similar or dissimilar to the target (de Gardelle & Summerfield, 2011). The adaptive gain model predicts a greater impact of flankers that are moderately

dissimilar to targets compared to the case when they are highly dissimilar to the target (Figure 4.7), capturing the human data as well as replicating the reported influence of surround tilt on the perception of the central tilt from previous tilt illusion studies (Solomon & Morgan, 2006, 2009).

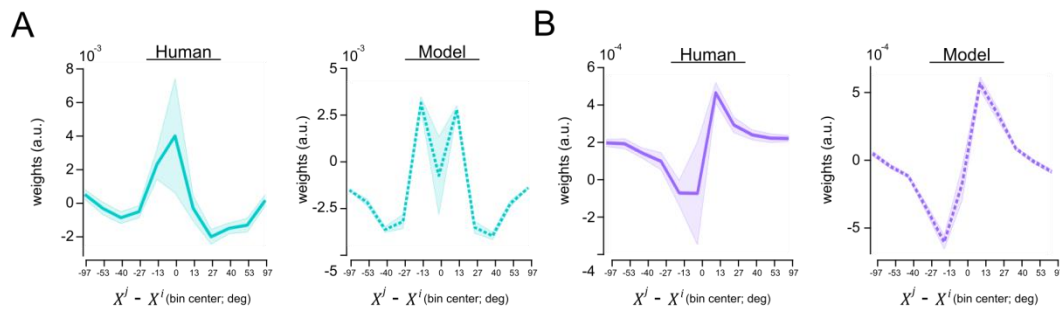


**Figure 4.7. Choice bias driven by the flankers.**

The angular difference between individual flankers and the target tilts  $X^j - X^i$  were being sorted into 11 bins ( $q$ ) from most CCW (i.e., the target is more CW than the flankers) to most CW (the target is more CCW than the flankers) and used for predict choice (panel A; using probit regression) and response times (panel B) for both humans (solid lines) and the best-fitting model (dotted lines). The x-axis shows the centre of each tilt bin, and the y-axis shows the resulting coefficients. The inverted-U-shape pattern in Panel A reveals that flankers that are moderately different from the target have a stronger impact on choice than when the flankers are dissimilar to the target for both humans and model. The response time data reveal the repulsive nature of this effect (Panel B solid purple line; note similarities with Figure 4.1A. Higher absolute magnitude of beta weights implies that moderately dissimilar flankers are more predictive of RT than more dissimilar flankers, similar to what was shown on choice in panel A. We showed that flankers that are moderately dissimilar to the target (flankers that are  $13^\circ$  tilted CW or CCW from the target) have the greatest impact on choice and RT. In other words, these flankers exert a more distracting influence on choices. We observed exactly the same pattern shown in humans, as well as replicating the prediction of the model in Figure 4.1A. Together, these analyses show that the human data resemble the model predictions over 4 different experimental datasets in Exp.1 and 2. Shaded areas indicate SEM.

#### 4.3.2. Functional brain imaging.

Established theories propose that a brain network that prominently includes the dorsal anterior cingulate cortex (dACC) is involved in the recruitment of control processes that

allow the brain to overcome distraction. Across a range of paradigms including the Eriksen Flanker task, the dACC responds with higher-amplitude BOLD signals on incompatible than compatible trials (Carter et al., 1998; Van Veen & Carter, 2002), and this effect is accentuated when the previous trial was compatible ('conflict adaptation'), as if the dACC is monitoring for conflict and signalling its onset (Kerns et al., 2004). However, the dACC is also implicated in decision processes more generally. For example, it signals the level of noise that corrupts an imperative stimulus during perceptual discrimination (Ho, Brown, & Serences, 2009), and its proximity to a choice point or category boundary (Grinband, Hirsch, & Ferrera, 2006); it responds to the relative economic value of an unchosen to a chosen option (Juechems, Balaguer, Ruz, & Summerfield, 2017; Tsetsos, Wyart, Shorkey, & Summerfield, 2014; Wunderlich, Rangel, & O'Doherty, 2009), to the value of switching to a new task or context (Barber & Carter, 2004; Kolling et al., 2012), and to the update signals that occur as decision values change (Rushworth, Noonan, Boorman, Walton, & Behrens, 2011). The search for a unifying theory of the dACC has been one of the most challenging and controversial themes in cognitive neuroscience over recent years (Heilbronner & Hayden, 2016; Kolling, Behrens, Wittmann, & Rushworth, 2016; Shahnazian & Holroyd, 2017; Shenhav, Botvinick, & Cohen, 2013).

To assess the role of the dACC and interconnected regions in adaptive gain control, we conducted a new experiment in which  $X^i$ ,  $\overline{X^j}$  and  $sX^j$  varied parametrically from trial to trial, rather than in a condition-wise fashion. A new cohort of humans (Exp. 3;  $N = 20$ ) performed this task whilst we acquired BOLD signals from across the brain using functional magnetic resonance imaging (fMRI). Behavioural results of this experiment replicated those from Exp. 2 (Figure 4.8), and so we focused on neural analysis to test whether brain signals indexed decision information in a way that was predicted by the adaptive gain model. We began by confirming previous reports that the dACC responds more vigorously when a target feature lies closer to a category boundary, i.e., in our experiment, when the target orientation is closer to vertical (Grinband et al., 2006). We first regressed  $|X^i|$  (i.e., proximity of the target to the category boundary) alone against BOLD signals occurring at the time of choice across

the entire brain (GLM1). Consistent with previous observations, we observed a negative effect of  $|X^i|$  in the dACC (peak: 2, 8, 54,  $t_{19} = 9.06$ ,  $p_{\text{fdr}} < 0.001$ ), as well as the anterior insula (AIC; peak: 34, 24, 2,  $t_{19} = 10.19$ ,  $p_{\text{fdr}} < 0.001$ ;  $p_{\text{fdr}}$  denotes significance after correction for multiple comparisons using false discovery rate) and superior parietal lobe (SPL; peak: 22, -56, 46,  $t_{19} = 8.19$ ,  $p_{\text{fdr}} < 0.001$ ; see Figure 4.9A). Extracting regions of interest from these areas in a leave-one-out fashion across participants (see Methods), we then plotted how the BOLD signal varied in quartiles of both (signed)  $X^i$  and  $(\bar{X}^j)$  (GLM2) and compared these signals to the predictions of (i) the adaptive gain model, (ii) an equivalent model with no adaptive gain, i.e., where all simulated neurons had identical tuning width, and (iii) a model in which BOLD signals were driven by conflict alone (Figure 4.9B). We found that the pattern of BOLD signals in all 3 regions closely resembled that predicted by the adaptive gain model, but not the other models (Figure 4.9C). Specifically, although BOLD responses were elevated when the  $X^i$  was close to zero (dACC:  $F_{1,19} = 52.37$ ,  $p < 0.001$ ; AIC:  $F_{1,19} = 53.4$ ,  $p < 0.001$ ; SPL:  $F_{1,19} = 48.94$ ,  $p < 0.001$ ), this effect was exaggerated on those trials where  $(\bar{X}^j)$  was far from zero but of compatible sign (i.e., greater BOLD response in dACC, AIC and SPL on congruent relative to incongruent trials; dACC:  $t_{19} = 3.03$ ,  $p = 0.0069$ ; AIC:  $t_{19} = 2.82$ ,  $p = 0.011$ ; SPL:  $t_{19} = 2.28$ ,  $p = 0.034$ ). No such modulation was observed when  $X^i$  was far from zero, as predicted by the adaptive gain model.

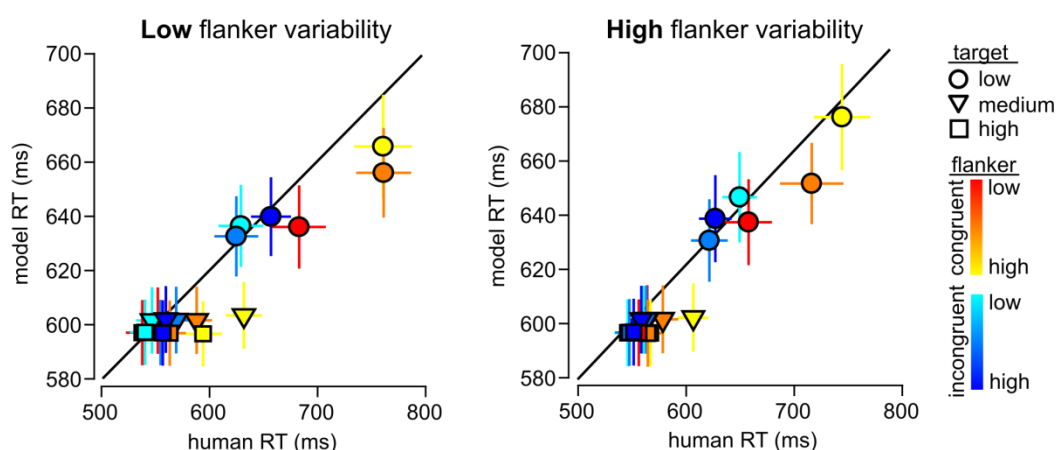


**Figure 4.8. 4 way interactions for Exp.3 (fMRI).**

The same cross-plots as Figure 4.6C

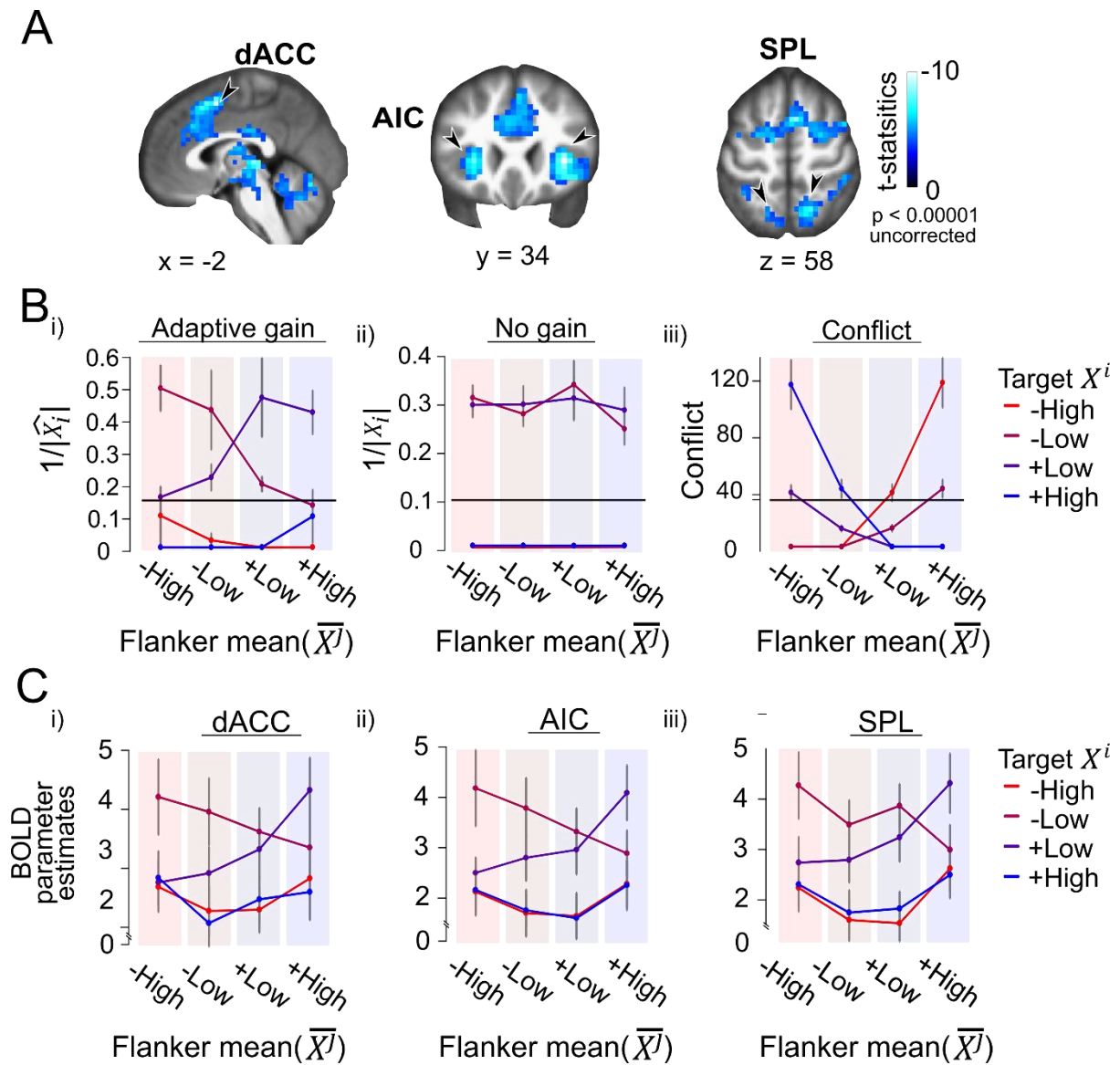


**Figure 4.9. Effect of target orientation on BOLD signal.**

(A) Brain areas correlating negatively with the absolute target decision variable, rendered onto a template brain in sagittal (left panel), coronal (middle) and axial (right) slices. Images were generated with an uncorrected threshold of  $p < 0.0001$ . (B) Parametric modulators  $X^i$  and  $\bar{X}^j$  were each being split into 4 quartiles (lines for  $X^i$ , x-axis or shaded area for  $\bar{X}^j$ ). Mean predictions ( $\pm$ SEM) from different models: Reciprocal gain-modulated decision variable  $1/\hat{X}^i$  (left panel). Reciprocal no-gain modulated target decision variable  $1/\hat{X}^i$  (middle panel). Estimated conflict was plotted for from each quartile bins (right panel). Coloured lines correspond to 4 levels of  $X^i$ . Shaded coloured background (groupings on x-axis) corresponds to 4 levels of  $\bar{X}^j$  (C) Mean BOLD ( $\pm$ SEM) signal beta values for each level of  $X^i$  and  $\bar{X}^j$  from the quartile bins. Left: dACC, Middle: AIC, and Right: SPL ROIs.

This suggests that a gain-modulated decision variable, rather than a conflict signal *per se*, is driving the dACC response. However, to quantify and compare the predictions of different models, we used Bayesian neural model comparison (see Methods) (Stephan et al., 2009). We fit the adaptive gain model and the rival conflict models on the trial RTs. Model estimates  $X^i$  and conflict computed in different ways (see Methods) from the best fitting parameters are then used to estimate BOLD signals. We computed, within the dACC, AIC and SPL, the posterior probability of the adaptive gain model conditioned on the BOLD signal using Random effects Bayesian model selection (Daunizeau et al., 2014), and compared the resulting estimates to those obtained for rival models. Both the exceedance probabilities and the expected frequencies strongly favoured the adaptive gain model over a decision model with no gain modulation, as well as over a family of conflict models (exceedance probabilities for the adaptive gain model in dACC: 0.992; AIC: 0.994; SPL: 0.996; expected frequencies for the adaptive gain model compared to chance level in dACC:  $t_{19} = 4.11$ ,  $p < 0.001$ ; AIC:  $t_{19} = 4.11$ ,  $p < 0.001$ ; SPL:  $t_{19} = 4.76$ ,  $p < 0.001$ ; Figure 4.10A). In other words, the dACC, along with AIC and SPL, code for a decision signal modulated in precisely the fashion predicted by the adaptive gain model.

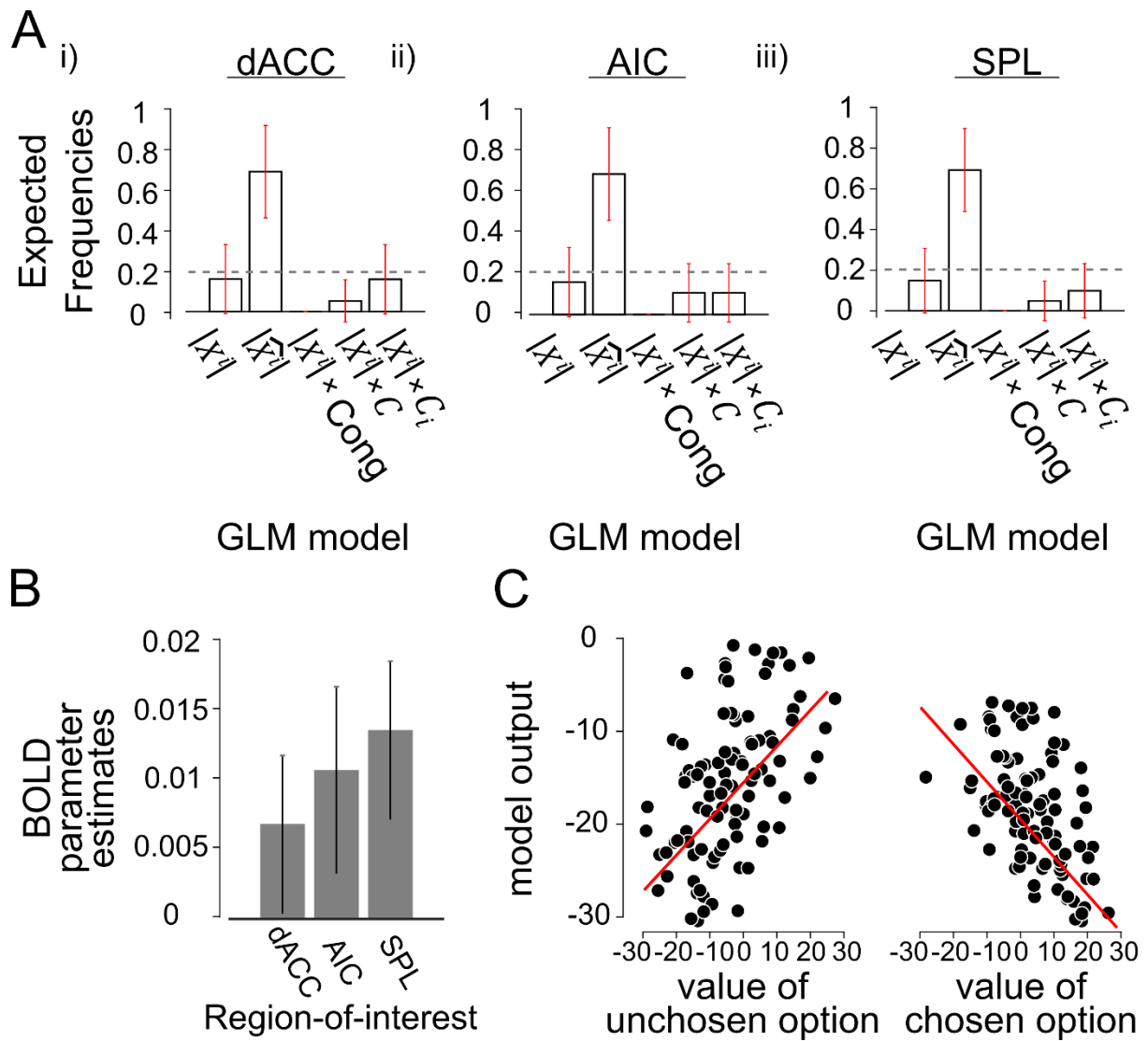


**Figure 4.10.**

(A) Model comparisons across GLM models. Expected frequencies for each GLM and each ROI were computed using Random effects Bayesian model selection. Grey dotted line displayed the chance level. Models compared: no-gain modulated target decision variable ( $|X^i|$ ), gain modulated target decision variable ( $|\widehat{X}^i|$ ), no-gain modulated target decision variable independently with three conflict models: binary classification of congruency - Cong, Conflict model 1 ( $C$ ), Conflict model 2 ( $C_i$ ); see Methods. (B) ROI analysis on dACC, AIC, and SPL on the absolute flanker mean decision variable  $|\overline{X^j}|$ . Bar plots correspond to the mean ( $\pm$ SEM) beta estimates for this predictor from GLM4. (C) Simulated (negative) model output as a function of the value of chosen and unchosen option (see Methods for details).

Finally, we addressed a concern that dACC is simply exhibiting a BOLD signal that correlates with the response production time on each trial (Grinband et al., 2006). Disentangling these factors is challenging, because (as described above) the model does an excellent job of predicting RTs. Nevertheless, when we included both model output  $1/|\widehat{X}^t|$  and RTs as competitive predictors in the model (GLM3), we still recovered a significant activation in the dACC (peak: 6, 20, 50,  $t_{19} = 4.47$ ,  $p_{\text{fdr}} = 0.046$ ), and AIC (left peak: -30, 16, 10,  $t_{19} = 5.96$ ,  $p_{\text{fdr}} = 0.038$ ; right peak: 34, 24, 6,  $t_{19} = 5.05$ ,  $p_{\text{fdr}} = 0.04$ ; Appendix Table 3.3). In other words, the dACC BOLD signal correlates better with the demand predicted by the adaptive gain model than it does with time taken to produce a response on each trial.

How does our model explain previously-reported findings, such as the observation that the dACC responds to conflict (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999), or to the relative value of an unchosen vs. a chosen option during economic choice (Boorman, Behrens, Woolrich, & Rushworth, 2009)? We have already shown that in the simple version of the Flanker task (c.f. Exp. 1a where  $sX^j = 0$ ), the model predicts a larger output  $|\widehat{X}^t|$  on compatible relative to incompatible trials. The model is thus in clear accord with a large literature indicating that dACC BOLD increases when target and distracters are incongruent in a simple version of the Flanker task (Hazeltine, Poldrack, & Gabrieli, 2000). We note that as described here, the adaptive gain model computes decision values independently on each successive trial, and thus in its current form would not predict conflict adaptation in the dACC. However, one could reasonably assume that adaptive effects may spill over from one trial to the next, i.e., that neural tuning width will be partly modulated by the previous trial. Under this assumption, the adaptive gain model will successfully predict that responses should be faster on two successive incongruent or two successive congruent trials (Gratton, Coles, & Donchin, 1992), just as it successfully account for the observation that during categorisation of a multi-element array, response times are faster if the target array is preceded by a prime array with an equivalent level of feature variance .

However, we also note another facet of our results: that BOLD signals in the dACC, AIC and SPL ROIs correlate negatively with  $|X^i|$  but positively with  $|X^j|$  (GLM4; Figure 4.10b;

dACC:  $t_{19} = 2.27$ ,  $p = 0.035$ , AIC:  $t_{19} = 3.03$ ,  $p = 0.003$ , SPL:  $t_{19} = 4.56$ ,  $p = 0.007$ ; this effect was also significant at the whole-brain level in voxels within the AIC and SPL, but not dACC; Table 3.5). If we consider the target to be a “chosen” option and the flankers as a competing, “unchosen” option, the ensemble of findings reported here is reminiscent of the well-described finding by which dACC signals scale positively with the decision value associated with an unchosen option (i.e., the flankers) and negatively with the value of a chosen option (i.e., the target). Building on this intuition, we tested more directly the coding of model-predicted value of a chosen and unchosen option in a further simulation in which decision values for two stimuli were drawn randomly and independently from two distributions, and model output was converted to a choice via a softmax function (see Methods). This allowed us to correlate model output (i.e., predicted BOLD) with the value of the chosen and unchosen option, revealing a negative correlation with the former and a positive correlation with the latter, as previously reported (Figure 4.10c). Our model thus unifies a number of disparate accounts that have emphasised a role for the dACC in tasks involving categorising perceptual stimuli, and choosing among economic prospects.

#### 4.4. Discussion

Good choices are based solely on information that is relevant to the choice at hand, and rational agents will successfully ignore distracting signals when making decisions (Luce, 1959). However, across perceptual, cognitive and economic domains, human participants are observed to deviate from this rational principle. A range of different theories have been proposed to account for human susceptibility to distraction, but thus far, no single model has emerged that can account for phenomena as diverse as visual illusions, susceptibility to conflicting contextual features, or economic decoy effects for single attribute stimuli (see Chapter 4 for the discussion of multiattribute case). Here, we describe one such account. An adaptive gain model has previously been shown to successfully account for diverse contextual influences of relevant decision information in perceptual decision-making, including

confirmatory biases in sequential sampling (Cheadle et al., 2014) and priming by second-order summary statistics in perceptual categorisation (Michael et al., 2014). Here, we show that not only can it account for classical effects of distraction across perceptual, cognitive and economic domains, including the tilt illusion and decoy effects for single attribute stimuli. Moreover, our model successfully predicts a range of previously unreported, counterintuitive findings in a well-studied cognitive paradigm, the Eriksen flanker task (Eriksen & Eriksen, 1974).

The effect of distraction is most often modelled under the assumption that irrelevant features are imperfectly filtered during decision-making, driving residual activation that corrupts decisions. When target and distracters prompt conflicting sensorimotor responses, the resulting competition slows response times and increases error rates (Botvinick et al., 2001). A very successful neurocognitive theory proposes that dedicated processing systems have evolved in the primate medial prefrontal cortex that detect this conflict, and that are responsible for mobilising control mechanisms (associated with the lateral prefrontal cortex) to help mitigate the resulting costs (Egner & Hirsch, 2005). Here, we propose a differing view: one that emphasises the benefits of consistent context rather than the costs of inconsistent context. In the adaptive gain model, contextual features offer guidance as to where to best allocate gain across feature space, ensuring that neurons that code for the most prevalent (or “expected”) features have the sharpest tuning and thus provide the most sensitive outputs. The adaptive gain model is thus motivated by the more general view that the nervous system has evolved to code efficiently for sensory inputs, reducing redundancy by dynamically adjusting the tuning properties of decision-relevant neurons to maximise sensitivity to expected features (Barlow, 1961; Simoncelli & Olshausen, 2001). In the flanker task, thus, the difference between compatible and incompatible trials arises at least in part because of a contextual facilitation mechanism at the decision level, akin to that described in sensory circuits (Stemmler, Usher, & Niebur, 1995), rather than because of an active cost of response competition. This idea is not without precedent in theories of control. In fact, the notion that a flexible response to stimulus conflict is dependent on adaptive expectation mechanisms

dates back to the original discovery of conflict adaptation by Gratton and colleagues more than 25 years ago (Gratton et al., 1992). We take the opportunity to highlight two major features of our behavioural data that cannot be accounted for by standard accounts that emphasise the cost of conflict alone. Firstly, in Exp. 1a-b, we found that a low-variance flanker array hastens response times on congruent trials, rather than prolonging response times on incongruent trials. This is consistent with an account that emphasises the benefit of consistent context rather than the cost of inconsistent context. Secondly, in Exp. 2a-b, we observed that the longest response times were in fact observed on compatible trials, not incompatible trials. We replicated this finding across two different classes of visual feature: tilt and colour. According to our model, this cost occurred when the gain field dictated by the context repulsed the target subjectively closer to the category boundary, rendering choices more uncertain. Although such negative flanker effects have been reported with heavily masked stimuli, where they can be explained by differing time courses of facilitatory versus inhibitory processes (Eimer, 1999), only rarely have such phenomena been reported for fully visible targets and distracters such as ours. Most interestingly, one such report occurred for a modified version of the flanker task where the targets were letters that were parametrically morphed between two possible identities, each corresponding to a possible flanker (Rouder & King, 2003). This report describes negative flanker effects when the target is most ambiguous, precisely paralleling our findings here for trials with small  $X^i$  and large  $\overline{X^j}$ , and a shift in the psychometric function that occurs with flanker identity in precisely the fashion predicted by our adaptive gain model (Cheadle et al., 2014).

Our behavioural findings were echoed in the neural data recorded from dACC, where BOLD signals were higher when targets fell closer to the category boundary, but these signals were positively modulated (yet higher) when the distracters mean was congruent but further from the boundary. Without further assumptions, a model based on conflict alone cannot account for these findings. We do not wish to argue that stimulus or response conflict do not ever incur an additional cost to accuracy and response times, or that such a cost is unable to drive the dACC. Nevertheless, in the current study, we found that such an account was not

required to explain our data, and that a model embodying this assumption fit our data more poorly.

Our findings present a challenge to some extant theories, but we acknowledge that our model is currently incomplete. For example, without further elaboration, our model cannot account for the previously-described below-chance responding on the flanker task that is observed under strong speed pressure (Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988). Furthermore, a large literature implicates the dACC in the mechanisms by which we update the value of actions in dynamically changing environments (Behrens, Woolrich, Walton, & Rushworth, 2007; Rushworth et al., 2011). Our experiments were conducted in stationary settings, and we do not doubt that these regions may play additional roles (potentially also related to gain control) when slower learning about a changing context is required. We also note an important shortcoming in our findings: we were unable to identify differing roles for the dACC, AIC and SPL, which seem to act as one in our study. We think it is likely that our BOLD data are simply indexing the output of a decision process that involves modulation by distracting context, but are unable to make strong claims about the interim processes by which the computations proposed by the model occur. We suspect that exploring the role of adaptive gain control in dynamically-changing environments may shed more light on the differing contributions made by these regions, and we hope to pursue this question in future studies.

## **Chapter 5 - Compare and contrast biologically-plausible explanations on multiattribute multialternative choice**

### **5.1. An overview of multialternative decisions**

Let us consider an everyday real-life decision on what to have for lunch. There could be several options available in a Cafe: a light sandwich meal, a standard main course, a vegetarian main course, or just a soup. Unlike binary decisions that have been studied extensively in a well-controlled laboratory setting, real-life decisions often involve comparing multiple alternative options. We are not only faced with many alternatives to choose from, but we also need to compare and integrate features and attributes for each option to get a good idea of how good an option is. For instance, a foraging animal evaluates fruit based on their colour and size, or a consumer compares products with differing price and quality. Normative choice theories propose that the preference toward an option should be independent of other less-preferred options in the choice set and independent of the context which this option is presented (Independence from Irrelevant Alternatives Axiom; IIA; see Chapter 1; Luce, 1959; von Neumann & Morgenstern, 1944). A rational agent that maximises their utility would make choices based on the relevant options and ignore irrelevant options.

Studies on decoy effects have featured the three types of canonical decoy effects in humans, have found individual differences in decoy effects, or have found an effect of time pressure on decoy effects. Computational models of decoy effects should be equipped with the ability to explain all of these effects. Therefore, in the following sections, I will first give an introduction on studies that investigated the three types of canonical decoy effects, then followed by discussing individual difference in decoy effects and the effect of time pressure on decoy effects. In the next section, I will talk about the computational models that have shown success in capturing in some of these effects.

### 5.1.1. Decoy effects

Researchers have discovered a predictable violation of IIA in multi-attribute decision making across species and experimental settings. Both humans and animals display preferences for one of two otherwise equally preferred options when a third “decoy” stimulus is introduced, even when the decoy is less attractive or unavailable (Cohen et al., 2018; Lea & Ryan, 2015; Louie et al., 2013; Shafir, Waite, & Smith, 2002; Trueblood, Brown, & Heathcote, 2014). For instance, the choice between a powerful but more expensive laptop computer  $Y$  and a less powerful but more economical model  $X$  can be systematically biased toward  $X$  by the presence of a third option  $Z$  that is either even less powerful and more economical (compromise effect) (Simonson, 1989), similar in power and price to  $Y$  (similarity effect) (Tversky, 1977), or that is less powerful but more expensive than  $X$  (attraction effect) (Huber, Payne, & Puto, 1982). These decoy stimuli can be described as points in a multi-dimensional (i.e., attribute) space. Similarity in values between options can be seen as the geometric distance between options in a multidimensional space. Figure 5.1 illustrates the configuration of decoys that were commonly used to test the aforementioned decoy effects.

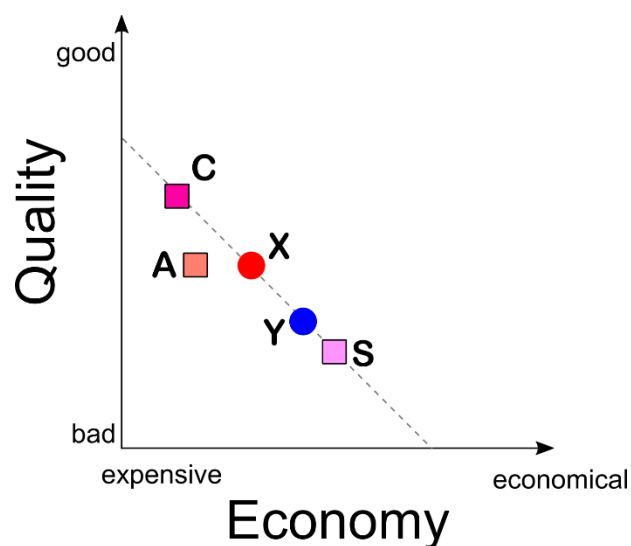


**Figure 5.1. An illustration of decoy effects on a two-dimensional spatial space**

An option can be characterised by multiple attributes which can then allow us to describe options in a multidimensional space. The figure shows options that have two attributes: economy (x-axis)

and quality (y-axis). Option X and Y (in circles) are the relevant options. They both lie on the preference-indifference-line (dotted line): options falling on this line are equally preferred when presented in a binary choice set. Other options: option C, A, S (pink squares) are decoys that correspond to a **C**ompromise, **A**ttraction and **S**imilarity decoy respectively. The presence of these decoys in a choice set would boost the preference for option X.

### 5.1.2. Individual differences on the impact of context.

Until recently, most of the empirical studies on decoy effects had studied these effects in isolation and often adopted a between-subject design. Trueblood and colleagues adopted a within-subject design and showed that all three decoy effects could be displayed within a single participant using both a perceptual task and an inference task (Trueblood et al., 2014). Berkowitsch and colleagues subsequently conducted a replication of these findings using a preferential task (Berkowitsch, Scheibehenne, & Rieskamp, 2014). They found that while the three decoy effects could co-exist within a single subject (i.e. replicating the findings of Trueblood et al., 2014). They also found that most subjects showed a stereotypical choice pattern that resulted in an intercorrelation of the effects at the cohort level (Figure 5.2). In that task, subjects were asked to make trinary choice judgements. Figure 5.2A shows the definitions of options. Each choice set contained two equally preferred options and a decoy, which should increase the favourability of the target option but not the competitor option (option A is the target option in the example of Figure 5.2A). Next, they computed the proportion of times participants selected the target option relative to the competitor options for each decoy trial type (i.e., the ‘relative choice share’; RCS). Figure 5.2B-D shows the RCS for each decoy type against each other. Each circle corresponds to the RCS for each participant. The results show that at the group-level, there is a positive correlation between the attraction and compromise effect but a negative correlation between the similarity and both the attraction and the compromise effect.

Liew and colleagues conducted a direct replication of an inference trinary choice task by Trueblood (2012) and the preferential choice task by Berkowitsch et al. (2014),

using online subjects. They failed to replicate the attraction effect from the Trueblood study. They then performed nonparametric clustering of individuals' choice patterns from the replication of the Berkowitsch et al. study and found that there are individual variabilities on the choice pattern across the cohort. In their data, around half of the subjects show all three decoy effects. Half of the subjects show weak attraction effects along with a reversed similarity effect and a reversed compromise effect (i.e. a compromise decoy that now favours the competitor option and a similarity decoy that now favours the target option). Some recent studies showed that decoy effects did not exist in specific tasks (e.g. Frederick, Lee, & Baskin, 2014) or that the effect was reversed (Cataldo & Cohen, 2018a; Spektor, Kellen, & Hotaling, 2018).

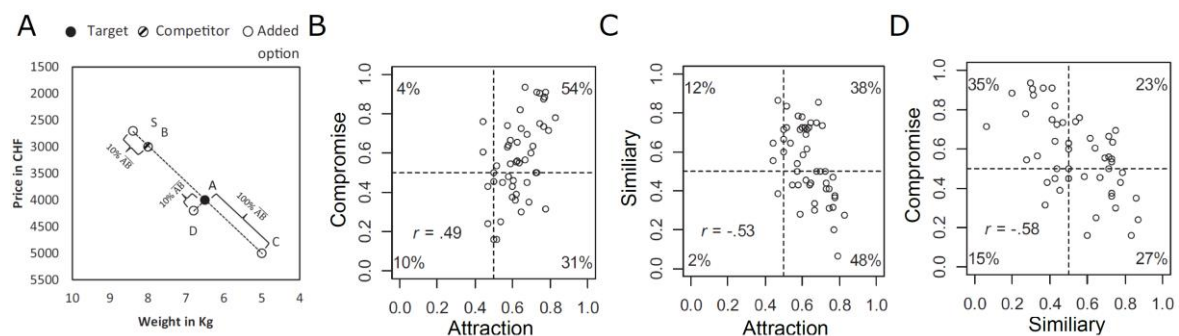


**Figure 5.2. Intercorrelation of decoy effects reported by Berkowitsch et al. 2014**

Figure copied from (Berkowitsch et al., 2014)(A) Definitions of decoys in the choice task. The decoy option (open circles) were carefully selected as a function of the distance between option A and B. (B-D) Each panel corresponds to a pairwise scatter plot of the relative choice share of the target under different decoy task types where each circle corresponds to a subject. The effect of time pressure on decoy effects.

An additional variable, namely time pressure, provides new insights on the nature of decoy effects because it decorrelates the predictions of alternative models. As mentioned in Chapter 1, Kahneman and colleagues considered violation of rationality as a consequence of a cognitive system that has two separate processors, a fast and a slow processor (Kahneman & Egan, 2011). The fast processor often adopts heuristics in order to make snap decisions, which can explain why perception and some economic decisions are

context-dependent since contexts can serve as heuristics. It implies that limiting the time for people to think should induce stronger contextual (e.g. decoy) effects. However, according to evidence accumulation based models, preferences evolve as evidence accumulates, meaning that decoy effects should be stronger when people have more time to deliberate. Pettibone (2012) and Trueblood and colleagues (2014) tested the strength of decoy effects when participants are placed under time pressure. They found that decoy effects are weakened when participants have more time to deliberate (Figure 5.3). This contradicts with the predictions made by heuristics-based models.

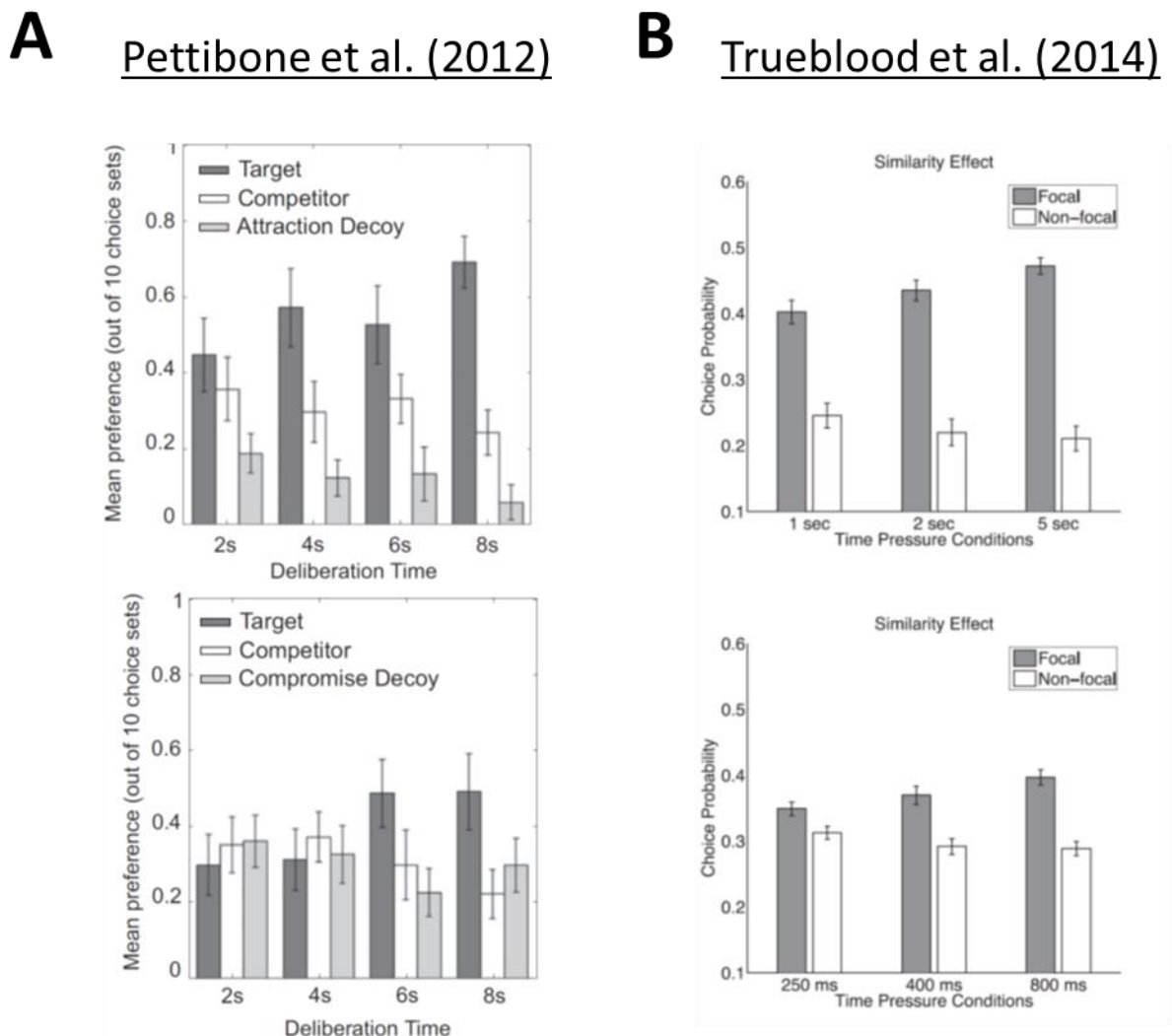


**Figure 5.3. Relative choice preference for the target, competitor, and decoy options under time pressure.**

(A) Figure copied from (Pettibone, 2012). Mean choice probabilities for the target (focal), the competitor (non-focal) and the decoy option (Attraction Decoy; top panel, Compromise Decoy; bottom panel) under different deliberation time conditions. (B) Figure copied from (Trueblood et al., 2014). Mean choice probabilities for the target (focal) and the competitor option (non-focal) of trinary choice tasks that contained only similarity decoys (inference task; top panel, perceptual task; bottom panel) under different time pressure conditions.

## 5.2. An overview of different classes of models in explaining context-dependent decisions

The findings described in the last section all highlight the need to have a computational model that can explain all of these paradoxical results, and more importantly, to conduct a within-subject study that investigates the decoy effect comprehensively within a large cohort. A variety of decision-making models strive to explain these context-dependent decoy effects. The goal of the rest of the chapter is to review some of the models, compare the qualitative predictions of those models, and provide quantitative fits on the behavioural predictions. In the following sub-sections, I will give a brief introduction of three classes of existing models that explain context-dependent decoy effects. The three classes of models include accumulator models (Section 5.2.1), “noise” models (Section 5.2.2), and models that are inspired by biological constraints (i.e. models that assume efficient computations; Section 5.2.3). Next, I will make a qualitative comparison of how different efficiency-based models predict decoy effects in a parametric two attribute space (Section 5.3). Lastly, I will describe a large-scale within-subject design behavioural study that investigated human choice patterns as a function of how the options are situated in a two-dimensional attribute space and show the quantitative fits of different models on this choice data (Section 5.4).

### 5.2.1. Explaining context-dependent decisions using diffusion processes

Accumulator models have shown success in explaining two-alternative forced choice tasks. These models also showed success in explaining decisions in the

multiattribute multialternative case by assuming stochastic selections of the active attribute during the integration process. Examples of models that assumed evidence accumulation processes are: the multiattribute decision field theory (Roe, Busemeyer, & Townsend, 2001), the multiattribute linear ballistic accumulator (Trueblood et al., 2014), and the multiattribute leaky competing integrator (Usher & McClelland, 2004). While these models are successful in explaining the decoy effects in human choice, these models make different or even opposite assumptions regarding the integration process, the relationship between the distance of a decoy to the target options, the strength of inhibition, etc. Accumulation-based models have been very influential in perceptual decision making since they provide an algorithmic explanation on how decisions evolves over time, and the model mechanism mirrors the neural dynamics during the decision-process. The recent success of these models in explaining value-based decisions and multiattribute multialternative choices have resulted in several recent reviews and model comparison work. These reviews aimed to differentiate between sequential accumulation models as well as evaluate their strengths and weaknesses in explaining the context-dependent effects and multi-attribute multi-alternative decisions (Busemeyer, Gluth, Rieskamp, & Turner, 2019; Hotaling & Rieskamp, 2018; Noguchi & Stewart, 2018; Turner, Schley, Muller, & Tsetsos, 2017). Despite this class of models is highly relevant for the topic I am investigating, I am placing a focus on models that assumed biologically plausible efficient computations. This is because there is not any extensive model comparison work on this type of models. Therefore, this chapter will not focus on sequential accumulation models.

### 5.2.2. Context-dependent decisions as a consequence of a noisy decision process

A different class of models proposes that context-dependent decisions can be explained by the imprecise representation of option values or the presence of noise during the decision process. For example, intransitive choices (an example of a transitive decision is that when one prefers A to B, B to C, she should also prefer A to C) can be understood as

the consequence of an adaptive computation. The adaptive computation “Selective Integration” assumes that the decision maker selectively integrates some pieces of information (e.g. the dominant or winning information) in the presence of noise that appears during the information integration stage (Tsetsos et al., 2016). Moreover, noise or imprecision can arise at different stages of the decision process (e.g., value encoding stage, information integration stage, choice stage), and they can have an independent impact on decision biases (Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016; Tsetsos et al., 2016). In the context-dependent decision literature, a few noise models have been proposed. Each model has different assumptions regarding the source of noise during the decision process. For example, before comparing options, one might need to first translate the value of an option into an expected utility. The computation of expected utility might be subject to inaccuracy. The logit and probit models can model this inaccuracy by adding a noise term to the value of an option. The two models differ in their assumptions on whether error or noise is additive (independent Gaussian noise; probit model; Equation 5-1) or multiplicative (higher noise for higher valued options; logit model; Equation 5-2) (Berkowitsch et al., 2014). A more elaborated model from Howes and colleagues showed that a model that assumed different distribution of the errors can replicate all three decoy effects. In particular, these errors can arise during both the computation of the expected utility and the computation of the ordinal ranking of options (Howes et al., 2016). Together, these models imply that noise that arises during the decision process can be a cause for the decoy effects. Without assuming any limitations in our information processing system (e.g., the presence of noise), we cannot conclude that humans are irrational.

$$U_i = V_i + N(0, \varepsilon)$$

Equation 5-1

$$U_i = V_i + V_i \cdot N(0, \varepsilon)$$

Equation 5-2

Both probit and logit models assume that the utility of an option  $i$ :  $U_i$  is a function of the subjective value of that option ( $V_i$ ) and some noise that is sampled from a zero-mean Gaussian distribution with a standard deviation of  $\varepsilon$ , i.e., it is sampled from  $N(0, \varepsilon)$ .

### 5.2.3. Efficiency-principled models in value-based decisions

One question that arises is: Why noise causes limitation in our information processing system? Unlike the theoretical agent assumed by most of the rational choice theories, humans or animals face neural constraints. For instance, not all neurons can be active at any time (Laughlin, 2001), neural activity is intrinsically variable (Dinstein, Heeger, & Behrmann, 2015), and the firing rate of different types of neurons is capped at a specific rate (Hasenstaub, Otte, Callaway, & Sejnowski, 2010). More recent work on context-dependent decisions has started to focus on the role of these neural constraints on valued-based choices. In other words, irrational decisions that are defined according to the normative choice theories might be a consequence of adaptive computations in the face of constraints from the neural system (Louie, Glimcher, & Webb, 2015; Summerfield & Tsetsos, 2015; Tsetsos et al., 2016).

Different neural computations were used to explain decoy effects (distractors that are defined by a single attribute or multiattributes) in a value-guided task. These computations include the divisive normalisation scheme (Landry & Webb, 2017; Louie et al., 2013) and the range normalisation scheme (Padoa-Schioppa, 2009; Soltani et al., 2012). Louie and colleagues (2013) showed that the simple distractor can be modelled by a normalisation algorithm which the value of the item in consideration is normalised by the summed value of all items on a given trial:

$$r_i = K \frac{V_i}{\sigma_H + \sum_j wV_j}$$

Equation 5-3

The mean firing rate for the option under consideration ( $r_i$ ) is a ratio between the monetary value of option  $i$  and the weighted sum from the values of all available options  $j$ . There are three free parameters in this model:  $K$ ,  $\sigma_H$ , and  $w$  which correspond to a gain term, a semi-saturation term and weight terms. This model successfully captures this monotonic effect of the distractor on choice, that is, both human choice data and the model prediction showed that when the distractor value approaches the value of target options, the low valued option competes with the distractor option, resulting in a higher probability of choosing the distractor compared to the low value option. Therefore, the relative choice probability for the high value option compared to the low-valued option increases.

Padoa-Schioppa proposed a different normalisation scheme. He proposed that the relative value coding scheme is based on the range of values being encoded (Padoa-Schioppa, 2009). In other words, the mean firing rate of a given option ( $r_i$ ) should be scaled by the difference in the mean firing rate for the highest valued option and that of the lowest valued option ( $V_{max} - V_{min}$ ) over a temporal period of time e.g., over a block or a session, rather than based on all available options:

$$r_i = \beta + \sigma_H \frac{V_i - V_{min}}{V_{max} - V_{min}}$$

Equation 5-4

Where  $V_i$  refers to the subjective value of the option  $i$ ,  $\beta$  refers to the baseline in the mean firing rate, and  $\sigma_H$  is the semi-saturation term. Together with his prior work, Padoa-Schioppa and colleagues showed that the orbitofrontal cortex (OFC) – a subdivision of the prefrontal cortex that is known for reward-guided decisions and proposed to be the key region for representing the value of goods – engages in range adaptation. They showed that the OFC neural activity for a given option decreases as the range of option values increases over a period of time, implying that the value codes in the OFC seem to adapt to

the recent value statistics provided by the context (where statistics refers to different ranges of value magnitudes).

One could consider that the brain might perform different forms of normalisation depending on what is required by the situation: divisive normalisation for modulating the neural activity associated with different reward statistics provided by the spatial context and range adaptation as the counterpart for reward statistics provided by the temporal context (Louie et al., 2015). Recent work shows that perhaps neurons in different sub-regions of the OFC might apply divisive normalisation for relative value codes (Yamada, Louie, Tymula, & Glimcher, 2018). It remained unexplored how the entire population of OFC neurons arbitrate between temporal and spatial contexts and what the role is of divisive normalisation in the parietal regions if value normalisation already happens in the OFC during the value encoding stage.

Soltani and colleagues proposed that instead of considering the minimum and maximum values as the scaling factor, relative value codes might depend on a more dynamic value range (Soltani et al., 2012). They showed that multi-attribute decoy effects could be explained by a new model that applies range normalisation based on a dynamic range of values in a choice set. In their range normalisation model, they assumed that the neural response ( $r$ ) can go from 0 to 1. The mapping of the subjective value of an option to the neural response depends on a threshold ( $c_t$ ) and saturation value ( $c_s$ ) i.e., the "dynamic range". Any subjective values below the threshold would result in a minimum neural response of 0 and any subjective values above the saturation point would only result in a maximum neural response of 1. Any subjective values between the threshold and the saturation value are linearly normalised into the range of 0 to 1 (Figure 5.4). To compute the threshold and the saturation value, the agent needs to know the value of the best ( $s_{max}$ ), the second best ( $s_{nmax}$ ), the worst ( $s_{min}$ ) and the second worst option ( $s_{nmin}$ ) from a given trial:

$$c_s = \begin{cases} \frac{s_{max} - f_s s_{min}}{1 - f_s} & , \text{if } f_s \geq 0 \\ s_{max} + f_s (s_{max} - s_{min}) & , \text{if } f_s < 0 \end{cases} \quad c_t = \begin{cases} \frac{s_{min} - f_t s_{nmin}}{1 - f_t} & , \text{if } f_t \geq 0 \\ s_{min} - f_t (s_{nmin} - s_{min}) & , \text{if } f_t < 0 \end{cases}$$

Equation 5-5

Where  $f_t$  and  $f_s$  are two free parameters that can be in the range of  $[-1, 0.999]$  and the sum of the two parameters must be smaller than 1. Although the model was formulated as if there were four options in the choice set to assign preferences to, they assumed that in the case of binary and trinary choices, some of the parameters from Equation 5-5 have shared values. In the case of binary choice, both the best option and the second worst option refer to the higher value option, and both the second best and the worst option refer to the lower value option. In the case of trinary choice, the second best and the second worst refer to the low value target option. Lastly, unlike the range adaptation scheme, this model implied that the observer makes a ranking judgment on the choice set, consistent with the Range-Frequency theory proposed by Parducci (1965).

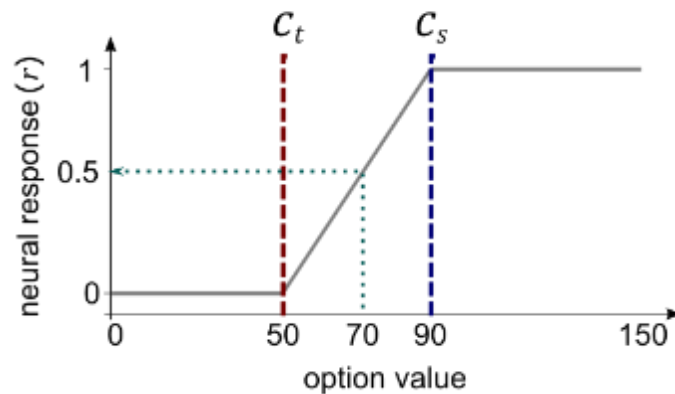


**Figure 5.4. An illustration of the range normalisation model from Soltani et al. (2012)**

The figure shows the mapping between the option value and the neural response given the threshold ( $c_t = 50$ ) and the saturation value ( $c_s = 90$ ). Option values that are below the threshold would elicit the same level of neural response ( $r = 0$ ). Options values that are above the saturation value would have a plateaued neural response ( $r = 1$ ). Option values that are within the dynamic range (between  $c_t$  and  $c_s$ ) are linearly mapped onto the neural response. For example, an option value of 70 results in  $r = 0.5$ .

As discussed in the previous chapter, an adaptive gain control mechanism can serve as an alternative explanation for context-dependent decisions across perceptual, cognitive and value-based tasks. The key idea of the model is that our brain engages in an adaptive gain process whereby decision information that is consistent with the context is represented with increased sensitivity; context can be defined by both the behavioural relevance and the statistical properties of the environment. The increase in sensitivity can be modelled by shifting the psychometric transfer function (equivalent to shifting the gain field) so that the peak of the gain field (for gain modulation) is maximal for decision information that is consistent with the context. In the population coding form of the model, I assume that neurons that preferably encode context-consistent decision information change their tunings as a form of increased sensitivity. Thus, I assumed a sharpening process in our model specifically, i.e. the area under each tuning curve is summed to unity. In the previous chapter, the population coding model consists of three parameters: maximum tuning width ( $\sigma_{max}$ ; a free parameter), tuning width envelope variance ( $\epsilon$ ; a free parameter) and a scaling parameter of the tuning width envelope ( $\tau$ ; fixed in value). All three parameters jointly control the degree of sharpening in the population codes (i.e., control the degree of context-dependencies of the decision-maker). The tuning width of each neuron  $k$  was defined as follows:

$$\sigma_k = \sigma_{max} - f(V_k|C, \epsilon)\tau$$

Where  $f(\cdot)$  is a probability density function of the standard normal.  $V_k$  refers to the value or decision value a neuron most preferred,  $C$  refers to the value of a distractor in this case. Lastly, to avoid having a  $\sigma_k$  that is 0, I rectified the tuning widths so that  $\sigma_k$  that fall below 0.1 will be corrected as 0.1. This model successfully captured all three decoy effects (Figure 5.5), the intercorrelation between decoy effects across the cohort (Figure 5.6), and the effect of time pressure (Figure 5.7) as I have reported in Li et al. (2018;). For the simulation on the interdependence in attraction, compromise, and similarity effects. I created a cohort of 40 simulated participants which were defined by differing the level of context sensitivities (i.e. varying the tuning selectivity parameter in the model). I then

simulated three choice tasks each with a decoy type in the position shown for A, C, and S in Figure 5.6E. The decoy values were chosen in the same way as reported by Berkowitsch and colleagues. I computed the choice probability using the subjective estimated value difference between the target and the competitor, and plotted the choice probabilities of each decoy task against each other to visualise the correlation and anti-correlation across subjects. This visualisation showed that the choice and intercorrelation pattern displayed in the Berkowitsch study can be explained by individual differences in context-sensitivity. To approximate the influence of time pressure on decoy effects, I simulated the decoy effects using the adaptive gain model by assuming that faster responses have a lower signal-to-noise ratio at the decision formation stage (Figure 5.7). I can recreate the attraction and compromise effect choice patterns under time pressure shown previously by Pettibone (2012), as well as the similarity effect during inferential and perceptual tasks reported by Trueblood and colleagues (Trueblood et al., 2014).

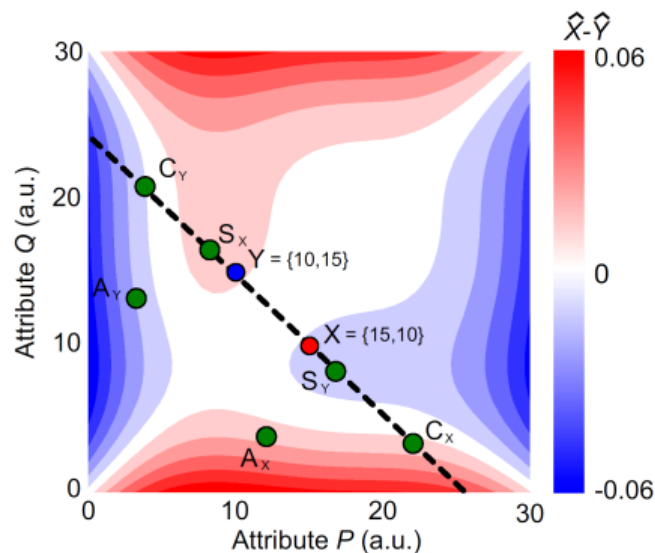


**Figure 5.5. A simulation of the decoy effects using the adaptive gain model.**

Figure copied from Li et al. (2018). The value difference between two equally preferred options (X and Y, red and blue dots) as a function of a third decoy option, Z (x and y axes). The axes correspond to the attribute values (e.g., inverse price P and quality Q). X and Y fall on the indifference line (dashed line). The red dot signals option X (where  $X_P = 15$ ,  $X_Q = 10$ , and the blue dot indicates option Y (where  $Y_P = 10$ ,  $Y_Q = 15$ . The colored surface shows the model-predicted

subjective value difference  $\hat{X} - \hat{Y}$  that is, the extent to which  $X$  is preferred over  $Y$  (red regions), and conversely the blue region in which the subjective value of  $X$  is smaller than  $Y$ , and thus  $Y$  is preferred. I have superimposed example decoy options (green dots) that produce the three context effects (compromise,  $C$ ; attraction,  $A$ ; and similarity,  $S$ ; subscript corresponds to a preference for either option  $X$  or option  $Y$ ). a.u., arbitrary units.

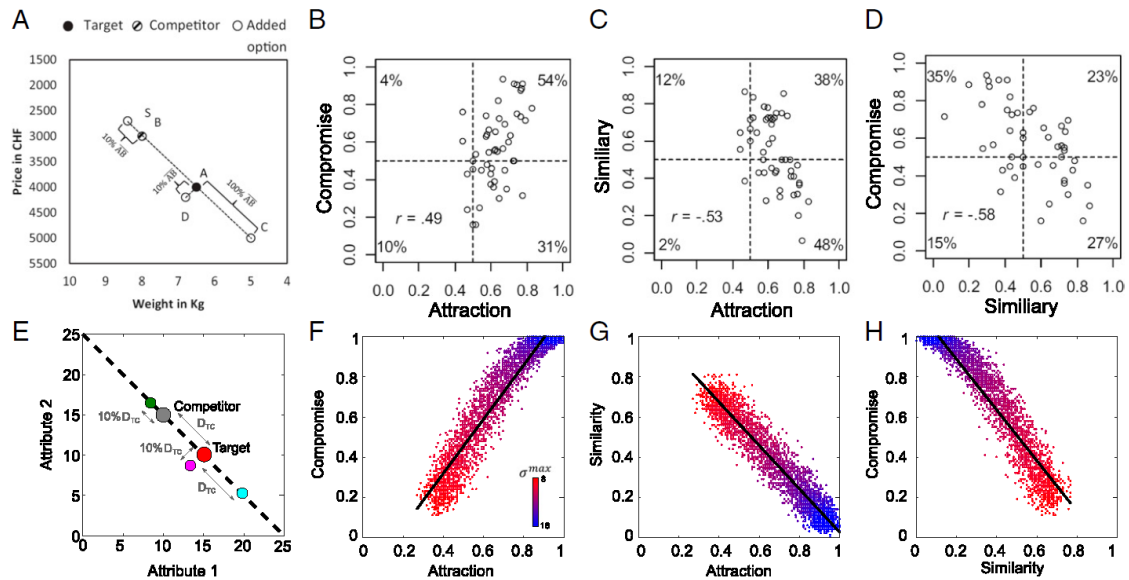


**Figure 5.6. Simulated intercorrelation of decoy effects using the adaptive gain model**

Figure copied from Li et al. (2018) (A-D) Repasting Figure 5.2 here. (E) An illustration of the simulated 2D attribute space with target and decoys. The target (red dot) is set to be {15,10}, and the competitor (gray dot) is {10,15}. Both lie on the indifference line (dashed line). The three types of decoy and their values are plotted in three different colored circles (green: similarity decoys, value = {8.5,16.5}; magenta: attraction decoy, value = {13.5,8.5}; cyan: compromise decoys, value = {20,5}). DTC refers to the distance between the target and the competitor. (F-H) Pairwise correlations among the strength of effect for each decoy type across the simulated cohort. Each unique colored dot corresponds to one simulated participant. The color corresponds to the tuning width  $\sigma_{max}$  (red = low tuning width  $\sigma_{max}$ , blue = high tuning width). The black line on each panel corresponds to the best least-squares fit. F shows a positive correlation between the compromise effect and the attraction effect across the population. G and H show a negative correlation between the similarity effect and the attraction or the compromise effect across the population. Across panels, we can see that participants who display a strong attraction effect also display a strong compromise effect but a weaker similarity effect, whereas those who display a weaker

similarity effect display stronger compromise and attraction effects. This is consistent with the data displayed in B–D from Berkowitsch et al.

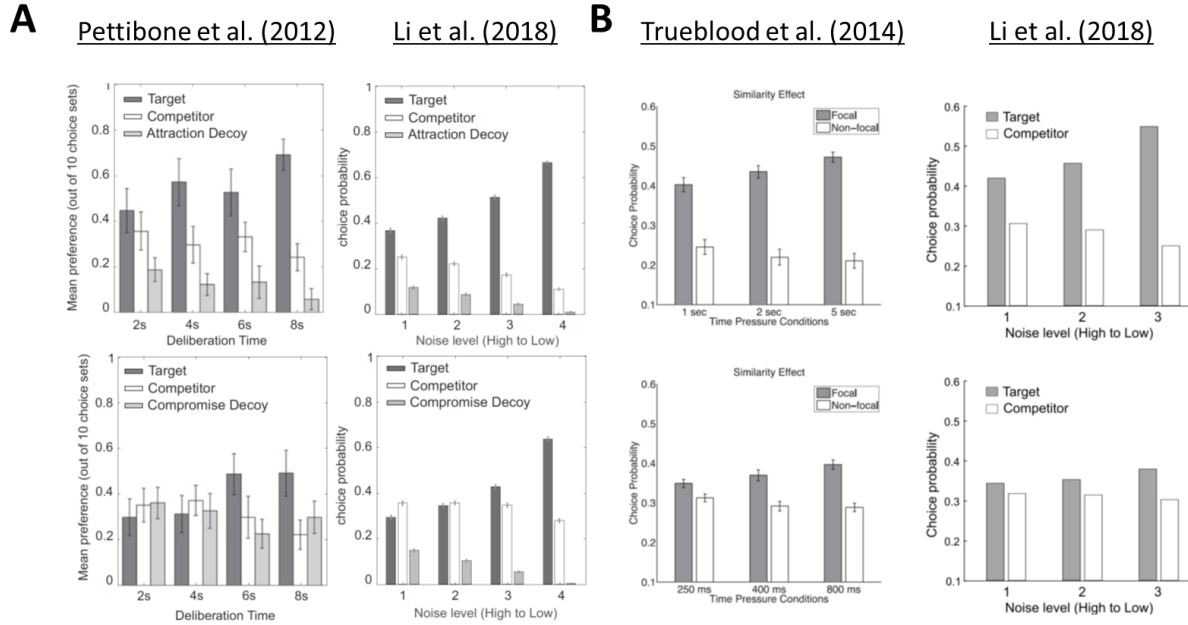


**Figure 5.7. Relative choice preference for the target, competitor, and decoy options under time pressure.**

Figure copied from Li, Michael, Balaguer, Castañón, & Summerfield (2018). The target (focal) and competitor (non-focal) options is present along with a decoy. (A & B) Left panels: repasting Figure 5.3A and B here; right panels: Simulations from the adaptive gain model. Decreasing noise (comparable to increasing deliberation time) would increase the preference for the target option relative to the competitor option.

One caveat on this formulation of the model is that the degree of sharpening does not scale monotonically with the parameters, i.e., parameters can trade-off with each other. For example, having a higher  $\epsilon$  means that the tuning width envelope should be quite flat rather than peaky (i.e., neurons are more similarly tuned in this case). This can be overcome by having a greater  $\tau$  and a smaller  $\sigma_{max}$ . Similarly, having a low  $\epsilon$  means that neurons that code for the context should be more sharply tuned, but having a high  $\sigma_{max}$  would reduce the sharpening. Therefore, I created a variant of the model to get around this problem:

$$\zeta_k = -f(V_k|C, \epsilon)$$

$$\sigma_k = \varsigma_{min} + (\sigma_{max} - \sigma_{min}) \frac{\varsigma_k - \varsigma_{min}}{\varsigma_{max} - \varsigma_{min}}$$

Equation 5-6

Here, without loss of generality, I define  $C$  as the mean value of all options. I move away from assuming that the context is driven solely by the distractor since human participants do not have full knowledge of which option in the choice set is the distractor in a trinary choice task that I will describe in later sections. The previous adaptive gain model was applied to a feature space that spans from 0 to 30 with a neuronal space that spans from -45 to 45 for our convenience. Now, I assumed that the value space spans from [0,1] and the neuronal space spans from [-3,3] to avoid any nonlinearities in the decoded subjective estimate that could be caused by the truncation of the tuning curves.

In the current formation of the model, I discarded the multiplicative scaling parameter  $\tau$ , and assumed that the inverse Gaussian function  $\varsigma_k$  is normalised so that the minimum and maximum tuning widths of a population of neurons are constrained by  $\sigma_{min}$  and  $\sigma_{max}$ . There are three parameters in this model:  $\sigma_{max}, \sigma_{min}, \epsilon$ . The  $\sigma_{min}$  and  $\sigma_{max}$  control the degree of gain (the height of the sharpened tuning curves), and  $\epsilon$  controls the breath of context-modulation along the value space.  $\sigma_{min}$  along with  $\epsilon$  are free parameters in our model, with a constraint that  $\sigma_{min}$  must be smaller than  $\sigma_{max}$  to keep our assumption that context results in sharpening of tuning curves (we could theoretically let both  $\sigma_{max}$  and  $\sigma_{min}$  to be free parameters and make no assumptions about whether context would induce an increase or a decrease in gain). We fixed  $\sigma_{max}$  to be a value of 2.5. This can be considered as the baseline tuning width, that is, when a neuron is not modulated by the context, it is widely tuned towards all possible values. Now, having a bigger difference in  $\sigma_{min}$  and  $\sigma_{max}$  means that there will be a greater degree of sharpening, i.e., the decision-maker is more context-dependent. Again, the population activity  $R$  can then be used for decoding the subjective estimate by summing the neural response from each neuron weighted by the feature value that neuron is coding for. To adjust for the scaling of the decoded subjective estimate introduced by the feature space

and the number of neurons, I can further multiply  $R$  with  $\frac{\rho}{M}$  where  $\rho$  represents the difference in the maximum and minimum value of the feature space, and  $M$  represents the number of neurons in the neuronal space. Figure 5.8 displays a schematic illustration of the model in a trinary choice task.

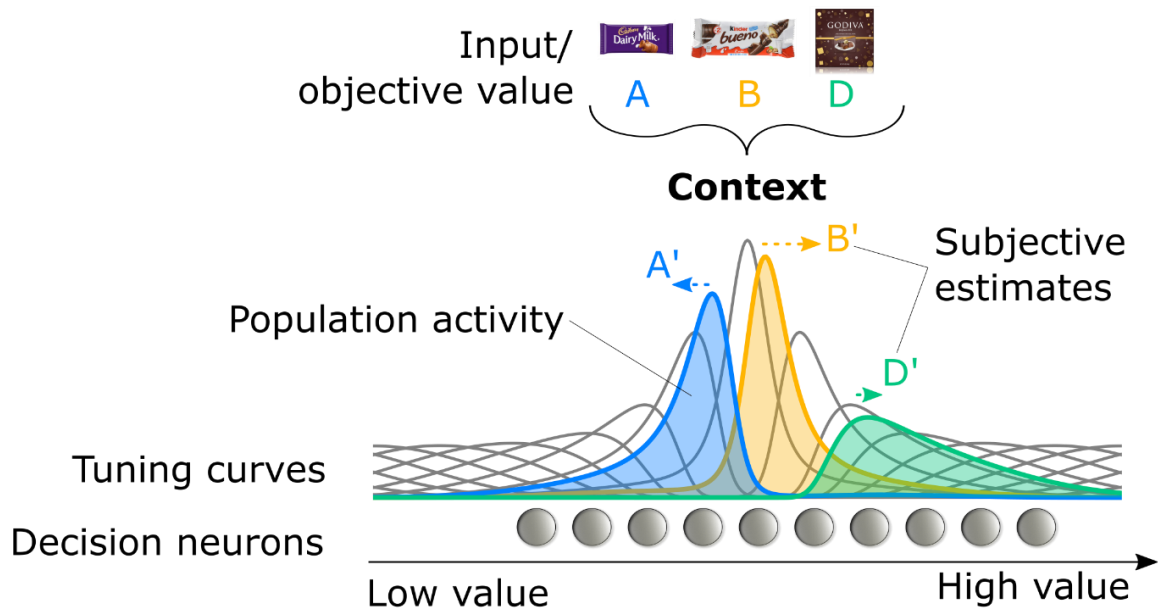


**Figure 5.8. Schematic illustration of the Adaptive gain model.**

The adaptive gain model assumed a set of linearly spaced neurons that encode decision information. Contextual signals can arise from the decoy option (see Chapter 3) or some other local statistics such as the mean value of all options. This illustration assumes three options: option A, B and D (an unavailable decoy). Each option activates the decision neurons and results in a profile of population activity. Options that share closer value with the context get heightened gain modulation, as shown by a greater magnitude of population activity for option B. The population activities are then used for decoding the subjective estimate. Coloured dashed arrows represents the shift in the decoded estimate from its original. In this one-dimension case, the decoy increases the chance the option B will get chosen.

### 5.3. A qualitative comparison of the predictions of efficiency-based models

To sum up, four efficiency-principled models: *i)* range adaptation, *ii)* range normalisation, *iii)* divisive normalisation, *iv)* adaptive gain have shown evidence for context-dependent value-based decisions. I further included the probit model as a baseline to examine the extent to which the capability of the model in explaining the context-dependent effect can be attributed to the inherent decision noise in the information processing system instead of due to the computation itself. However, how decoy effects can arise from efficiency-principled and biologically-plausible computations remains unexplored. For each model, the goal is to simulate how different models predict the effect of parametric decoy values in the attribute space.

#### 5.3.1. Parametric influence of a two-dimensional decoy.

This simulation aims to contrast model predictions about how irrelevant decoys influence decisions when options are characterised with two attributes. Since not all models were developed to explain multiattribute decisions, simulations on the multiattribute multialternative decisions require extra assumptions. I will first outline the rationale behind these assumptions before turning to the results of the simulation. Firstly, is the computation of the subjective estimates happening at the option-level or the attribute level? In this chapter, I assume that subjective estimates must be computed at the attribute level. This is because if the subjective estimates are computed based on integrated option values (i.e., the total objective value of an option), this boils down to a one-dimensional distractor effect which has already been discussed in other existing works (Gluth, Spektor, & Rieskamp, 2018). Recent work in humans shows that multiattribute decisions can best be explained by a class of models which utilise competition of attribute values across options, rather than competitions only at the option level (Hunt, Dolan, & Behrens, 2014), providing more motivation for assuming local, attribute-level computations. The second assumption is about whether the comparison process happens

at the attribute level or the option level (i.e., when does decision occur during the process). Noguchi and colleagues argued that models on multiattribute decisions should be based on attribute-wise comparisons rather than based on alternative-wise comparisons. They conducted an eye-tracking study to measure whether participants are more likely to switch their fixations within an option (implying integration of attribute values of an option) or more likely to switch their fixations across options within an attribute (implying a comparison of values across option within a single attribute; see Figure 5.9 Left panel). They compared fixation switching between alternatives and fixation switching within alternatives and found that for all three decoy effects, participants showed a higher likelihood in switching fixations between alternatives than within alternatives, implying an attribute-wise comparison process. However, this likely depends on the task and other factors such as whether there is a greater switch cost for comparing between attributes or comparing between options. For example, in a different study, participants ubiquitously sampled numbers from the same category before switching to a different category. In this study, subjects needed to fixate on a series of masked numbers in order to make a categorical judgement of whether the mean of numbers from category A is higher/lower than the mean of numbers from category B (Figure 5.9 right panel). Therefore, I make no a priori assumption of whether human participants make attribute-wise comparisons or not.

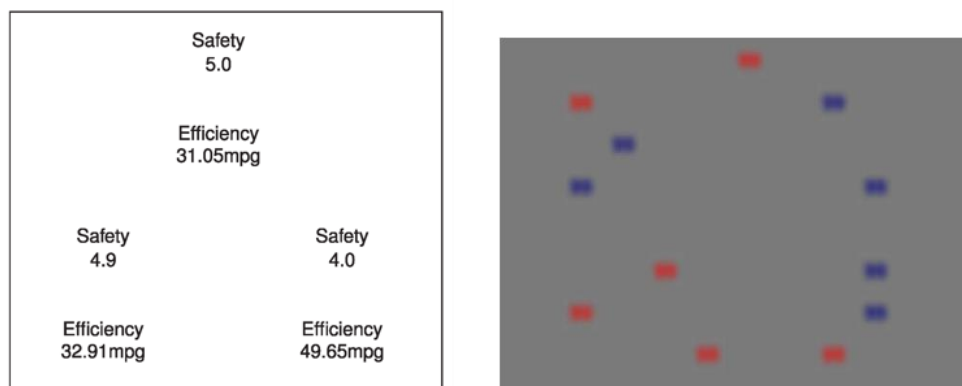


**Figure 5.9. Eye-tracking studies showing both attribute-wise and alternative-wise comparison process**

(Left panel) Figure reprinted from Noguchi and Stewart (2014). An example trial from the Noguchi eye-tracking study. Each trial consists of three options (top, bottom left and bottom right). Each option is characterised with two attributes: Safety and Efficiency in this example. (Right panel) Figure reprinted from (Vandormael, 2017). Masked numbers from different colour categories are displayed. Participants need to fixate on the numbers in order to reveal their identity.

Figure 5.10 illustrates an example of why it is essential to make a distinction on the assumption of the two types of comparisons during the decision process. Consider we have three options that are characterised by two attributes: economy (i.e., inverse price) and quality. Depending on whether the decision maker made a (noiseless) attribute-wise comparison or alternative wise comparison, the absolute choice probability for each option

|              |                                                                                   |                                                                                   |                                                                                   |
|--------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
|              |  |  |  |
| Economy      | 4 0%                                                                              | 6 100%                                                                            | 5 0%                                                                              |
| Quality      | 6 100%                                                                            | 4 0%                                                                              | 5 0%                                                                              |
| Option value | 10 33%<br>50%                                                                     | 10 33%<br>50%                                                                     | 10 33%<br>0%                                                                      |

|   |                 |
|---|-----------------|
| % | attr P(option)  |
| % | alter P(option) |

would be different. It implies that the relative choice ratio between options would be different, and it may or may not necessarily induce a violation of IIA. For example, the relative choice probability of orange to banana differs when one performs an attribute-wise comparison. For the same level of noise that corrupts the valuation on each option, the impact of noise on each option would be different as a function of which type of comparison one uses. For example, a positive noise term on the valuation of banana would result in the banana being chosen more often when one is doing an alternative-wise comparison but this would be less likely when one is making an attribute-wise comparison.

**Figure 5.10. Choice probability under attribute-wise and alternative wise comparisons**

An example of how the choice probability for each option (apple, orange and banana) can change as a function of whether it is an attribute-wise comparison (burgundy fonts) or an alternative-wise comparison (green fonts).

As a result, each model of interest will have two subsidiaries where the subjective estimates of the attribute values are either integrated before comparison or compared

before integration. For each model, I simulated the influence of a distracter  $Z$  on two equally preferred items  $X$  and  $Y$  that are characterised by two attributes, such as (inverse) price  $P$  and quality  $Q$ . I assume that the axes  $P$  and  $Q$  exhibit equal scaling and that  $X$  and  $Y$  fall on the line of isopreference which lies perpendicular to the identity line. We use  $X = [0.4, 0.6]$  and  $Y = [0.6, 0.4]$  as the two equally preferred target options, and  $Z_P$  (price attribute for  $Z$ ) and  $Z_Q$  (quality attribute for  $Z$ ) can vary between 0 to 1. We simulated the subjective estimate of the values for each option in a within-attribute manner. After computing subjective estimates for the attribute values for each option, these attribute values will either be integrated before being compared or the other way round.

For alternative-wise comparisons, I assumed that the agent first integrates the subjective estimates of all attributes for each option before comparison. For attribute-wise comparisons, I assume that there is an independent comparison process for each attribute. The choice probability for option  $X$  attribute  $P$  being greater than the value of the same attribute from other options is given by:

$$p_X^P = \int_{-\infty}^{\infty} \phi(t|X, \sigma_n) \Phi(t|Y, \sigma_n) \Phi(t|Z, \sigma_n) dt$$

Equation 5-7

Where  $\phi(\cdot)$  and  $\Phi(\cdot)$  refer to the probability density function and the cumulative probability distribution of the standard normal. Here, I assume that all options are conditionally independent. As a result, the probability that  $X$  is greater than both  $Y$  and  $Z$  is given by the product of cumulative probabilities that  $Y$  and  $Z$  is smaller than  $t$  and the probability that  $X$  equals  $t$  (i.e. the probability density function), integrating over all values of  $t$ . Lastly, for attribute-wise comparison, the choice probability is computed based on the subjective estimate of the value from each attribute. The overall choice probability of option  $X$  is simply the average of attribute choice probabilities:

$$P(X) = \frac{1}{2} [P_X^P + P_X^Q]$$

Equation 5-8

In this chapter, I assumed that attributes have equal-scaling and have equal weights on the decisions. Thus, the integration of attribute values into an option value is based on the sum of values from both attributes. Several existing models of multiattribute multialternative decisions assumed integration of attribute values are based on a weighted sum computation across attributes:

$$\hat{X} = \sum_{i=1}^J w_i X_i$$

Equation 5-9

Where  $J$  refers to the total number of attributes.  $w_i$  refers to the weight corresponds to the  $i$ th attribute, and the sum of weights across attributes would be 1. I took the simplest assumption that in the two dimensional attribute case, both attributes are weighted or integrated equally ( $w_i = 0.5$ ). To visualise the impact of an irrelevant decoy that can have any values in a two dimensional attribute space, I plotted a heat map that its colour-coded surface represents a relative preference toward either option  $X$  or option  $Y$ . To get the relative preference, I computed the relative choice share (RCS) difference. The RCS for option  $X$  is computed as follows:

$$\frac{P_X}{P_X + P_Y}$$

Equation 5-10

I zero-centred the relative choice share by taking the difference between the RCS for option  $X$  and RCS for option  $Y$ :

$$\frac{P_X - P_Y}{P_X + P_Y}$$

Equation 5-11

A positive value in RCS difference means option  $X$  is preferred to option  $Y$ , and vice versa. A rational agent that is not influenced by a third irrelevant option should have 0 RCS difference.

**Predictions on two-dimensional attribute decoy effect.** For each of the models of interest, we can focus on the following qualitative benchmarks: 1) Which decoy effects can the model predict, 2) Can the model predict all three decoy effects in one parameterisation, 3) What is the relationship between the distance of the decoy to the target and the strength of decoy effects. To look for these qualitative effects, I have conducted an exhaustive search of the entire parameter space (see Appendix Table 4.1 for the range of parameters used for searching and the list of parameters used for plotting Figure 5.11). Results of the exhaustive search are summarised in Table 5-1. Figure 5.11 shows the choice pattern from an example set of parameters that can recreate one or more decoy effects.

Previous work by Soltani and colleagues suggest that decoys that are further from the target options should induce a stronger decoy effect (Soltani et al., 2012). As Figure 5.11 highlights, all models show that decoys that are further from the target options would induce a stronger decoy effect (darker colour surface), so we should focus on which decoy effects each model can predict. Both the adaptive gain and the range normalisation model can explain all three decoy effects using one set of parameters under both comparison processes. As for range adaptation, it can explain all three decoy effects only when option values are integrated before comparisons. Furthermore, the simulations showed that the similarity effect could be explained by noise during the decision process (probit model). However, it predicts a context-dependent effect only under attribute-wise comparisons but not alternative-wise comparisons. This is because when I assume random noise on a trial-by-trial basis, this random noise can drive the decision to either direction, which means it gets cancelled out when summing over attributes (alternative-wise comparison). Therefore, the probit model under alternative-wise comparison predicts no patterned change in preference over two equally preferred options. It is interesting to see that divisive normalisation makes a very similar prediction to the probit noise model, implying that a violation of IIA under the current form of divisive normalisation is mainly caused by the appearance of noise during the comparison process rather than divisive normalisation itself.

**Table 5-1. A summary of the model prediction on parametric decoys in a two attribute dimensional space \***

|                               | Attribute-wise Comparisons |   |   |       |       |       |    | Alternative-wise Comparisons |   |   |       |       |       |    |
|-------------------------------|----------------------------|---|---|-------|-------|-------|----|------------------------------|---|---|-------|-------|-------|----|
|                               | A                          | C | S | A & C | A & S | C & S | SP | A                            | C | S | A & C | A & S | C & S | SP |
| <b>Probit</b>                 |                            |   | X |       |       |       |    |                              |   |   |       |       |       |    |
| <b>Adaptive Gain</b>          | X                          | X | X | X     | X     | X     | X  | X                            | X | X | X     | X     | X     | X  |
| <b>Divisive Normalisation</b> |                            |   | X |       |       |       |    |                              |   | X |       |       |       |    |
| <b>Range Adaptation</b>       | X                          |   | X |       | X     |       |    | X                            | X | X | X     | X     | X     | X  |
| <b>Range Normalisation</b>    | X                          | X | X | X     | X     | X     | X  | X                            | X | X | X     | X     | X     | X  |

\* A cross corresponds to the model's success in explaining: **A** = Attraction; **C** = Compromise; **S** = Similarity; **A & S** = Attraction and Similarity and so on; **SP** = whether the model can create all three decoy effects within a single parameterisation of the model

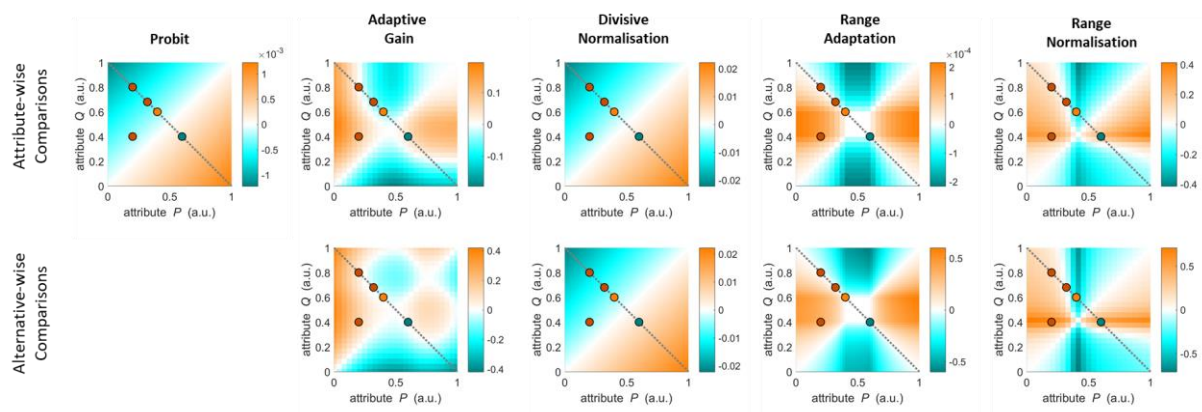


**Figure 5.11. Model predictions on the relative choice share difference**

The model predicted relative choice share difference between two equally preferred options ( $X$  and  $Y$ , orange and teal dots) as a function of a third decoy option  $Z$  ( $x$  and  $y$  axes for attribute inverse price  $P$  and quality  $Q$ ) for each model of interests (Columns). Brown dots correspond to the Attraction, Compromise and Similarity decoy used for testing the qualitative effect of each model (results in Table 4-1). The dashed line represents the indifference line. The orange dot indicates option  $X$  (where  $X_P = 0.4$ ,  $X_Q = 0.6$ ), and the blue dot indicates option  $Y$  (where  $Y_P = 0.6$ ,  $Y_Q = 0.4$ ). The coloured surface shows the model-predicted relative choice share difference

$\frac{p(X)-p(Y)}{p(X)+p(Y)}$ , that is the extent to which  $X$  is preferred over  $Y$  (orange regions) and conversely the teal regions referred to cases where  $Y$  is preferred. Top row corresponds to the model prediction assuming attribute-wise comparisons. Bottom row corresponds to the model prediction assuming alternative-wise comparisons. Parameters used for creating the corresponding surface plot is listed in Appendix Table 5.1, and the combination of parameters that can result in the same qualitative effect is displayed in Appendix Figure 5.1

## 5.4. A quantitative comparison of efficiency-based models

The three decoy effects have been tested extensively by a large body of experimental work (e.g., Table 1 from Trueblood et al.(2014) for a summary) and the effects have been used as a benchmark for evaluating the explanatory power of theoretical models on multiattribute multialternative decisions. However, more recent work has shown that the decoy effects can be reversed, e.g., (Cataldo & Cohen, 2018a, 2018b; Spektor et al., 2018), not exist under some cases (Frederick et al., 2014), or be weaker than we think (Liew, Howe, & Little, 2016). It is not clear under what circumstances the reversal of decoy effects could occur or when decoy effects would not appear. Cataldo and Cohen (2018) proposed that theories on context-dependent decisions should contain the power to explain when these context-dependent effects occur, not occur or are reversed. Moreover, no empirical work has systematically examined context effects with parametric decoy values that span the entirety of attribute space. In the following, I will describe a novel task that examines how parametric decoy values influence the choice patterns using a large-scale data set.

### 5.4.1. Parametric multiattribute multialternative experiment

#### Methods.

*Task.* Participants were recruited for a three-part study. In the first part, the participant completed a “property rental price guessing game”, where they were asked to guess the rental market price of some real estate properties from the US. On each trial, an image of a house was shown along with a slider for 1 minute (Figure 5.12A). Their task was

to guess the market rental price of the property, that is how much an average person is willing to pay per month to rent this property. I also randomised the initial value of the slider button on each trial to average out any rating bias caused by the initial value or the location of the slider button. There were a total of 250 unique house images. Each image is presented twice in randomised order (i.e., a total of 500 trials per participant).

When the participant finished the first part, I compared their ratings to the average market price ratings from a previous pilot. In that rating pilot, a separate cohort of subjects ( $N = 30$ ) rated 450 unique house images twice; these images were sampled from Zillow, an online real estate company. 250 images with the lowest between-subject variability were retained for the current task. The mean ratings from these 250 images were then used as the basis for the performance comparison. I used the Spearman's rank correlation to correlate the current participants' rating against the mean rating from the pilot. Participants who showed a correlation coefficient greater than 0.7 were invited for the second part – a “best-value property hunting game”. Thus, participants with a similar ordinal ranking of houses were qualified for the choice task independent of the range or the distribution of the absolute rating they provided.

In the second task, participants were told that they are a real estate agent who needs to recommend a client the best-valued house offered from us-- a real estate company. On each trial, three houses (i.e., options) are displayed for 1 minute. Underneath each image star ratings and a numerical cost are displayed, representing the two attribute dimensions: economy and quality (Figure 5.12D). Participants were told that the house images were from the rating task and the star ratings were also constructed based on their market rental price rating from this rating task; the cost refers to how much I am charging someone to rent this house. Their task was to press three keyboard buttons (left, down, right arrow button) to indicate their preference from the best-valued house to the worst-valued house. I explicitly instructed participants that the best-valued house refers to houses that have high market value but are charged with low rental price. At the end of

each block, participants were told how many trials' recommendations were correct, which turned into an actual bonus payment at the end of the experiment.

Not known to the participants, the options were carefully calibrated. I first filtered out 90 houses that subjects rated inconsistently across the two ratings in the first part, leaving 160 unique houses for constructing the choice set. I assumed that each house could be described by two attributes: quality and economy (Figure 5.12C). I assumed that the quality dimension is reflected by how much the participant thinks the house worth in market price, discretised into 10 bins (i.e. represented by the number of stars in the star ratings). I then construct the cost or economy dimension, which refers to how much I --the experimenter, are charging someone to rent a particular house. Importantly, I also discretised cost into 10 bins so that the bin edges matched perfectly between both dimensions. Options that fall into the diagonal grids in Figure 5.12C represent options that have similar quality and costs, and I call them iso-value options. Grids that are under the diagonal contain inferior options in the choice set and grids that are above the diagonal contain superior options in the choice set.

Each trial consists of three options, composed of two iso-valued option drawn from the red and blue grid (Figure 5.12C) and one decoy option that can have a cost that can be different from its rating. Figure 5.12E illustrates an example of an iso-value option, a superior and an inferior decoy. For example, a house that falls into the 4<sup>th</sup> market price decile bin might be assigned with a cost that is sampled from the 5<sup>th</sup> market price bin. In other words, the cost of renting the house is higher than the market price, which means it is a bad valued house/ an inferior option. Conversely, a good valued house (a superior option) should have a cost that is lower than the market price. This manipulation allowed us to have 10 decile bin costs x 10 decile bin market ratings = 100 decoy types defined by the bin they are in, covering the entire attribute space (Figure 5.12C). Most of the decoy types in the literature were defined under the iso-preference line, i.e. the majority (e.g. 8 unique trials per grid) of decoy types are sampled from under the diagonal grids. However, I chose to additionally include some trials where the decoy types are sampled from above

the diagonal grids, because different models make unique predictions across the entire two-dimensional space. Previous literature has failed to take into account the entirety of this space and has usually only focussed on a particular choice set where the decoy has the same attribute values across the entire experiment. This results in a total of 530 trials. When participants finished the second task, they can opt-in to complete more trials in the third part. Not explicitly mentioned to the subjects, the third part is just a repetition of the second part with randomised trial order and randomised option presentation.

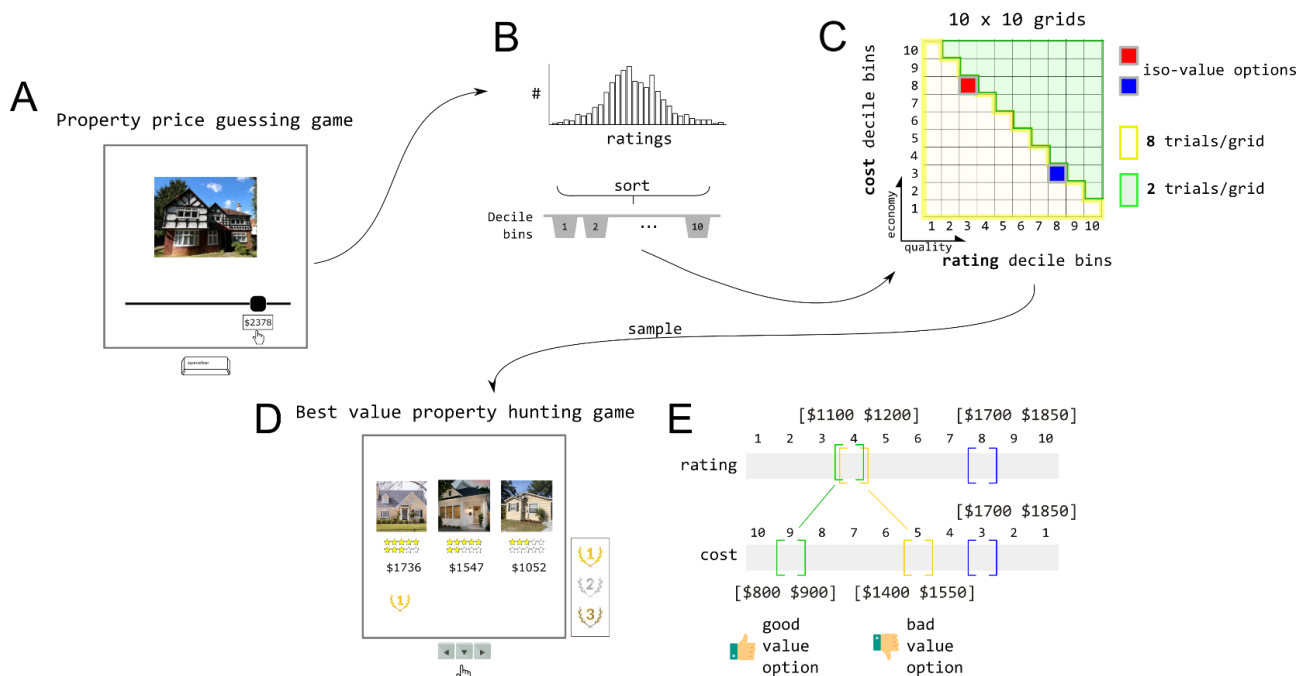


**Figure 5.12. Schematic illustration of the experiment**

(A) Participants first completed a game in which they need to rate the market rental price of the displayed image by dragging a slider with a mouse. Once they have confirmed their ratings, they can proceed to the next trial by pressing a spacebar. (B) Ratings of the first part are the basis for constructing options in the second and third part game. For each unique house image, I computed the difference of their ratings across two trials and then ranked all images according to this consistency measure. The 160 most consistently rated unique images are then sorted into decile bins based on their mean ratings. (C) Each house image option consists of two attributes: quality (x-axis; from low to high quality) and economy (y-axis; from high to low cost, i.e., low to high economy). The quality of a house is represented by the subjective rating from the first part. The cost of the house was assigned by us. Diagonal grids contain options that are iso-valued because

the cost assigned to these options should match roughly their ratings (i.e., the house is rented out for the market price). In particular, each trial contains at least an option-pair from the red and blue grid (each contains 30 unique house images) and a third option that can be sampled from the any of the grids. (D) On each trial, subjects need to make three keyboard responses (Left, Down, Right arrow key) corresponding to a choice for the left, middle and right option. They need to press those buttons in the order of the best to the worst valued option. At the end of each block, subjects were told how many trials of recommendation were preferred by the client. (E) A schematic illustration of different types of options. For the blue bracket, it shows an example of an iso-value option. The cost of renting this option is similar to the market rental price of this option (rating). For green and yellow brackets, they refer to the option where the cost can be higher (yellow; bad value option) or lower (green; good value option) than the market rental price. These brackets correspond to the grids in the colour shaded area in panel C.

*Participants.* A total of 358 US-located participants were recruited to participate in a three-part property game study on Amazon Mechanical Turk. After the first part (the rating task), participants who passed the performance threshold were invited to join the second and third part of the study (the choice task). In the end,  $N = 231$  participants continued to complete the second part. Out of which,  $N = 149$  participants volunteered to complete the third part. I collected data in two batches. In the first batch, we pay subjects \$4 for completing each part, and they can earn a maximum of \$10 performance-based bonus each from the second and third part of the study. To reduce the dropout rate, I increased the base payment to \$5 in the second batch and participants can earn up to \$12 and \$18 as the performance-based bonus in the second and third part of the study.

*Model fitting and comparison.* Instead of fitting each model on a trial-by-trial basis, I “downsample” the trials by taking the mean ratings and mean cost from each of the  $10 \times 10$  grids in Figure 5.12C as inputs. Each model outputs 100 choice probabilities, with each choice probability referring to the preference for option  $X$  over the preference for option  $Y$  (Equation 5-7 for computing the choice probability for option  $X$  and  $Y$  and Equation 5-11 for computing the RCS). Lastly, I compute the sum of squared errors between the empirical RCS difference and model predicted RCS difference:

$$\sum_{i=1}^{10} \sum_{j=1}^{10} \left[ \left( \frac{P_{ij}^X - P_{ij}^Y}{P_{ij}^X + P_{ij}^Y} \right) - \left( \frac{\gamma_{ij}^X - \gamma_{ij}^Y}{\gamma_{ij}^X + \gamma_{ij}^Y} \right) \right]^2$$

Equation 5-12

Where  $\gamma_{ij}^X$  represents the model predicted choice probability for option  $X$  in grid  $i$  and grid  $j$ . I used Bayesian Adaptive Direct Search (BADS; available at <https://github.com/lacerbi/bads>) optimization method to find the best-fitting parameters (Acerbi & Ma, 2017). Model comparisons were performed using the Variational Bayesian Analysis (VBA) Matlab toolbox: (available at <http://mbb-team.github.io/VBA-toolbox/>). Since all models consist of three free parameters, I assumed model complexity is the same for all models. I passed the negative sum of squared errors directly into the toolbox for random-effect Bayesian model selection. This method of model selection takes into account the individual variations on the generative model and computes the posterior probability of each model for each individual subject and lastly computes the expected frequency and exceedance probability for each model – two quantities that measure how likely a model is used more frequently relative to other models in the comparison set at the individual level and at the group level (Stephan et al., 2009).

*Cross model recovery.* For each model of interest, I am interested in whether I can recover the generative model of the choices. For each model ('true model'), I randomly sampled 50 empirical subjects' choice sets and randomly sampled parameter values uniformly from different ranges (see Table 4-2) as model inputs. I converted the model inputs according to the model parameters and the model algorithm (assuming alternative-wise comparisons) to generate model predicted choice probability. Taking these artificial choice probabilities, I fitted all 8 models using BADS (4 models: Adaptive Gain, Divisive Normalisation, Range Adaptation, Range Normalisation, each model uses 2 types of comparison). Lastly, I performed Bayesian model comparisons to find out which model best describes individual data at the group-level. The goal of this analysis is to discover: i) whether model parameters are recoverable, ii) whether models shared explained variance. For example, given that model B is the generative model, and model A shares features of

model B, I would expect model A's posterior probability to be greater than 0. This results in a significant deviation of model expected frequency from chance level.

**Table 5-2. Parameters used for generating artificial choices**

| Adaptive Gain                  | Divisive Normalisation    | Range Adaptation        | Range Normalisation     |
|--------------------------------|---------------------------|-------------------------|-------------------------|
| $\sigma_{min} = U(0.005, 2.5)$ | $K = U(0.001, 10)$        | $\beta = U(0.001, 10)$  | $f_t = U(-1, 0.99)$     |
| $\epsilon = U(0.1, 3)$         | $\sigma_H = U(0.001, 10)$ | $\sigma = U(0.001, 10)$ | $f_s = U(-1, 0.99)$     |
| $\sigma_n = U(0.01, 2)$        | $\sigma_n = U(0.01, 5)$   | $\sigma = U(0.01, 10)$  | $\sigma_n = U(0.01, 2)$ |

#### 5.4.2. Results

**Rating task.** In the experiment, I presented subjects ( $N = 358$ ) a set of US property images ( $m = 250$ ) twice and asked them to rate the market rental price (\$/per month) of the property on a continuous rating scale (from \$0 to \$2500). Participants were told that they would be invited for the choice task to get a chance to earn a generous bonus (max. \$30), only if their ratings followed the market price trend. I encouraged participants to make better estimates by also telling them that this would help them to perform better in the subsequent task (and earn a better bonus). I also point out to the subjects that their task was to guess how much an average person would pay for the house, instead of how much they like or are willing to pay for the house. To distinguish between conscientious participants and others, I impose a  $\rho = 0.7$  as the threshold for inviting subjects for the choice task. Figure 5.13 displays the histogram on the spearman's  $\rho$  from the cohort of subjects. The histogram shows that the performance is variable but, the majority of participants gets a  $\rho > 0.7$ , meaning that the majority of participants would rate a low-valued house with a low rating and a high-valued house with a high rating.

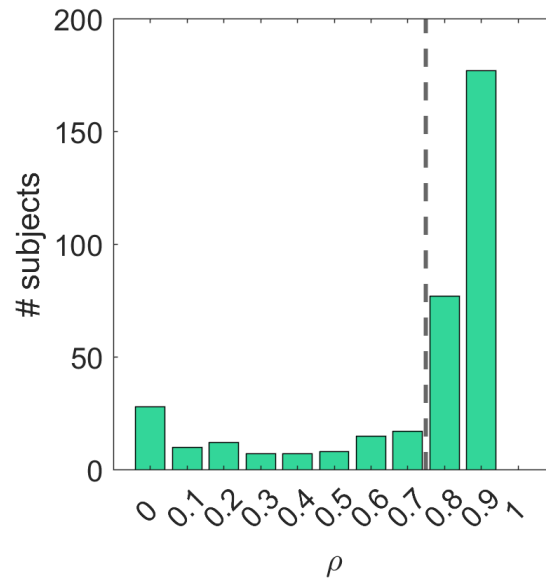


**Figure 5.13. A histogram on the Spearman's Rank-order correlation.**

Each subject's ratings were correlated with the ratings from a pilot rating experiment using Spearman's  $\rho$  -- a metric that measures whether the rank of the ratings for different houses matches with the rank of the ratings from the pilot experiment. The dashed line corresponds to the threshold I impose for inviting participants for the subsequent choice tasks.

After excluding participants who dropped out and who did not pass the performance threshold,  $N = 231$  have completed at least one choice task. Subsequent analyses were based on this pool of subjects. One question I asked is whether in the task, the rating displays scalar variability, i.e., larger magnitudes would result in a greater estimation noise (variability in the deviation of estimates; Petzschner et al., 2015). Previous behavioural and neural work in humans and non-human primates has suggested that judgements on both psychophysical magnitudes (distance, length, sizes) and symbolic magnitudes (symbolic numbers, arithmetic) follows the Weber-Fechner law and the Stevens' power law (Livingstone et al., 2014; Nieder & Miller, 2003; Petzschner et al., 2015). Steven's power law proposes a power law relationship between the represented magnitude by our sensory system and the physical magnitude. While for the Weber-Fechner law, it proposes that the represented magnitude is a function of the physical magnitude with a compressive scaling (i.e., Steven's power-law with a power exponent

smaller than one). In particular, the Weber-Fechner law also predicts scalar variability, an effect that has been found in numerosity experiments as well as in price estimation experiments (Dehaene & Marques, 2002). However, recent work by Polania and colleagues showed that the variability of the ratings does not follow the power law (compressive or anti-compressive functions) but an inverted-U shape: variability in the ratings is the greatest near the middle of the rating scale (Polania et al., 2019).

I computed the mean rating of each house image by averaging the rating across the two ratings and the variability of the ratings by taking the absolute difference of the two ratings and plotting them against each other in Figure 5.14. Next, I fitted zero, first and second-degree polynomials to the group averaged mean ratings and group averaged variabilities (Figure 5.14A) and also to the mean ratings and variability per individuals (Figure 5.14B):

$$= p_0 + p_1 R + p_2 R^2$$

Equation 5-13

Where  $V$  refers to the vector of variabilities:  $V = \{V_1, V_2, \dots V_m\}$ ,  $R$  refers to the vector of mean ratings  $R = \{R_1, R_2 \dots R_m\}$  for each unique house  $m$ , Lastly,  $p$  refers to the polynomial coefficients. The resulting least-square fitted line is displayed in Figure 5.14A in the green solid line. It shows that our rating data display scalar variability as predicted by the Weber-Fechner law. However, I went a step further and fitted each subject individually. I took the individual second-degree polynomials ( $p_2$ ) and split according to quartiles and plotted the least square lines from individuals in four panels (Figure 5.14B). The corresponding statistics on the polynomial coefficients are summarized in Table 5-3. As the figure and the table show, two groups of subjects displayed a concave parabola mean rating-variability pattern (negative  $p_2$  significant from 0; with  $Q_1$  group displaying a stronger nonlinear pattern than  $Q_2$  group), replicating the reported findings by Polania and colleagues.  $Q_3$  group displayed a positive linear relationship between mean ratings and

rating variability ( $p_1$  differs significantly from 0; i.e. scalar variability predicted by the Weber-Fechner Law). The last group of subjects display a convex parabola pattern, i.e., they have a lower variability in their ratings for middle-valued houses. This is the opposite of what Polania and colleagues found in their experiments. Taken together, this implies that there are individual differences on the noisy internal representation of different value magnitudes. Also, it is important to point out a few differences between our current rating task and the rating task in Polania and colleagues. Polania and colleagues tested fewer items ( $M = 64$ ). Their task is a preference rating task (i.e., participants were asked to rate how much they prefer an item rather than to guess the objective market price of the item), while our task is a mixture between inference and preference rating task. Lastly, to reduce any adaptations during the rating task, they made sure that participants were familiar with all items prior to ratings. All of these behavioural pattern could be explained by the Bayesian Observer model proposed by Polania and colleagues if assuming that the prior takes the shape as described by the mapping between mean rating and rating variability.

**Table 5-3. *t*-statistics on the polynomial coefficients**

| $Q_1$                            |                                  |                                  | $Q_2$                             |                                 |                                  |
|----------------------------------|----------------------------------|----------------------------------|-----------------------------------|---------------------------------|----------------------------------|
| $t_{50} = -19^{***}$<br>d = 3.83 | $t_{50} = 22^{***}$<br>d = 4.35  | $t_{50} = -8^{***}$<br>d = 1.59  | $t_{50} = -25^{***}$<br>d = 4.9   | $t_{50} = 20^{***}$<br>d = 3.99 | $t_{50} = -3.3^{**}$<br>d = 0.66 |
| $Q_3$                            |                                  |                                  | $Q_4$                             |                                 |                                  |
| $t_{50} = 0.6$<br>d = 0.12       | $t_{50} = 5.5^{***}$<br>d = 1.09 | $t_{50} = 6.3^{***}$<br>d = 1.25 | $t_{50} = 9.95^{***}$<br>d = 1.97 | $t_{50} = -9^{***}$<br>d = 1.78 | $t_{50} = -9.4^{**}$<br>d = 1.87 |

$Q_1$ - $Q_4$  corresponds to the four quantiles for subjects and the statistics for the four panels in figure 4.12B. The  $t$ -values and Cohen's  $d$  represents the statistics and effect size for the three polynomial coefficients:  $p_0$ ,  $p_1$  and  $p_2$ . Asterisk represents the significance threshold:  $p < 0.01^{**}$ ;  $p < 0.001^{***}$

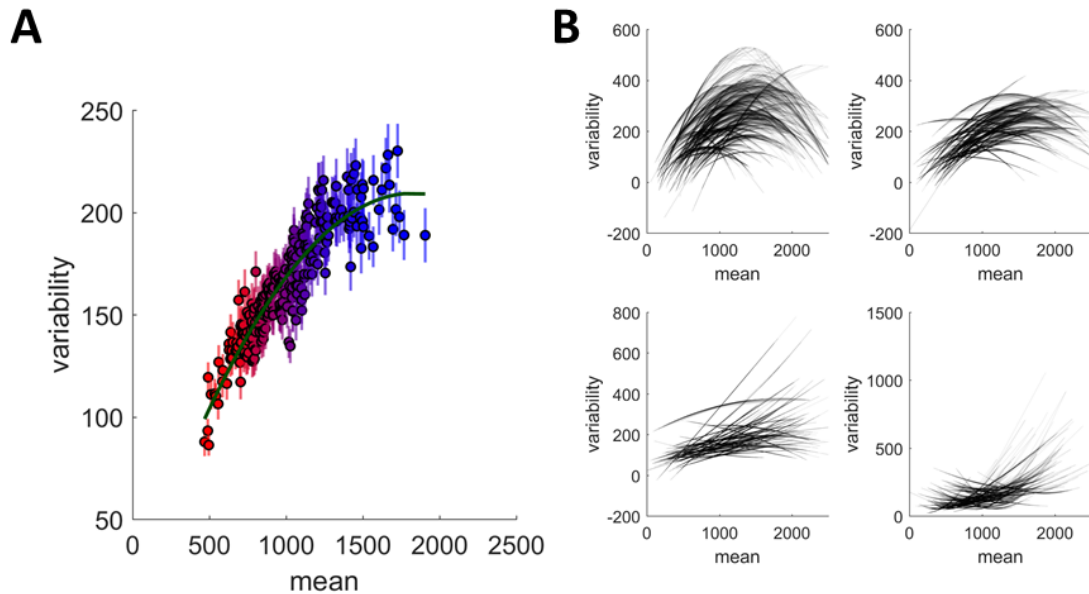


**Figure 5.14. Cross-plot for mean ratings and rating variability.**

(A) Cross-plot for mean ratings and rating variability. Participants rate 250 images (each dot) twice in the rating task. Mean ratings (average of the two ratings; x-axis) and variability of the ratings (absolute difference of the two ratings; y-axis) were plotted against each other. Horizontal lines (too small to be seen) and vertical lines correspond to the standard error of mean above and below the mean ratings and rating variability respectively. Colour codes represent images with low averaged ratings (red) to high averaged ratings (blue). The green line corresponds to the best fitting least-square line from a second-degree polynomial fit on the group averaged mean ratings and rating variability. (B). Individual variabilities in the relationship between mean ratings and rating variability. Each panel is plotting the best fitting least-square line from a second-degree polynomial fit on each individual's mean rating and rating variability. Each panel corresponds to a subset of subjects who have a second-degree polynomial coefficient that falls into the four quantile bins. The figure shows that subjects display different mean-variability patterns. In particular, half of the subjects (top panels) display the in rating variability pattern reported in Polania, Woodford & Ruff (2019), which was tested using a much smaller set ( $M = 64$ ) of goods and very different type of consumer goods (grocery items).

**Choice task.** In the choice task, participants were shown with three options, and they had to rank those options from the best-valued to the worst-valued house. I refer to the order of recommendation as the first, second and third choice. I collapsed the choice data from both the first choice task and the second choice task. Figure 5.15 shows the

choice bias in the choice task. I first looked at whether there is a bias on the spatial display of the options across individuals. I found that subjects were likely to pick the middle option first and the left or right option second (Figure 5.15A). There was a significant main effect of the display:  $F(2.31, 531.48) = 3300, p < 0.001, \eta_p^2 = 0.94$ , a significant main effect of choice:  $F(2.63, 604.53) = 4951, p < 0.001, \eta_p^2 = 0.96$  and a significant interaction effect:  $F(3.31, 760.88) = 15.89, p < 0.001, \eta_p^2 = 0.065$ . In our task, the option display was fully randomised. Nevertheless, I checked whether the bias on the option display interact with specific options. Figure 5.15B shows that participants generally have a strong bias toward picking option *Y* (the high rating, high cost option) as the best option (first choice). Together, I found a significant main effect of options:  $F(2.57, 591.18) = 1561.15, p < 0.001, \eta_p^2 = 0.87$ , a significant main effect of choice:  $F(2.48, 570.36) = 2567.64, p < 0.001, \eta_p^2 = 0.92$ , and a significant interaction effect:  $F(1.45, 334) = 116, p < 0.001, \eta_p^2 = 0.34$ . Unlike other decoy experiments in the literature, which often contained a calibration 2AFC task to find out the specific combination of attribute values of an option that would result in indifferent preferences, I did not perform any calibrations but just assumed that if one considers an option that worth \$100 ('quality'), charging him \$100 ('cost') to acquire this item is a fair trade. In other words, I objectively matched the numerical scale of the quality and cost attribute dimension and a priori assumed that the subjective preference scale would also match. However, the empirical data shows that many subjects have a stronger preference for one of the option, implying that the mapping between subjective preference of the two attribute dimension might not match perfectly. However, this should not influence our interpretation of the decoy effects, which I will show in the later sections that the baseline preference on one of the options can be corrected. Lastly, I check whether the participants understood the choice task. On each trial, the choice set contained a decoy and two iso-value options. Since I want to mainly explore the value space under the iso-value grids (Figure 5.15C), it results in  $45 \times 8 = 360$  trials in which choice sets contained an inferior decoy and  $45 \times 2 = 90$  trials in which choice sets contained a superior decoy and  $10 \times 8 = 80$  trials in which choice sets contained an iso-value decoy. A perfect performance agent should pick the inferior decoy last and the

superior decoy first (thus the performance should be capped at ~75% since 80 trials consists of iso-valued decoys). I then check the proportion of trials participants make the “correct” decisions by ranking an inferior decoy last and a superior decoy first. Figure 5.15C displays the accuracy of each individual subject. For later analyses on the decoy effects, we have excluded  $N = 24$  subjects who have an accuracy lower than 34%, i.e., lower than chance level.

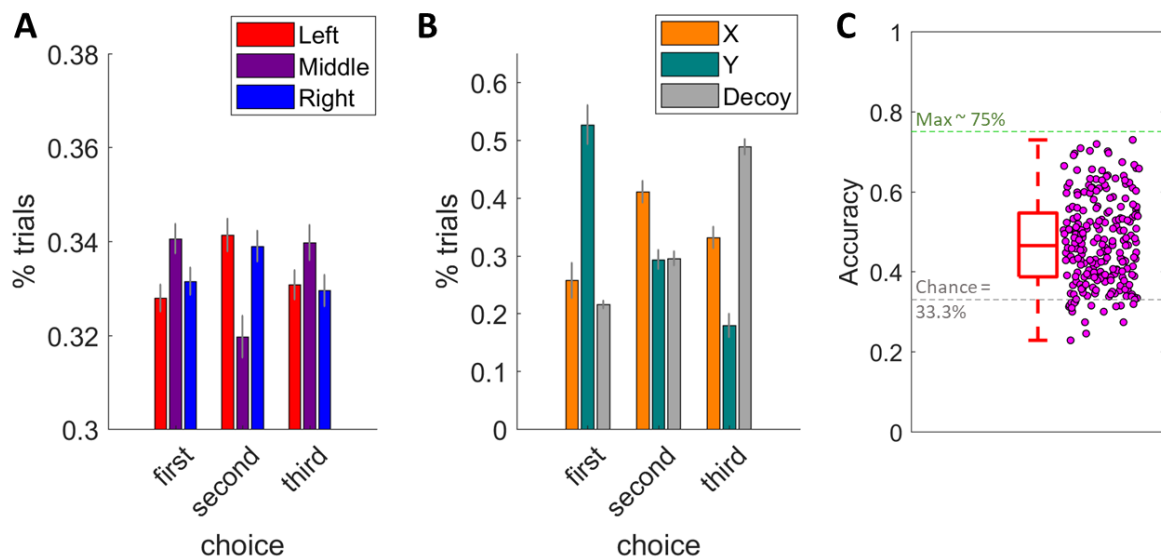
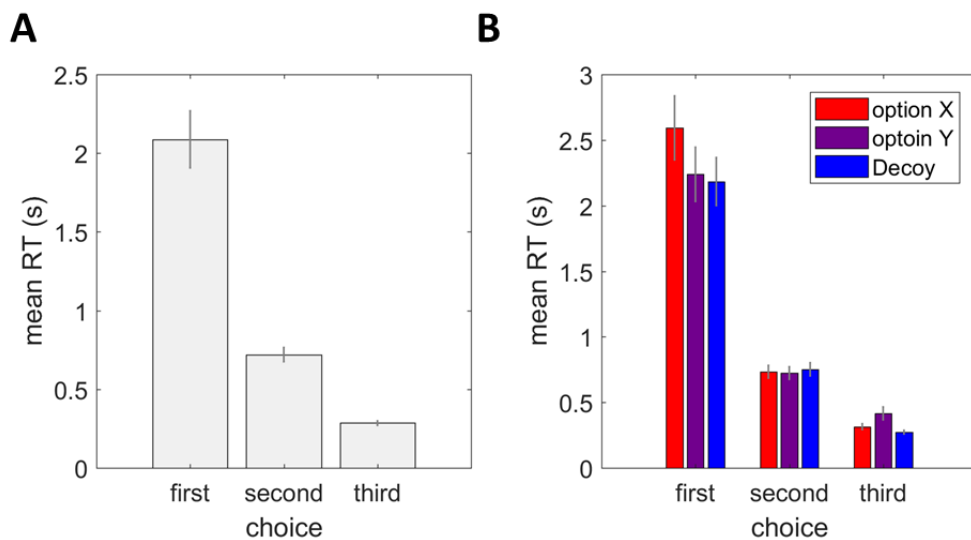


**Figure 5.15. Choice bias and performance**

(A) Choice bias on the left/middle/ right option (coloured bars). On each trial, subjects make three choices- from the most recommended to the least recommended option (x-axis). Option X, Y and Decoy's spatial displays are randomised on each trial. An agent who is unbiased by the spatial location of the option should select any of the options 1/3 of the time (y-axis) for any given choice. Grey lines correspond to the 95% confidence interval. (B) Choice bias for option X and Y. Participants generally are able to recommend an inferior decoy last, however insufficiently. 68% of trials contain an inferior decoy (where the decoy is definitively worse than option X and Y), yet the actual percentage of trials that the decoy is being selected least is much lower. Grey lines correspond to the 95% confidence interval. (C) Boxplot of the accuracy. Magenta dots correspond to the accuracy level of each individual subject. On each trial, the decoy can either be an inferior or a superior decoy. A correct choice is defined as when the subject picks the superior decoy first or pick the inferior decoy last on each trial. Grey dotted line corresponds the chance level performance which is 1/3. Green dotted line corresponds to the maximum performance one could achieve 75% (since 15% of trials consist of an iso-valued decoy).

I also recorded participants' reaction time (RT) since previous studies have shown that longer deliberation time results in stronger decoy effects. I first checked if the RT varies between choices (Figure 5.16). Participants take less time to make their second and third choice:  $F(1.07,245) = 310, p < 0.001, \eta_p^2 = 0.57$ . Next, I checked whether RT varies for different options, taking into account that on average participants would pick option Y first. I found that participants took a significantly longer time for option X compared to other options when they select option X as their first choice:  $t_{230} = 5, p < 0.001$ . This is consistent with the idea that people in general have a stronger preference for option Y. As a result, I have a significant main effect of choice:  $F(1.33,303) = 15.28, p < 0.001, \eta_p^2 = 0.063$ , a significant main effect of options:  $F(1.07,244) = 352, p < 0.001, \eta_p^2 = 0.61$ , and



a significant interaction effect:  $F(1.43,326) = 15.91, p < 0.001, \eta_p^2 = 0.065$ .

**Figure 5.16. Reaction time.**

(A) Mean RT for each choice. (B) mean RT as a function of each choice (x-axis) and each option (coloured bars).

When a participant finishes the first choice task, they are invited for the second choice task. In the end, we have  $N = 149$  subjects' data who have completed the second choice task. Not explicitly mentioned to them, I present them the same choice set from the first choice task but with randomised trial order and display. I then measure whether these

subjects make the same choice in the second choice task as in the first choice task for the same choice set. **Figure 5.17** shows the results of the choice consistency measures. I looked at the proportion of trials the participant made the same first, second and third choice (panel A). I further collapsed the data and plotted the proportion of trials the participants recommend option *X* before option *Y* from the same choice set (panel B). The results show that participants show considerable choice consistency in their preferences.

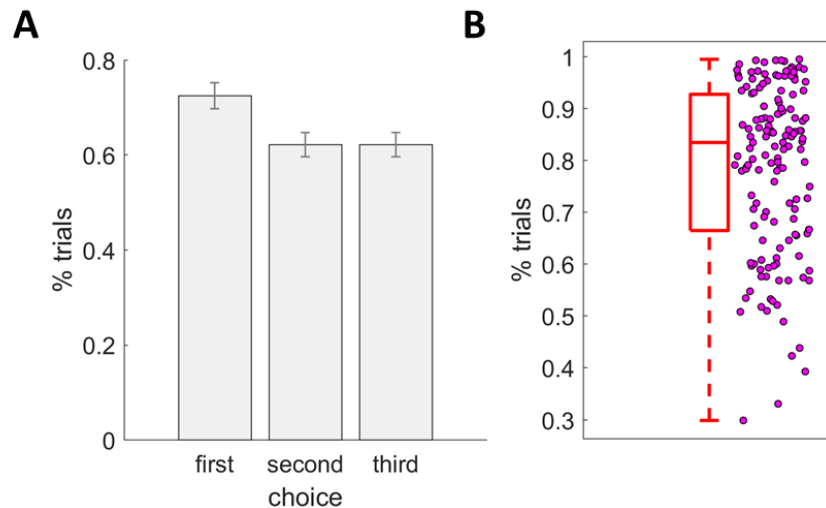
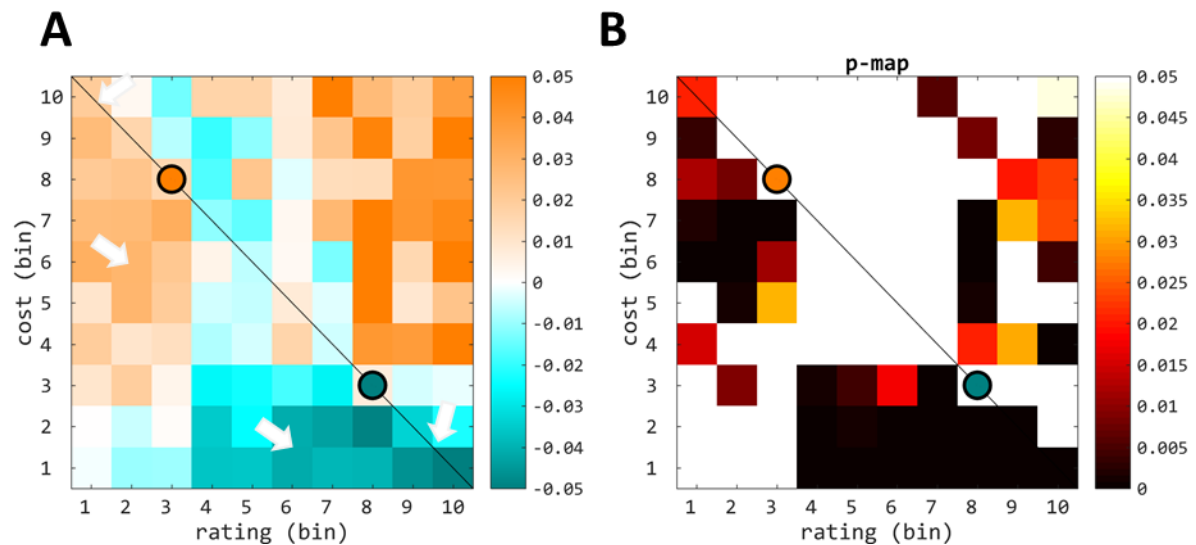


**Figure 5.17. Choice consistency.**

(A) Choice consistency for each of the choice (first, second and third recommendations) (B) Choice consistency for option *X* over option *Y* i.e., the proportional of trials that *X* is selected before *Y*.

**Decoy effects.** Existing studies of decoy effects have focused on three types of decoys: the attraction, similarity and compromise effects. Many of these studies have looked at only one or two effects (Trueblood et al., 2014). Here, I not only looked at all three decoy effects at once, I additionally mapped out the choice pattern of the entire attribute space, instead of focusing on a portion of the attribute space. Since trial numbers and choice set were carefully constructed so that they are evenly distributed along both attribute dimensions (Figure 5.12), I simply computed the choice probability within each grid. Participants made three choices on each trial but I only care about their relative preference on the option *X* over option *Y*, so I collapsed their three choices into a binary choice variable that describes whether the participant has recommended option *X* before option *Y*. I first measured the probability that a participant chooses *X* over *Y* across all

trials. This probability denotes one's general preference for  $X$  over  $Y$  averaged across all



decoy types. I then averaged the binary choice variable across trials within each of the grids and take the difference of this choice probability against the overall choice probability on option  $X$  as our final choice probability measure. This final choice probability measure tells us how one's preference changes with respect to their overall preference on the option  $X$ . For example, if one generally prefers option  $X$  over  $Y$  70% over all trials, but when this  $X$  and  $Y$  are paired with a specific decoy  $Z$ , the choice probability drops to 60%, it means that this person has switched his preference on option  $X$  to  $Y$ . Figure 5.18A displays the final choice probability for the 10 x 10 grids. Figure 5.18B shows the p-value of an independent t-test on each of the grids (see Appendix Figure 5.3 for the Cohen's  $d$ ). Both panels show strong evidence of the attraction and compromise effects (indicated by the white arrows). However, I found little evidence of the similarity effect. In the later sections, I am going to probe into whether the lack of similarity effect is caused by individual differences (e.g., if half of the subjects shows a normal similarity effect while the other half shows a reversed similarity effect). In the Appendix Figure 5.4 and 5.5, I display 25 subjects' individual choice maps which are the most similar to the group average and 25 subjects' maps which are the most dissimilar to the group average.

**Figure 5.18. Group average on relative choice share.**

(A) Normalised relative choice share averaged across the entire cohort. Colour surface represents a preference for option X (orange circle) or option Y (teal circle). (B) The  $p$ -value computed using a two-tailed t-test on each of the 10 x 10 grid. The coloured surface refers to choice sets that significantly induce a preference shift at a threshold of  $p < 0.05$

**Deliberation time and decoy effects.** Next, I ask whether having a longer deliberation time would result in a stronger decoy effect. I performed a tercile split on the RT for the first choice and compared the mean choice probability of the first tercile and the mean choice probability of the third tercile. If having a longer deliberation time would result in stronger decoy effects, we should see a significant positive difference in the choice probabilities. We found some evidence that longer RT resulted in an increase in the strength of context effects, especially for attraction and compromise decoys (Figure 5.19). This is consistent with the findings from previous studies on the effect of time pressure on decoy effects.

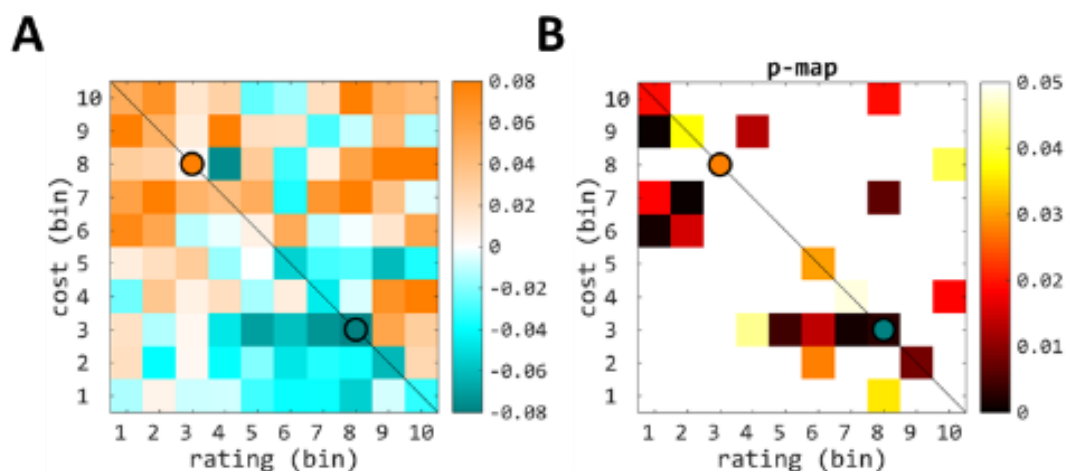


**Figure 5.19. Relative choice share difference between slow and fast reaction time trials.**

(A). Relative choice share difference for fast and slow RT trials defined by first and third tercile.  
(B) Significance testing using a t-test.

**Intercorrelation between decoy types and individual differences.** Previous work has shown that not all participants display all three decoy effects. Moreover, there is a stereotypical correlation between decoy effects: namely a positive correlation between the attraction and compromise effect, and a negative correlation between the similarity

effect and other decoy effects (Berkowitsch et al., 2014). One motivation for using parametric decoys in our task is precisely because what constitutes a compromise and a similarity decoy were arbitrarily defined in the literature. Compared to other decoy tasks, Berkowitsch and colleagues made a principled effort by defining decoys based on some distance between the decoy and the target options (e.g. a similarity decoy lies along the iso-preference line near the target option while maintaining a 10% distance between the target and the competitor option; Figure 5.2A). Even with this definition, it is not clear how one should define decoys that are, e.g. 30% distance from the target option, and so forth. In order to perform this correlation analysis, it requires us to first categorise parametric decoys into one of the three decoy types; while this contradicts our original motivation, I took some similar principles as categorisation guidelines, so that I can make comparisons between our data and findings from the literature.

I first assumed option  $X$  (low rating, low cost option) is the target option and option  $Y$  is the competitor option. The cost dimension is the attractive dimension for option  $X$ , therefore decoys were defined as attraction decoys if they have a higher cost than option  $X$  and a lower than option  $Y$ . To define compromise and similarity decoys, these options should not only fall close to the iso-value line, but they also should differ in terms of their distance to the target option. Therefore, I first compute the Euclidean distance between option  $X$  and option  $Y$  ( $D_{XY}$ ). I look for decoys where their summed rating and cost match roughly with the summed rating and cost for the target option (deviation  $< 0.1$ ). These decoys then get assigned as compromise decoys if the distance of this decoy with the target option (e.g.  $D_{XD}$ ) is greater than 20% of the distance between the target and the competitor option ( $|D_{XD} - D_{XY}| > 0.2D_{XY}$ ). Unlike Berkowitsch and colleagues, I cannot use 100% as the threshold due to how our choice sets were constructed. Decoys that fall out of the 20% distance were defined as similarity decoys. I repeated the same procedure to define decoy types for option  $Y$ , treating option  $Y$  as the target option and option  $X$  as the competitor option. Figure 5.20A shows the results of this decoy type categorisation across all subjects. I also inverted the sign of the choice probabilities for option  $Y$ , so that

a positive choice probability refers to a preference toward the target option and a negative choice probability refers to a preference toward the competitor option. Based on findings from the previous literature, I would expect a positive choice probability for attraction and compromise decoys and a negative choice probability for similarity decoys. Figure 5.20B-D shows the scatter plot of choice probabilities for each pair of decoy types. I computed the Pearson correlation and found a positive correlation between the attraction and compromise effect (  $r = 0.524, p < 0.001$  ), replicating what Berkowitsch and colleagues found in their data. However, I found a positive (instead of negative) correlation between similarity and other decoy effects (attraction effect:  $r = 0.35, p < 0.001$  ; compromise effect:  $r = 0.43, p < 0.001$  ). Importantly, I can see that some subjects show a normal similarity effect (negative choice probabilities for the target option) while others show a reversed similarity effect (positive choice probabilities for the target option). This explains the lack of similarity effect at the cohort level. One possible reason for this is that whether a decoy is classified as a similarity decoy instead of a compromise decoy depends on where one puts the cutoff for considering a decoy as “too similar” to the target options, which can be a subjective decision for each individual. Despite the failure to replicate the negative correlation effect from our data, there was a clear segregation between subjects who are susceptible to context effects (bluish dots) and subjects who are generally not influenced by context effects (reddish dots).

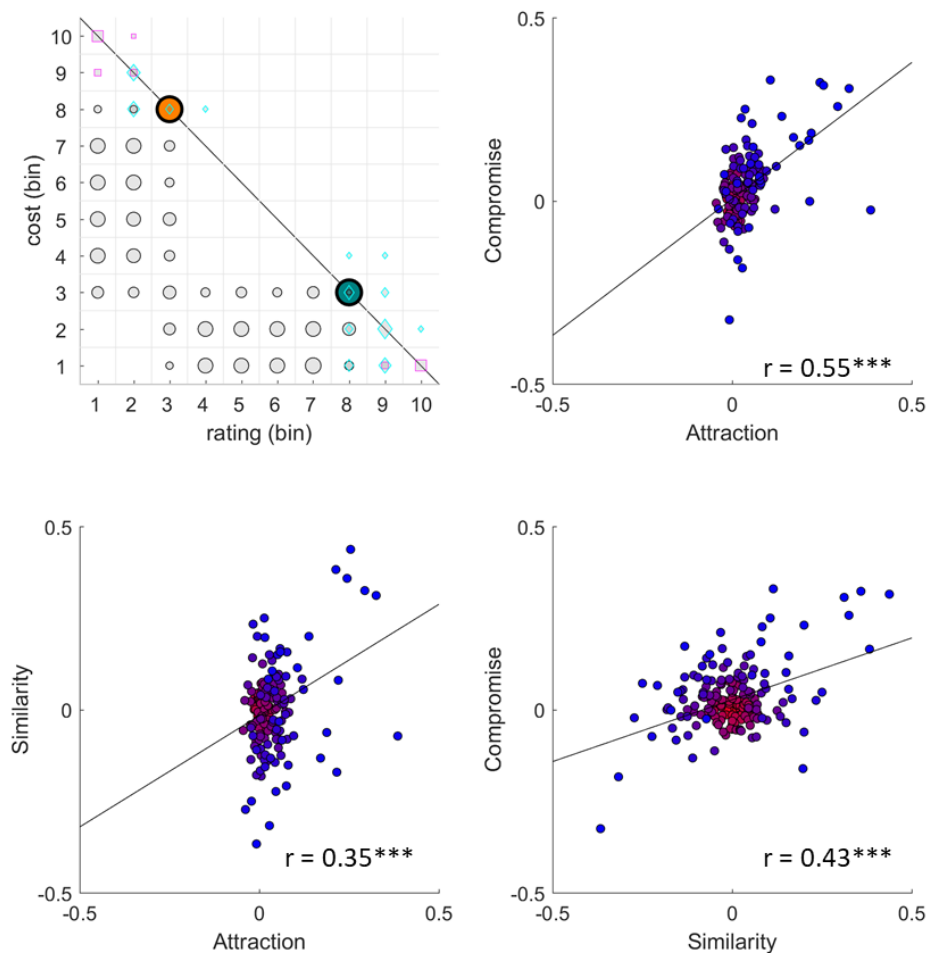


**Figure 5.20. Individual relative choice share pattern.**

(A) Choice set selected for plotting individual choice patterns. Different symbols correspond to different decoy types: magenta squares = Compromise decoys, cyan diamonds = Similarity decoys, grey circles = Attraction decoys. The size of the symbols refers to the density of trials that belong to a specific decoy type across subjects. (B – D) Scatter plots of the relative choice share for pairs of decoy type (x and y-axis). Positive relative choice shares refer to a preference for the target option. Negative relative choice shares refer to a preference for the competitor option. Coloured circles represent the overall strength of the context effect (sum of absolute relative choice shares from all decoy types), with red circles representing someone who is overall not impacted by decoys (relative choice shares close to 0) and blue circles representing someone who is very much influenced by decoys (in any directions). The bottom right text denotes the correlation coefficient of the relative choice shares for each pair of decoy types. \*\*\*:  $p < 0.001$

**Measuring model contributions on decoy effects.** Lastly, I performed formal quantitative model comparisons to ask i) which model can best explain the choice data at the group-level ii) if individuals adopt different strategies for the same task (different types of comparison and different efficient algorithms). I had four models of interest: Adaptive Gain, Divisive Normalisation, Range Adaptation and Range Normalisation, and we assumed that subjects could perform either alternative-wise comparisons or attribute-wise comparisons, meaning we have eight unique models for model fitting. I fitted each model independently by minimising the sum of squared errors between the model predicted choice probabilities and empirical choice probabilities, and took the negative sum of squared errors to conduct Bayesian model selection (see Methods). Bayesian model comparison across models showed that the majority of subjects were best explained by the adaptive gain model assuming attribute-wise comparisons since the posterior probability of this model is the highest compared to other models (Figure 5.21A and Figure 5.21C). Figure 5.21B shows that while the adaptive gain model is the model with the highest posterior probability, some models have a non-zero expected model frequency, meaning that the posterior probability of those models from some subjects are not close to 0. These models include the range normalisation (assuming attribute-wise comparisons) and the adaptive gain model (assuming alternative-wise comparisons), although the range normalisation model failed to reach chance level performance. Figure 5.22 shows the relative choice pattern for recreated model choices using the best fitting parameters from each of the models. Appendix Figure 5.4 shows the recovered intercorrelation between decoy types using the best fitting adaptive gain model.

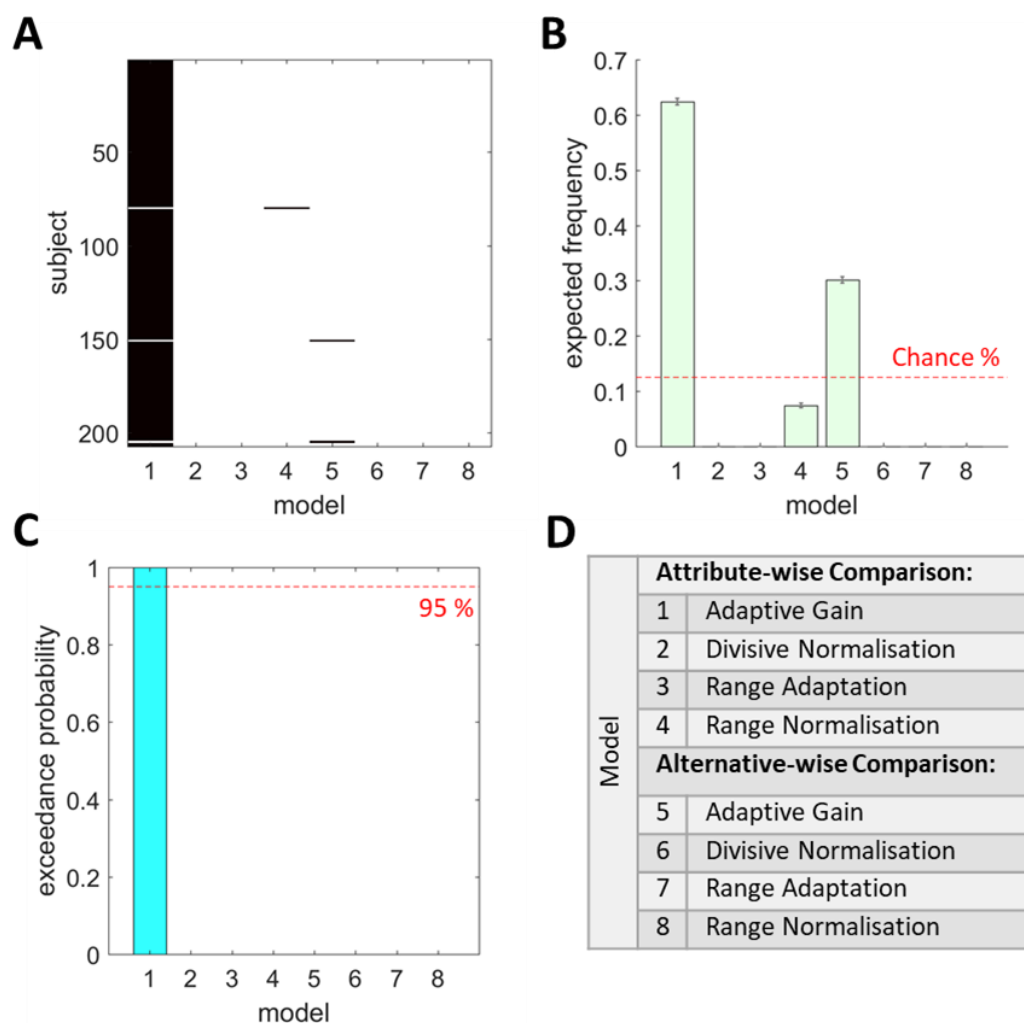


**Figure 5.21. Model comparisons across models using random effects Bayesian model selection.**

(A) Models with the highest posterior probability (shaded in black) for each individual subject (x-axis). (B). The expected frequencies for each model. Red line indicates the chance level (i.e.,  $1/8$  for a total of 8 models). Grey lines correspond to the 95% confidence interval. (C) The exceedance probability for each model, i.e., the probability of a given model is more frequent than the other models in comparisons. (D) The legends for model 1- 8.

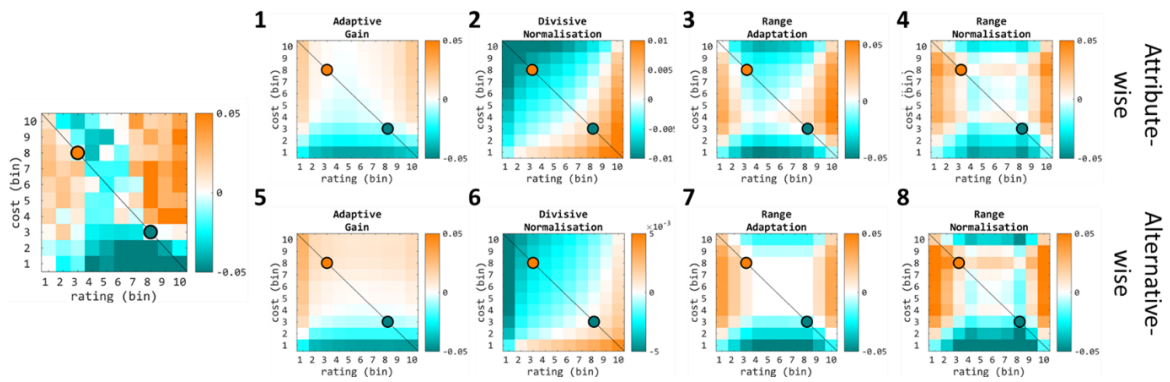


**Figure 5.22. Model recovery.**

Recovered relative choice pattern using the best fitting parameters. Left-panel: empirical choice probability (same as Figure 5.18). Panel 1 to 8 corresponds to the models illustrated in Figure 5.11, with the top row referring to models that assumed attribute-wise comparisons and the bottom row referring to models that assumed alternative-wise comparisons.

As the figure shows, there are some similarities across model predictions, I next performed a cross model recovery procedure (Methods) on all eight models. I assumed 50 hypothetical subjects and generated artificial choices for these subjects. Each subject is defined by a set of ground truth parameters. I took these artificial choices and fitted all models like the way I fitted human choice data, and lastly perform model comparisons. The idea was to identify whether each model allows itself to be recovered as the true generative model and to explore whether a particular model shares components with other models. If a model is recoverable, Bayesian model comparison should indicate that the same model is the best fitting model compared to other models. Figure 5.23 shows the results for Bayesian model comparison from the cross model recovery procedure. The results showed that models generally failed to recover the true model if I assume alternative-wise comparisons. It is worth noting that their counterparts on attribute-wise comparisons were often considered as the true model, implying there might not be an actual or observable difference between the two types of comparison processes. In particular, out of all models, the adaptive gain model and the range normalisation were best recovered. On the other hand, when range adaptation is the true generative model, the cross model recovery shows that a significant group of subjects show range

normalisation as their best model, implying that range normalisation and range adaptation share computational features. This is unsurprising since range normalisation encompasses range adaptation by definition. Interestingly, when divisive normalisation is the true generative model, the cross model recovery shows that some subjects' choices were best explained by the adaptive gain model. This is partially explained by the fact that two parameters in the model cannot recover themselves  $\kappa:r < 0.01, \sigma_n:r < 0.01$ .

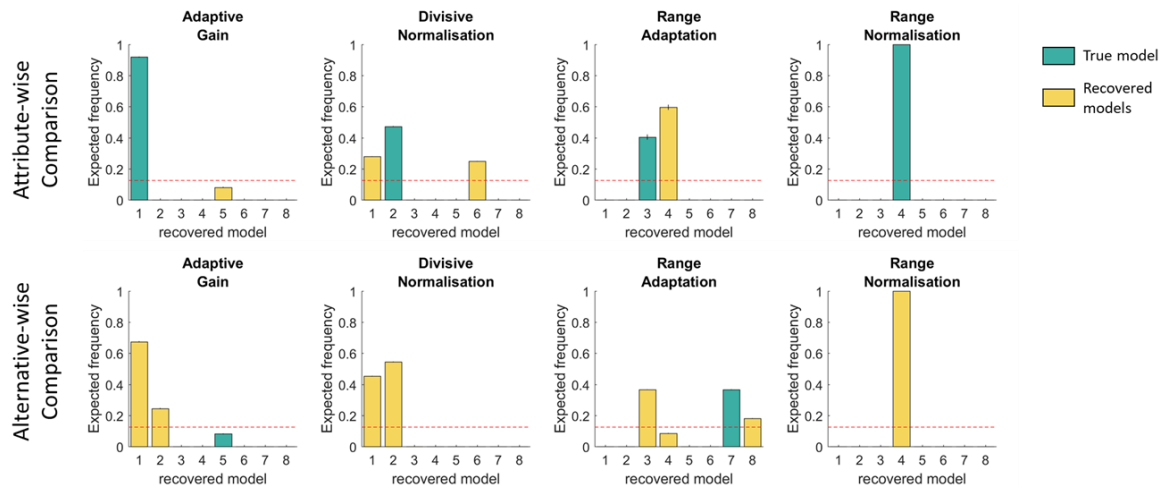


**Figure 5.23. Cross model recovery**

Model recovery between all model of interests. Each bar represents the expected model frequency. The dashed line represents the chance level (25%). Each panel's title represents the true generative model of model choices which subsequently were used for model fitting. Recovered model 1-8 denotes the 1/5) Adaptive Gain 2/6) Divisive Normalisation 3/7) Range Adaptation 4/8) Range Normalisation Model, with model 1-4 refers to models that assumed attribute-wise comparisons and model 5-8 refers to models that assumed alternative-wise comparisons.

### 5.4.3. Discussion.

Violations of IIA and decoy effects have been examined exhaustively in the decision-making literature. Different theoretical frameworks have been proposed in psychology and economics to explain the decoy effects. Building upon computational simulations by Li and colleagues (2018), I showed that a violation of IIA could be caused by decision noise alone,

and also by efficient computations that were developed to surmount the biophysical constraints of the brain. Efficient computations exploit statistical regularities in the environment to allow neural responses to adjust adaptively. Different computations assume different summary statistics of the environment as the driver for adaptation. For example, divisive normalisation cares about the sum of values of a choice set, range adaptation cares about the minimum and maximum values of a choice set, the current formulation of the adaptive gain assumes that the mean of option values drives the adaptation. Different assumptions result in different exhibitions of IIA. In our current data set, I found strong evidence of the attraction and compromise effect but no similarity effect at the group level. Choice correlation results showed that a portion of subjects displayed a reversed similarity and a reversed compromise effect. Together, the results showed that decoy effects are a lot richer and perhaps context-specific compared to what the majority of the literature has proposed. This aligns well with the recent reports that questioned the fidelity of decoy effects (Cataldo & Cohen, 2018b, 2018a; Frederick et al., 2014; Liew et al., 2016). I nevertheless showed that the human data could best be explained by the adaptive gain model which assumed that the agent compared between options at the attribute-level. Finally, I successfully recovered the group-level choice patterns and the choice correlation patterns with this model.

Finally, I would like to point out the limitations of the current chapter. First, I assumed that the observer has access to the true value of the items. Bayesian theorists would argue that the observer does not have access to what they have observed (e.g. orientation or the value of a good). Instead, the observer has to infer the value (Polania et al., 2019; Rigoli, Mathys, Friston, & Dolan, 2017; Sanborn & Chater, 2016). This implies that once the inference is performed, not only does the observer get the inferred value of a good, he also has access to his uncertainty on the inferred value. It is clear that the Bayesian approach got around the paradoxical question: Why would the observer compute a subjective estimate when he already has access to the true estimate. The Bayesian approach also implies one's decision confidence can simply be a readout of the posterior

probability. The drawback of this approach is that in many cases, the posterior distribution can only be approximated by sampling techniques. In the limit of infinite samples, the true posterior can be recovered, and it is in principle performing efficient coding (Bhui & Gershman, 2018), but infinite sampling is just impossible for any information processing system. However, it can be possible that the posterior probability could be approximated with sufficient samples, and the deviation of rationality could potentially be explained by a poor approximation of the posterior probability due to the lack of sampling. Recent work on the Decision by Sampling model, a model which suggests that the subjective value of an item is computed based on sampling from our memory, can explain all three decoy effects (Noguchi & Stewart, 2018). It remains unclear how the brain would perform decision by sampling. Thus, this chapter did not consider Decision by Sampling as one of the efficient computations. Secondly, the current chapter takes the most primitive form of divisive normalisation as our model of interests since this model successfully captured the one-dimensional distraction effect and has a strong biological basis. This model failed to recover parameters and failed to explain both attraction and compromise effects in our simulation setting. However, a similar model based on divisive normalisation was recently shown to capture all decoy effects if the divisive normalisation is conducted in a pairwise manner with the decoy option and at the attribute level (Daviet, 2018; Landry & Webb, 2017). The basic pairwise divisive normalisation algorithm states that the subjective value for two otherwise equally preferred options ( $X$  and  $Y$ ) in the presence of option  $Z$  can be computed by:

$$v(X, Y; Z) = \sum_{i=1}^J \left| \frac{X_i}{\sigma + X_i + Z_i} \right| - \sum_{i=1}^J \left| \frac{Y_i}{\sigma + Y_i + Z_i} \right|$$

Where the subscript  $i$  refers to attribute  $i$  and  $J$  is the total number of attributes.  $\sigma$  is the only free parameter in the model added to the normalisation term. This parameter also specifies the margin along the indifference line. If the choice set contains a decoy located in that margin, it would result in a similarity effect (or reversed compromised/attraction effect). This model is an attractive model for its simplicity. However, as I have mentioned in the main text, participants did not seem to have a full

knowledge of which option is the decoy option (option  $Z$ ). The second reason for us to use the primitive form of divisive normalisation is that this model successfully captured the one dimensional distractor effect. Like many of the existing models, I assumed that the value of an item is separable across dimensions and there is no independent mechanism for processing one or more attributes (i.e. the subjective value of a good is simply the sum or the mean of its attributes). As a result, I set out to explore whether the same divisive normalisation algorithm that was applied to the single attribute case is applicable to the multiattribute case. However, future work should consider this pairwise divisive normalisation model as a candidate model and investigate how the brain could perform pairwise normalisation when the distractor is not apparent and when the number of options and attributes increases. Second, I assume that an agent only engages in one efficient computation in the current task. As I have discussed previously, different models have assumed different summary statistics to be the relevant input for context modulation. Therefore, there is no reason why the brain would not engage in multiple efficient computations, especially given the recent work on OFC which suggests the possibility that different subpopulation of OFC neurons perform different relative value codes. Future work should consider factorial model comparisons to explore whether combining efficient computations results in novel behavioural predictions, and if they contribute to different forms of irrationality. Furthermore, the current chapter is only considering static models rather than dynamic models. This is a reasonable assumption given our choice to have a task where the participants have to make three choices on each trial, so that I do not have to discard trials in which the participants selected the decoy. This implies that it is hard to extract useful RT data since participants consistently take a much longer time for their first decision, making it hard to study the actual decision time for selecting a specific option. A recent study by Tajima, Drugowitsch, Patel, and Pouget (2019) built a recurrent network to approximate a biologically plausible optimal evidence accumulation process during multialternative decisions. In particular, momentary evidence accumulated by neurons at each time point were subject to divisive normalisation. This model successfully captured the neural response in the LIP (Louie, Grattan, & Glimcher, 2011) and the choice patterns

in a trinary choice task (Louie et al., 2013). Lastly, the model is able to replicate the similarity effect in the one-dimension case (i.e. a third option that is similar to option A would reduce the choice probability of option A), and that similarity effects became stronger as evidence accumulation goes by. It is not clear whether this model can capture preference shifts when the choice set contains other distractor types since I have shown that divisive normalisation can only recreate the similarity effect but failed to capture other decoy effects. Nevertheless, future work should consider how different efficient computations can embed in neural circuits and how bias evolve from the neural population dynamics.

## Chapter 6 - General Discussion

### 6.1. Summary

In **Chapter 1**, I described various optimality (in perceptual decision making) or rationality (in economic decision making) principles under different normative theories. To provide a knowledge basis regarding the source of suboptimality, I gave an overview of different types of efficiency that have been discussed in the cognitive science and neuroscience literature in **Chapter 2**. Finally, I used modern computational theories as examples to show how apparent sub-optimality can be understood as a consequence of a biological system that optimises for the aforementioned efficiencies.

In **Chapter 3**, I described an experimental study that investigated the cause of “robust averaging” – a consistently observed behaviour where humans downweight useful information during multi-element averaging tasks. Not only did we successfully replicate the robust averaging effect using orientations, but we also showed that this could be explained by a simple psychophysical model where feature values in the decision space are transduced according to the power-law before averaging. We found that the optimal transduction is determined by the stimulus distribution in the presence of noise that arises during or after averaging.

In **Chapter 4**, I investigated why humans take into account irrelevant information, a seemingly opposite behaviour compared to what was described in Chapter 3. Across visual science, cognitive psychology and behavioural economics, scientists have reported that irrelevant information (information which should not impact the goal) has a consistent influence on human behaviour. Namely, humans display the tilt illusion effect, conflict/congruency effect in flanker tasks and show a violation of independence of irrelevant alternatives (IIA) in value-guided tasks. I showed that an adaptive gain model in population codes could explain all three behavioural observations. In particular, the model assumes that the distracting information serves as a context for neural gain modulations. We collected

functional imaging data and showed that BOLD signals from various brain regions were best explained by the gain-modulated decision variable, providing reliable evidence of context-modulation in the brain.

Various forms of biologically constrained adaptive computations have been proposed to explain sensory adaptation effects and relative value codes both in behaviour and at the neural level. It remains unclear which form of computation is used in the cortical circuits. Extending the work described in Chapter 4, I showed that the adaptive gain model could predict a variety of qualitative benchmarks reported in the decoy effect literature. Given the richness of decoy effects, I asked in **Chapter 5** whether different efficiency-principled computations can explain decoy effects, and showed that all efficient computations could induce a violation of IIA. However, only some models predicted the full range of decoy effects. I also systematically tested decoy effects using parametric decoy values with real-life real estate property images and found consistent evidence of individual variabilities in some of the decoy effects. Together the described findings highlight the importance of studying the generalizability of decoy effects across different individuals and contexts.

## 6.2. General Discussion and Future Direction

In Chapter 3, we showed that performance in the averaging task was best explained by nonlinear transduction of the decision variable into a decision value signal. We made a number of assumptions in our task and in the modelling approach. Firstly, we assumed that the observer made a perfect mapping between feature values and decision value (from raw orientations in the sensory space to orientations in the decision space i.e. relative to the category boundary). Related to this point, we assume that early noise minimally impacts task performance. We considered both assumptions as reasonable since model fitting can faithfully recreate human performance without introducing early noise in the current task. As such, our conclusion that the nonlinear transduction can serve as a protection of post-decision noise as a function of stimulus distribution is limited to tasks where early noise is not a limiting factor. However, if early noise is a limiting factor in the task, it can potentially change

how the observers perform the task (e.g. Castañón et al., 2019) and change the relationship between noise and decision strategies. Future experiments should consider directly manipulating different sources of noise to see if and how decision strategies change.

Both Chapter 4 and 5 showed that the influence of irrelevant information across perceptual, cognitive and economic decision making tasks could be explained by the adaptive gain model in population codes. We introduced a few assumptions in the model. I am now going to discuss the applicability of these assumptions. First, we assumed that context results in the sharpening of tuning curves. The assumption of sharpening was motivated by the single unit recording literature which shows that attention can result in the sharpening of neuronal tuning for individual neurons (although some research shows the contrary i.e. decreases in firing rate with directed attention ;Reynolds & Heeger, 2009). The idea of sharpening is motivated further by the fact that when neurons are tuned towards the context (e.g. increase their response gain towards stimuli that repeatedly occur and decrease their sensitivity towards stimuli that do not occur often), the population firing becomes more informative regarding the identity of the frequently occurring stimuli and neurons become more sensitive towards changes in the frequently occurring stimuli. Here, we did not only assume that neural tunings are sharpened, but also that the area under each tuning curve is summed to unity, meaning that the firing over a population of neurons remained constant across contexts. This is an assumption that should be challenged and tested in future experiments, and it remains open whether response gain or sensitivity is the (more important) driver of decision biases.

Early work in repetition suppression has proposed various models of neural tunings to explain why presenting repeated stimuli results in a decrease in BOLD signal in fMRI tasks. One of the models proposed is a sharpening model by Desimone (1996) and Wiggs & Martin (1998). The sharpening model of repetition suppression assumes that neurons become more selective towards the repeated stimuli while the mean firing rate remains the same (i.e. a decrease in sensitivity towards non-repeat stimuli and no response gain). This implies overall firing over a population of neurons should decrease with increasing repetition of the same stimuli. It is possible through lateral inhibition dynamics for some neurons to increase and

others to decrease in firing (Grill-Spector et al., 2006). Other models of neuronal tuning include the fatigue model and the facilitation model. The fatigue model proposes that neurons show a proportional reduction in their firing to repeated presentation of the same stimuli. This is consistent with the notion of redundancy reduction in efficient coding, and can be explained by a simulation where there is local divisive normalisation computations in pools or pairs of neurons (Westrick, Heeger, & Landy, 2016). Recently, methodological advances have been made, enabling us to infer the possible neural mechanism of repetition suppression from fMRI responses (Alink, Abdulrahman, & Henson, 2018). Alink and colleagues constructed computational models that simulate neural responses under different adaptation regimes (e.g. sharpening, adaptation, etc.) and fit the BOLD voxel responses. Their modelling approach in BOLD responses can potentially allow us to make better conclusion on the research question I asked in Chapter 5. In Chapter 5, I asked whether different efficient computations predict different behavioural patterns in a trinary choice task. Different efficient computations should predict different voxel patterns since they have very different neural tuning assumptions. For example, Westrick and colleagues (2016) showed that neural responses in V1 during visual adaptation could be explained by a model where the neural responses for each pair of neurons remained constant by introducing divisive normalisation. The model predicts tuning functions similar to the predictions of the fatigue model. This implies the possibility to distinguish different efficient computations using non-invasive functional imaging techniques and the computational modelling proposed by Alink and colleagues. In our model, we also assumed a fixed number of neurons and they are linearly spaced. Others have proposed other tuning changes as a consequence of context-modulation. These alternative proposals include tuning shifts (i.e. the neuron changes the identity of preferred stimulus), density changes (a portion of neuronal space is cluttered with more neurons) and combinations of them (Ganguli & Simoncelli, 2014; Schwartz, Hsu, & Dayan, 2007; Wei & Stocker, 2015). In particular, Wei and Stocker (2015) showed that changing the density of neurons, tuning curve, shape and response gain could result in indistinguishable qualitative predictions regarding perceptual biases, suggesting that behavioural data might not be adequate for distinguishing between different efficient neural adaptations.

Moreover, our model assumed that the decoding of population activity is independent of the prior or stimulus distribution, i.e. non-Bayes optimal. This is the major qualitative difference between our model and the Wei and Stocker's Bayesian Observer Model. Both the adaptive gain model and the Bayesian Observer model made similar qualitative predictions regarding the relationship between sharpening/ prior shape and the consequent perceptual bias. Future work could compare the difference in quantitative fits of the two models over a variety of tasks (as Wei and Stocker (2018) have reported) and make an effort to formulate qualitatively distinct predictions regarding the role of stimulus distributions in neural decoding.

Lastly, it remains an open question of where and how adaptive gain control is implemented at the brain circuit level. For simplicity, we assumed that the neurons of our population coding model have Gaussian tuning curves. This assumption might not necessarily be true for certain brain regions' neurons. However, there is some evidence that brain regions such as the lateral parietal area might have Gaussian tuning curves (Ma et al., 2006; Platt & Glimcher, 1999). As such, we assumed that the neurons in our model are decision-neurons that are tuned to the decision value (or a common scale as "magnitude"). In Chapter 4, we showed that a network of brain regions could be best explained by the gain-modulated decision variable. However, we were not able to distinguish the role of these different brain regions. Future experiments that test how temporal dependencies or temporal contexts result in adaptive gain control might allow us to unpack which parts of the brain are involved in specific gain control or efficient computations. These experiments can consider using a task where decisions depend on a stream of sequential events. Alternatively, an experiment that requires the observer to learn the structure of the dynamically changing task statistics could be developed. Furthermore, recent work has shown that it is possible to relate brain signals from non-invasive functional imaging to the predictions of a circuit-level dynamical computational model (Hunt, Behrens, Hosokawa, Wallis, & Kennerley, 2015). Therefore, devising a circuit level computational model would allow us to map out the evolution of gain

control signals across populations of neurons to understand the source of the identified suboptimalities

.

## Reference

- Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In *Advances in neural information processing systems* (pp. 1836–1846).
- Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, 7(10), 1057.
- Alink, A., Abdulrahman, H., & Henson, R. N. (2018). Forward models demonstrate that repetition suppression is best modelled by local neural scaling. *Nature Communications*, 9(1), 3854.
- Allais, M. (2008). *Allais paradox*. Springer.
- Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: the "oblique effect" in man and animals. *Psychological Bulletin*, 78(4), 266.
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157–162.
- Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3), 183.
- Balaguer, J., Spiers, H., Hassabis, D., & Summerfield, C. (2016). Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron*, 90(4), 893–903.
- Barber, A. D., & Carter, C. S. (2004). Cognitive control involved in overcoming prepotent response tendencies and switching between tasks. *Cerebral Cortex*, 15(7), 899–912.
- Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In *Sensory Communication*. MIT Press.
- Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3), 241–253.
- Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, 321(5890), 851–854.
- Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci*, 10(9), 1214–21. doi:10.1038/nn1954
- Bengio, Y., LeCun, Y., & others. (2007). Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*, 34(5), 1–41.
- Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143(3), 1331.
- Bernoulli, D. (2011). Exposition of a new theory on the measurement of risk. In *The Kelly Capital Growth Investment Criterion: Theory and Practice* (pp. 11–24). World Scientific.
- Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, 125(6), 985.
- Blakemore, C., Carpenter, R. H., & Georgeson, M. A. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature*, 228(5266), 37–9.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4), 700.
- 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), 733–43. doi:10.1016/j.neuron.2009.05.014
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol Rev*, 108(3), 624–52.
- Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, 402(6758), 179.
- Bower, G. H., & Springston, F. (1970). Pauses as recoding points in letter series. *Journal of Experimental Psychology*, 83(3p1), 421.
- Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3), 389.
- Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, 113(27), 7459–7464.
- Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, 1–44.
- Bulthoff, H. H. (1996). Bayesian decision theory and psychophysics. *Perception as Bayesian Inference*, 123.
- Busmeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of Multi-Attribute, Multi-Alternative, Value-based decisions. *Trends in Cognitive Sciences*.

- Bushong, B., Rabin, M., & Schwartzstein, J. (2015). A model of relative thinking. *Unpublished Manuscript, Harvard University, Cambridge, MA.*
- Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7), 2183–2203.
- Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1), 51.
- Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280(5364), 747–749.
- Castañón, S. H., Moran, R., Ding, J., Egner, T., Bang, D., & Summerfield, C. (2019). Human noise blindness drives suboptimal cognitive inference. *Nature Communications*, 10(1), 1719.
- Cataldo, A. M., & Cohen, A. L. (2018a). Reversing the similarity effect: The effect of presentation format. *Cognition*, 175, 141–156.
- Cataldo, A. M., & Cohen, A. L. (2018b). The comparison process as an account of variation in the attraction, compromise, and similarity effects. *Psychonomic Bulletin & Review*, 1–9.
- Charnov, E. L., & others. (1976). Optimal foraging, the marginal value theorem.
- Chau, B. K., Kolling, N., Hunt, L. T., Walton, M. E., & Rushworth, M. F. (2014). A neural mechanism underlying failure of optimal choice with multiple alternatives. *Nat Neurosci*, 17(3), 463–70. doi:10.1038/nn.3649
- Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castanon, S., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, 81(6), 1429–41. doi:10.1016/j.neuron.2014.01.020
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43(4), 393–404.
- Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, 45(7), 891–900.
- Clifford, C. W. (2002). Perceptual adaptation: motion parallels orientation. *Trends in Cognitive Sciences*, 6(3), 136–143.
- Cohen, D., Volovich, M., Zeevi, Y., Louie, K., Levy, D. J., & Rechavi, O. (2018). Asymmetric Overlap in Neuronal Sensation Constrains Rational Choice in *C. elegans*. *BioRxiv*, 257535.
- Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9), 910–923.
- Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and Brain Sciences*, 24(1), 154–176.
- Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, 96, 1–25.
- Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology*, 10(1), e1003441.
- Daviet, R. (2018). Methods for statistical analysis and prediction of discrete choices. University of Toronto.
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6), 1204–1215.
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704.
- Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*, vol. 806. Cambridge, MA: MIT Press.
- Dayan, P., & Solomon, J. A. (2010). Selective Bayes: attentional load and crowding. *Vision Res*, 50(22), 2248–60. doi:10.1016/j.visres.2010.04.014
- De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, 70(5), 789–794.
- de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proc Natl Acad Sci U S A*, 108(32), 13341–6. doi:10.1073/pnas.1104517108
- Dehaene, S. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, 7(4), 145–147.
- Dehaene, S., & Marques, J. F. (2002). Cognitive neuroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *The Quarterly Journal of Experimental Psychology Section A*, 55(3), 705–731.
- Deneve, S., Latham, P. E., & Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8), 740.
- Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences*, 93(24), 13494–13499.

- Dinstein, I., Heeger, D. J., & Behrmann, M. (2015). Neural variability: friend or foe? *Trends in Cognitive Sciences*, 19(6), 322–328.
- Dragoi, V., Sharma, J., & Sur, M. (2000). Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, 28(1), 287–298.
- Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6), 1398–1411.
- Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nat Neurosci*, 8(12), 1784–90. doi:10.1038/nn1594
- Eimer, M. (1999). Facilitatory and inhibitory effects of masked prime stimuli on motor activation and behavioural performance. *Acta Psychol (Amst)*, 101(2–3), 293–313.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a non-search task. *Perception and Psychophysics*, 16, 143–149.
- Frederick, S., Lee, L., & Baskin, E. (2014). The limits of attraction. *Journal of Marketing Research*, 51(4), 487–507.
- Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9), 1325–1352.
- Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, 26(10), 2103–2134.
- Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, 51(7), 771–781.
- Geisler, W. S., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, 5(6), 508.
- Gershman, S. J. (2018). How to never be wrong. *Psychological Bulletin and Review*. doi:https://doi.org/10.3758/s13423-018-1488-8
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273–278.
- Geurts, L. S., Chetverikov, A., van Bergen, R. S., Zhou, Y. J., Bertana, A., & Jehee, J. F. (2018). Optimality is critical when it comes to testing computation-level hypotheses. *Behavioral and Brain Sciences*, 41.
- Gibson, J. J. (1937). Adaptation with negative aftereffect. *Psychological Review*, 44, 222–244.
- Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.
- Gilbert, C. D., & Wiesel, T. N. (1990). The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat. *Vision Research*, 30(11), 1689–1701.
- Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7), 926.
- Gluth, S., Spektor, M. S., & Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *ELife*, 7, e39659.
- Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, 5(1), 10–16.
- Gold, J. I., & Stocker, A. A. (2017). Visual decision-making in an uncertain and dynamic world. *Annual Review of Vision Science*, 3, 227–250.
- Golz, J., & MacLeod, D. I. (2002). Influence of scene statistics on colour constancy. *Nature*, 415(6872), 637.
- Gorea, A., Belkoura, S., & Solomon, J. A. (2014). Summary statistics for size over space and time. *Journal of Vision*, 14(9), 22–22.
- Gratton, Gabriele., Coles, Michael. G. H., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *J Exp Psychol Gen*, 121(4), 480–506.
- Gratton, Gabriele., Coles, Michael. G. H., Sirevaag, Erik. J., Eriksen, Charles. W., & Donchin, Emanuel. (1988). Pre-and poststimulus activation of response channels: a psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3), 331.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.
- Gregory, R. L. (1968). Perceptual illusions and brain models. *Proc. R. Soc. Lond. B*, 171(1024), 279–296.
- Gregory, R. L. (1997). Visual illusions classified. *Trends in Cognitive Sciences*, 1(5), 190–194.
- Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012).
- Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, 10(1), 14–23.
- Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, 49(5), 757–63. doi:10.1016/j.neuron.2006.01.032
- Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, 72(7), 1825–1838.

- Hasenstaub, A., Otte, S., Callaway, E., & Sejnowski, T. J. (2010). Metabolic cost as a unifying principle governing neuronal biophysics. *Proceedings of the National Academy of Sciences*, 107(27), 12329–12334.
- Hazeltine, E., Poldrack, R., & Gabrieli, J. D. (2000). Neural activation during response competition. *J Cogn Neurosci*, 12 Suppl 2, 118–29. doi:10.1162/089892900563984
- Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Vis Neurosci*, 9(2), 181–97.
- Heilbronner, S. R., & Hayden, B. Y. (2016). Dorsal Anterior Cingulate Cortex: A Bottom-Up View. *Annu Rev Neurosci*, 39, 149–70. doi:10.1146/annurev-neuro-070815-013952
- Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214), 1158–1161.
- Ho, T. C., Brown, S., & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *J Neurosci*, 29(27), 8675–87. doi:10.1523/JNEUROSCI.5984-08.2009
- Hotaling, J. M., & Rieskamp, J. (2018). A quantitative test of computational models of multialternative context effects. *Decision*.
- Howe, C. Q., & Purves, D. (2002). Range image statistics can explain the anomalous perception of length. *Proceedings of the National Academy of Sciences*, 99(20), 13184–13188.
- Howes, A., Warren, P. A., Farmer, G., El-Deredy, W., & Lewis, R. L. (2016). Why contextual preference reversals maximize expected value. *Psychological Review*, 123(4), 368.
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154.
- Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9(1), 90–98.
- Hunt, L. T., Behrens, T. E., Hosokawa, T., Wallis, J. D., & Kennerley, S. W. (2015). Capturing the temporal evolution of choice across prefrontal cortex. *Elife*, 4, e11945.
- Hunt, L. T., Dolan, R. J., & Behrens, T. E. (2014). Hierarchical competitions subserving multi-attribute choice. *Nature Neuroscience*, 17(11), 1613.
- Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5), 690.
- Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4), 169–188.
- Juechems, K., Balaguer, J., Ruz, M., & Summerfield, C. (2017). Ventromedial Prefrontal Cortex Encodes a Latent Estimate of Cumulative Reward. *Neuron*, 93(3), 705–714.
- Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow* (Vol. 1). Farrar, Straus and Giroux New York.
- Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99–127). World Scientific.
- Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, 303(5660), 1023–1026.
- Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, 13(2), 150–158.
- Knill, D. C., Kersten, D., & Yuille, A. (1996). Introduction: A Bayesian formulation of visual perception. *Perception as Bayesian Inference*, 1–21.
- Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6), 229–235.
- Kohn, A., & Movshon, J. A. (2004). Adaptation changes the direction tuning of macaque MT neurons. *Nature Neuroscience*, 7(7), 764.
- Kolling, N., Behrens, T. E. J., Wittmann, M. K., & Rushworth, M. F. S. (2016). Multiple signals in anterior cingulate cortex. *Current Opinion in Neurobiology*, 37, 36–43.
- Kolling, N., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2012). Neural mechanisms of foraging. *Science*, 336(6077), 95–98.
- Kording, K. (2007). Decision theory: what 'should' the nervous system do? *Science*, 318(5850), 606–10. doi:10.1126/science.1142998
- Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971), 244.
- Kőszegi, B., & Szeidl, A. (2012). A model of focusing in economic choice. *The Quarterly Journal of Economics*, 128(1), 53–104.
- Landry, P., & Webb, R. (2017). Pairwise Normalization: A Neuroeconomic Theory of Multi-Attribute Choice.
- Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, 11(4), 475–480.

- Lea, A. M., & Ryan, M. J. (2015). Irrationality in mate choice revealed by túngara frogs. *Science*, 349(6251), 964–966.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436.
- Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2), 279–311.
- Li, V., Castañón, S. H., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology*, 13(8), e1005723.
- Liew, S. X., Howe, P. D., & Little, D. R. (2016). The appropriacy of averaging in the study of context effects. *Psychonomic Bulletin & Review*, 23(5), 1639–1646.
- Livingstone, M. S., Pettine, W. W., Srihasam, K., Moore, B., Morocz, I. A., & Lee, D. (2014). Symbol addition by monkeys provides evidence for normalized quantity coding. *Proceedings of the National Academy of Sciences*, 111(18), 6822–6827.
- Louie, K., Glimcher, P. W., & Webb, R. (2015). Adaptive neural coding: from biological to behavioral decision-making. *Current Opinion in Behavioral Sciences*, 5, 91–99.
- Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward value-based gain control: divisive normalization in parietal cortex. *Journal of Neuroscience*, 31(29), 10627–10639.
- Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1217854110
- Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279.
- Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci*, 9(11), 1432–8. doi:10.1038/nn1790
- Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17(3), 347.
- Machens, C. K., Gollisch, T., Kolesnikova, O., & Herz, A. V. (2005). Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, 47(3), 447–456.
- Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. *Probabilistic Models of the Brain*, 13–36.
- Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research*, 38(18), 2817–2832.
- Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24(12), 2351–2360.
- Michael, E., de Gardelle, V., Nevado-Holgado, A., & Summerfield, C. (2013). Unreliable evidence: 2 sources of uncertainty during perceptual choice. *Cerebral Cortex*, 25(4), 937–947.
- Michael, Elizabeth., de Gardelle, V., & Summerfield, C. (2014). Priming by the variability of visual information. *Proc Natl Acad Sci U S A*, 111(21), 7873–8. doi:10.1073/pnas.1308674111
- Michael, Elizabeth., Moran, R., Balaguer, J., Ruz, M., Usher, M., & Summerfield, C. (In prep.). Adaptation to feature variability in the human visual cortex.
- Miller, G. (1956a). Human memory and the storage of information. *IRE Transactions on Information Theory*, 2(3), 129–137.
- Miller, G. A. (1956b). Information and memory. *Scientific American*, 195(2), 42–47.
- Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, 125(4), 486.
- Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, 45(2), 205–231.
- Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37(1), 149–157.
- Noguchi, T., & Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, 132(1), 44–56.
- Noguchi, T., & Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review*.
- Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), 607.
- Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23), 3311–3325.

- Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4), 481–487.
- O'Toole, B., & Wenderoth, P. (1977). The tilt illusion: Repulsion and attraction effects in the oblique meridian. *Vision Research*, 17(3), 367–374.
- Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, 29(44), 14004–14014.
- Parducci, A. (1965). Category judgment: a range-frequency model. *Psychological Review*, 72(6), 407.
- Parducci, A. (1974). Contextual effects: A range-frequency analysis. *Handbook of Perception*, 2, 127–141.
- Park, I. M., & Pillow, J. W. (2017). Bayesian efficient coding. *BioRxiv*, 178418.
- Pettibone, J. C. (2012). Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgment and Decision Making*, 7(4), 513.
- Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 19(5), 285–293.
- Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741), 233.
- Pleskac, T. J. (2012). Decision and choice: Luce's choice axiom. *International Encyclopedia of Social and Behavioral Sciences*.
- Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., & Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5), 503–519.
- Polania, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, 22(1), 134.
- Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2), 125.
- Rahnev, D., & Denison, R. N. (2018). Suboptimality in Perceptual Decision Making. *Behavioral and Brain Sciences*, 1–107.
- Ramachandran, V. S. (1988). Perception of shape from shading. *Nature*, 331(6152), 163.
- Rangel, A., & Clithero, J. A. (2012). Value normalization in decision making: theory and evidence. *Curr Opin Neurobiol*, 22(6), 970–81. doi:10.1016/j.conb.2012.07.011
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922.
- Reinagel, P. (2001). How do visual neurons respond in the real world? *Current Opinion in Neurobiology*, 11(4), 437–442.
- Reutsckaja, E., Lindner, A., Nagel, R., Andersen, R. A., & Camerer, C. F. (2018). Choice overload reduces neural signatures of choice set value in dorsal striatum and anterior cingulate cortex. *Nature Human Behaviour*, 2(12), 925.
- Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, 61(2), 168–185.
- Rigoli, F., Mathys, C., Friston, K. J., & Dolan, R. J. (2017). A unifying Bayesian account of contextual effects in value-based choice. *PLOS Computational Biology*, 13(10), e1005769.
- Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2), 370.
- Rouder, J. N., & King, J. W. (2003). Flanker and negative flanker effects in letter identification. *Percept Psychophys*, 65(2), 287–97.
- Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, 70(6), 1054–69. doi:10.1016/j.neuron.2011.05.014
- Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12), 883–893.
- Sanborn, S., Bourgin, D. D., Chang, M., & Griffiths, T. L. (2018). Representational efficiency outweighs action efficiency in human program induction. *ArXiv Preprint ArXiv:1807.07134*.
- Schoemaker, P. J. (1991). The quest for optimality: A positive heuristic of science? *Behavioral and Brain Sciences*, 14(2), 205–215.
- Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, 8(7), 522.
- Schwartz, O., Sejnowski, T. J., & Dayan, P. (2009). Perceptual organization in the tilt illusion. *J Vis*, 9(4), 19 1-20. doi:10.1167/9.4.19

- Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8), 819.
- Scolari, M., & Serences, J. T. (2009). Adaptive allocation of attentional gain. *Journal of Neuroscience*, 29(38), 11933–11942.
- Shafir, S., Waite, T. A., & Smith, B. H. (2002). Context-dependent violations of rational choice in honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*). *Behavioral Ecology and Sociobiology*, 51(2), 180–187.
- Shahnazian, D., & Holroyd, C. B. (2017). Distributed representations of action sequences in anterior cingulate cortex: A recurrent neural network approach. *Psychonomic Bulletin & Review*, 1–20.
- Shannon, C. E., & Weaver, W. (1963). The mathematical theory of communication. 1949. *Urbana, IL: University of Illinois Press*.
- Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2), 217–240.
- Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nat Neurosci*, 19(10), 1286–91. doi:10.1038/nn.4384
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2), 129.
- Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society*, 1(1), 25–39.
- Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu Rev Neurosci*, 24, 1193–216. doi:10.1146/annurev.neuro.24.1.1193
- Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16(2), 158–174.
- Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3), 665–690.
- Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, 152, 181–198.
- Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, 360(6389), 652–656.
- Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, 119(4), 807.
- Solomon, J. A., May, K. A., & Tyler, C. W. (2016). Inefficiency of orientation averaging: Evidence for hybrid serial/parallel temporal integration. *Journal of Vision*, 16(1), 13.
- Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, 11(12), 13–13.
- Solomon, J. A., & Morgan, M. J. (2006). Stochastic re-calibration: contextual effects on perceived tilt. *Proc Biol Sci*, 273(1601), 2681–6. doi:10.1098/rspb.2006.3634
- Solomon, J. A., & Morgan, M. J. (2009). Strong tilt illusions always reduce orientation acuity. *Vision Research*, 49(8), 819–824.
- Soltani, A., De Martino, B., & Camerer, C. (2012). A range-normalization model of context-dependent choice: a new model and evidence. *PLoS Computational Biology*, 8(7), e1002607.
- Spektor, M. S., Kellen, D., & Hotaling, J. M. (2018). When the Good Looks Bad: An Experimental Exploration of the Repulsion Effect. *Psychological Science*, 0956797618779041.
- Stemmler, M., Usher, M., & Niebur, E. (1995). Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. *Science*, 269(5232), 1877–80.
- Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, 46(4), 1004–1017.
- Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, 53(1), 1–26.
- Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4), 578.
- Stone, J. V. (2018). Principles of Neural Information Theory: Computational Neuroscience and Metabolic Efficiency (1st ed.). Sebtel Press.
- Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11), 745.
- Summerfield, C., & Li, V. (2018). Perceptual suboptimality: bug or feature? A commentary on Rahnev and Denison - Suboptimality in Perceptual Decision Making. *Behavioral and Brain Sciences*, 41(e245). doi:https://doi.org/10.1017/S0140525X18001437
- Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences*, 19(1), 27–34.
- Sutton, R. S., & Barto, A. G. (2011). Reinforcement learning: An introduction.

- Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *BioRxiv*, 595843.
- Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6), 401.
- Tomassini, A., Morgan, M. J., & Solomon, J. A. (2010). Orientation uncertainty reduces perceived obliquity. *Vision Research*, 50(5), 541–547.
- Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736), 575.
- Trueblood, J. S. (2012). Multialternative context effects obtained using an inference task. *Psychonomic Bulletin & Review*, 19(5), 962–968.
- Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, 121(2), 179.
- Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences*, 109(24), 9659–9664.
- Tsetsos, K., Moran, R., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, 113(11), 3102–3107.
- Tsetsos, K., Usher, M., & Chater, N. (2010). Preference reversal in multiattribute choice. *Psychol Rev*, 117(4), 1275–93. doi:10.1037/a0020580
- Tsetsos, K., Wyart, V., Shorkey, S. P., & Summerfield, C. (2014). Neural mechanisms of economic commitment in the human medial prefrontal cortex. *ELife*, 3, e03701.
- Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2017). Competing theories of multialternative, multiattribute preferential choice.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, S251–S278.
- Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4), 1039–1061.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323.
- Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, 111(3), 757.
- Van den Berg, R., & Ma, W. J. (2012). Robust averaging during perceptual judgment is not optimal. *Proceedings of the National Academy of Sciences*, 109(13), E736–E736.
- van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, 7, e34963.
- Van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*, 14(4), 593–602.
- Vandormael, H. (2017). Optimal decision agents? Biases during voluntary information sampling using eye-tracking (PhD Thesis). University of Oxford.
- Vandormael, H., Castañón, S. H., Balaguer, J., Li, V., & Summerfield, C. (2017). Robust sampling of decision information during perceptual choice. *Proceedings of the National Academy of Sciences*, 201613950.
- Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1), 92.
- Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik* (Vol. 9). Voss.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ.: Princeton University Press.
- Wald, A., & Wolfowitz, J. (1949). Bayes Solutions of Sequential Decision Problems. *Proc Natl Acad Sci U S A*, 35(2), 99–102.
- Wang, Z., Stocker, A. A., & Lee, D. D. (2016). Efficient neural codes that minimize lp reconstruction error. *Neural Computation*, 28(12), 2656–2686.
- Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, 18(10), 1509.
- Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, 114(38), 10244–10249.

- Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6), 598.
- Westrick, Z. M., Heeger, D. J., & Landy, M. S. (2016). Pattern adaptation and normalization reweighting. *Journal of Neuroscience*, 36(38), 9805–9816.
- Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, 8(2), 227–233.
- Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, 17(3), 431.
- Wu, T., Dufford, A. J., Mackie, M. A., Egan, L. J., & Fan, J. (2016). The Capacity of Cognitive Control Estimated from a Perceptual Decision Making Task. *Scientific Reports*, 6.
- Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences*, 106(40), 17199–17204.
- Yamada, H., Louie, K., Tymula, A., & Glimcher, P. W. (2018). Free choice shapes normalized value signals in medial orbitofrontal cortex. *Nature Communications*, 9(1), 162.
- Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. *J Exp Psychol Hum Percept Perform*, 35(3), 700–17. doi:10.1037/a0013553
- Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7), 301–308.
- Zhang, K., Ginzburg, I., McNaughton, B. L., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, 79(2), 1017–1044.

## Appendix

**Appendix Table 3.1. ANOVA results**

| Variables                 | d.f. | F values | p-value     |
|---------------------------|------|----------|-------------|
| Mean                      | 20   | 99.37    | $p < 0.001$ |
| Variance                  |      | 292.59   | $p < 0.001$ |
| Session                   |      | 108.41   | $p < 0.001$ |
| Mean x Variance           |      | 39.2     | $p < 0.001$ |
| Mean x Session            |      | 4.14     | $p = 0.55$  |
| Variance x Session        |      | 2.23     | $p = 0.15$  |
| Mean x Variance x Session |      | 4.57     | $p = 0.045$ |

**Appendix Table 4.1. Summary of ANOVA results on RT.<sup>^</sup>**

| Experiment | Variable                                                       | Degrees of freedom | F value | $\eta_p^2$ | p-value           |
|------------|----------------------------------------------------------------|--------------------|---------|------------|-------------------|
| 1a         | Cong                                                           | 1, 36              | 4.38    | 0.11       | $p = 0.043^*$     |
|            | $sX^j$                                                         | 1.87,67.14         | 3.26    | 0.08       | $p = 0.048^*$     |
|            | $\text{Cong} \times sX^j$                                      | 1.76,63.39         | 9.72    | 0.21       | $p < 0.001^{***}$ |
| 1b         | Cong                                                           | 1,34               | 22.56   | 0.4        | $p < 0.001^{***}$ |
|            | $sX^j$                                                         | 1.95,66.17         | 3.37    | 0.09       | $p = 0.042^*$     |
|            | $\text{Cong} \times sX^j$                                      | 1.96,66.79         | 9.67    | 0.22       | $p < 0.001^{***}$ |
| 2a         | Cong                                                           | 1,26               | 1.51    | --         | $p = 0.23$        |
|            | $sX^j$                                                         | 1.76,45.73         | 11.58   | 0.31       | $p < 0.001^{***}$ |
|            | $ X^i $                                                        | 1.81,46.97         | 72.55   | 0.74       | $p < 0.001^{***}$ |
|            | $ \overline{X}^j $                                             | 1.25,32.53         | 108.211 | 0.81       | $p < 0.001^*$     |
|            | $\text{Cong} \times sX^j$                                      | 1.62,42.2          | 1.41    | --         | $p = 0.25$        |
|            | $\text{Cong} \times  X^i $                                     | 1.45,37.66         | 21.33   | 0.45       | $p < 0.001^{***}$ |
|            | $\text{Cong} \times  \overline{X}^j $                          | 1.54,40.01         | 65.72   | 0.72       | $p < 0.001^{***}$ |
|            | $sX^j \times  X^i $                                            | 3.41,88.61         | 25.71   | 0.5        | $p < 0.001^{***}$ |
|            | $sX^j \times  \overline{X}^j $                                 | 3.06,79.46         | 10.31   | 0.28       | $p < 0.001^{***}$ |
|            | $ X^i  \times  \overline{X}^j $                                | 1.52,91.6          | 12.04   | 0.32       | $p < 0.001^{***}$ |
|            | $\text{Cong} \times  X^i  \times  \overline{X}^j $             | 3.42,88.88         | 10.18   | 0.28       | $p < 0.001^{***}$ |
|            | $sX^j \times  X^i  \times  \overline{X}^j $                    | 5.36,139.32        | 5.19    | 0.17       | $p < 0.001^{***}$ |
|            | $\text{Cong} \times  X^i  \times sX^j$                         | 3.14,81.7          | 6.3     | 0.19       | $p = 0.001^{***}$ |
|            | $\text{Cong} \times  \overline{X}^j  \times sX^j$              | 3.43,89.1          | 15.42   | 0.37       | $p < 0.001^{***}$ |
|            | $\text{Cong} \times  X^i  \times  \overline{X}^j  \times sX^j$ | 5.34,138.84        | 5.48    | 0.17       | $p < 0.001^{***}$ |
| 2b         | Cong                                                           | 1,29               | 1.57    | --         | $p = 0.22$        |
|            | $sX^j$                                                         | 1.98,57.55         | 13.78   | 0.32       | $p < 0.001^{***}$ |
|            | $ X^i $                                                        | 1.85,53.59         | 11.01   | 0.28       | $p < 0.001^{***}$ |
|            | $ \overline{X}^j $                                             | 1.45,42.01         | 134.79  | 0.82       | $p < 0.001^*$     |
|            | $\text{Cong} \times sX^j$                                      | 1.82,52.85         | 14.13   | 0.33       | $p < 0.001^{***}$ |
|            | $\text{Cong} \times  X^i $                                     | 1.67,48.5          | 5.76    | 0.17       | $p = 0.007^{**}$  |
|            | $\text{Cong} \times  \overline{X}^j $                          | 1.84,53.28         | 5.9     | 0.17       | $p = 0.006^{**}$  |
|            | $sX^j \times  X^i $                                            | 3.59,104.07        | 3.46    | 0.11       | $p = 0.013^*$     |
|            | $sX^j \times  \overline{X}^j $                                 | 2.94,85.29         | 0.4     | --         | $p = 0.75$        |
|            | $ X^i  \times  \overline{X}^j $                                | 3.38,98.02         | 2.45    | --         | $p = 0.061$       |
|            | $\text{Cong} \times  X^i  \times  \overline{X}^j $             | 3.64,105.64        | 6.48    | 0.18       | $p < 0.001^{***}$ |
|            | $sX^j \times  X^i  \times  \overline{X}^j $                    | 6.32,183.36        | 2.94    | 0.09       | $p = 0.008^{**}$  |
|            | $\text{Cong} \times  X^i  \times sX^j$                         | 3.48,100.95        | 2.43    | --         | $p = 0.061$       |
|            | $\text{Cong} \times  \overline{X}^j  \times sX^j$              | 3.1,89.85          | 4.35    | 0.13       | $p = 0.006^{**}$  |
|            | $\text{Cong} \times  X^i  \times  \overline{X}^j  \times sX^j$ | 5.67,164.48        | 4.03    | 0.12       | $p = 0.001^{***}$ |

<sup>^</sup>ANOVA statistics for congruency (Cong), flanker variance ( $sX^j$ ), target decision variable ( $X^i$ ), flanker decision variable ( $\overline{X}^j$ ), and their interactions on reaction time for Experiment 1 and 2.  $^*p < 0.05$ ,  $^{**}p < 0.01$ ,  $^{***}p < 0.001$

**Appendix Table 4.2. Summary of ANOVA results on accuracy.**

| Experiment | Variable                                             | Degrees of freedom | F value | $\eta_p^2$ | p-value           |
|------------|------------------------------------------------------|--------------------|---------|------------|-------------------|
| 1a         | Cong                                                 | 1, 36              | 3.14    | --         | $p = 0.085$       |
|            | $sX^j$                                               | 1.9,68.41          | 0.48    | --         | $p = 0.61$        |
|            | Cong $\times$ $sX^j$                                 | 1.91,68.89         | 3.5     | 0.09       | $p = 0.038^*$     |
| 1b         | Cong                                                 | 1,34               | 21.24   | 0.38       | $p < 0.001^{***}$ |
|            | $sX^j$                                               | 1.96,66.48         | 0.14    | --         | $p = 0.87$        |
|            | Cong $\times$ $sX^j$                                 | 1.67,60.95         | 2.56    | --         | $p = 0.09$        |
| 2a         | Cong                                                 | 1,26               | 1       | --         | $p = 0.33$        |
|            | $sX^j$                                               | 1.93,50.15         | 2.96    | --         | $p = 0.063$       |
|            | $ X^i $                                              | 1.98,51.5          | 17.27   | 0.4        | $p < 0.001^{***}$ |
|            | $ \bar{X}^j $                                        | 1.73,45.03         | 34.57   | 0.57       | $p < 0.001^*$     |
|            | Cong $\times$ $sX^j$                                 | 1.62,42.23         | 0.84    | --         | $p = 0.42$        |
|            | Cong $\times$ $ X^i $                                | 1.8,46.9           | 4.88    | 0.16       | $p = 0.014^*$     |
|            | Cong $\times$ $ \bar{X}^j $                          | 1.66,43.28         | 21.02   | 0.45       | $p < 0.001^{***}$ |
|            | $sX^j \times  X^i $                                  | 3.36,87.37         | 2.4     | --         | $p = 0.067$       |
|            | $sX^j \times  \bar{X}^j $                            | 3.39,88.12         | 4.27    | 0.14       | $p = 0.005^{**}$  |
|            | $ X^i  \times  \bar{X}^j $                           | 3.24,84.36         | 3.72    | 0.13       | $p = 0.012^*$     |
|            | Cong $\times$ $ X^i  \times  \bar{X}^j $             | 2.68,69.8          | 2.18    | --         | $p = 0.1$         |
|            | $sX^j \times  X^i  \times  \bar{X}^j $               | 6.04,157.15        | 3.71    | 0.12       | $p = 0.002$       |
|            | Cong $\times$ $ X^i  \times sX^j$                    | 2.82,73.43         | 1.1     | --         | $p = 0.35$        |
|            | Cong $\times$ $ \bar{X}^j  \times sX^j$              | 3.22,83.79         | 4.35    | 0.14       | $p = 0.006^{**}$  |
|            | Cong $\times$ $ X^i  \times  \bar{X}^j  \times sX^j$ | 4.8,124.9          | 1.2     | --         | $p = 0.31$        |
| 2b         | Cong                                                 | 1,29               | 0.48    | --         | $p = 0.5$         |
|            | $sX^j$                                               | 1.62,46.98         | 10.46   | 0.27       | $p < 0.001^{***}$ |
|            | $ X^i $                                              | 1.95,56.61         | 2.81    | --         | $p = 0.07$        |
|            | $ \bar{X}^j $                                        | 1.22,35.45         | 166.1   | 0.85       | $p < 0.001^*$     |
|            | Cong $\times$ $sX^j$                                 | 1.92,55.54         | 2.17    | --         | $p = 0.13$        |
|            | Cong $\times$ $ X^i $                                | 1.97,57.14         | 3.02    | --         | $p = 0.057$       |
|            | Cong $\times$ $ \bar{X}^j $                          | 1.27,36.74         | 0.55    | --         | $p = 0.5$         |
|            | $sX^j \times  X^i $                                  | 3.75,108.8         | 1.08    | --         | $p = 0.37$        |
|            | $sX^j \times  \bar{X}^j $                            | 3.08,89.33         | 3.06    | 0.095      | $p = 0.031^*$     |
|            | $ X^i  \times  \bar{X}^j $                           | 3.04,88.13         | 0.87    | --         | $p = 0.46$        |
|            | Cong $\times$ $ X^i  \times  \bar{X}^j $             | 2.96,85.95         | 2.59    | --         | $p = 0.059$       |
|            | $sX^j \times  X^i  \times  \bar{X}^j $               | 5.28,153.19        | 1.79    | --         | $p = 0.11$        |
|            | Cong $\times$ $ X^i  \times sX^j$                    | 3.41,98.94         | 1.18    | --         | $p = 0.32$        |
|            | Cong $\times$ $ \bar{X}^j  \times sX^j$              | 3.19,92.6          | 0.05    | --         | $p = 0.99$        |
|            | Cong $\times$ $ X^i  \times  \bar{X}^j  \times sX^j$ | 5.5,159.64         | 1.17    | --         | $p = 0.32$        |

^ ANOVA statistics for congruency (Cong), flanker variance ( $sX^j$ ), target decision variable ( $X^i$ ), flanker decision variable ( $\bar{X}^j$ ), and their interactions on accuracy for Experiment 1 and 2.  $*p < 0.05$ ,  $**p < 0.01$ ,  $***p < 0.001$

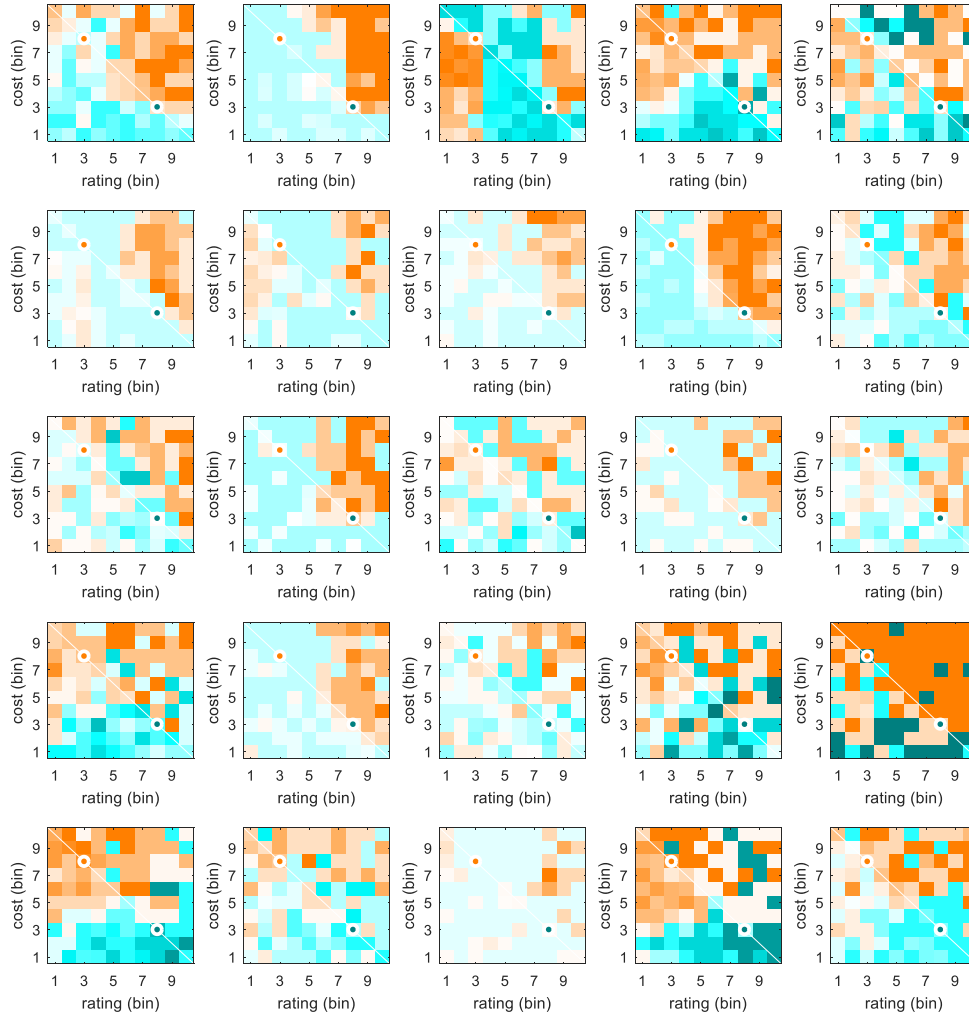
**Appendix Table 4.3. Voxels correlating with  $1/\widehat{X}^t$  after partialling out reaction time from GLM3 at a threshold of  $p < 0.05$  (FDR corrected) with a cluster extent  $> 10$  voxels.**

| Cluster FDR corrected p | Number of voxels in cluster | Peak t-value | X   | Y   | Z  | Anatomical label               |
|-------------------------|-----------------------------|--------------|-----|-----|----|--------------------------------|
| 0.038                   | 41                          | 5.96         | -30 | 16  | 10 | Left Insula                    |
|                         |                             | 5.83         | -38 | 20  | 2  |                                |
| 0.042                   | 33                          | 5.05         | 34  | 24  | 6  | Right Insula                   |
|                         |                             | 4.65         | 34  | 24  | -6 |                                |
| 0.046                   | 28                          | 4.47         | 6   | 20  | 50 | Anterior cingulate             |
|                         |                             | 4.32         | 10  | 24  | 38 |                                |
| 0.000                   | 103                         | 7.03         | -42 | -56 | 26 | Left inferior parietal lobule  |
|                         |                             | 6.43         | -42 | -64 | 30 |                                |
|                         |                             | 5.36         | -42 | -76 | 34 |                                |
| 0.003                   | 68                          | 5.69         | -18 | 20  | 54 | Superior frontal gyrus         |
|                         |                             | 4.95         | -18 | 44  | 42 |                                |
|                         |                             | 4.55         | -14 | 56  | 30 |                                |
| 0.000                   | 146                         | 5.51         | 2   | -44 | 38 | Posterior cingulate            |
|                         |                             | 5.31         | -6  | -52 | 14 |                                |
|                         |                             | 5.05         | 14  | -40 | 34 |                                |
| 0.036                   | 35                          | 5.08         | 54  | -68 | 26 | Right inferior parietal lobule |
|                         |                             | 4.94         | 50  | -60 | 26 |                                |
|                         |                             | 4.65         | 62  | -60 | 22 |                                |

**Appendix Table 4.4. Significant voxels for different regressors from GLM4 at a threshold of  $p < 0.05$  (FDR corrected) and with a cluster extent  $>10$  voxels.**

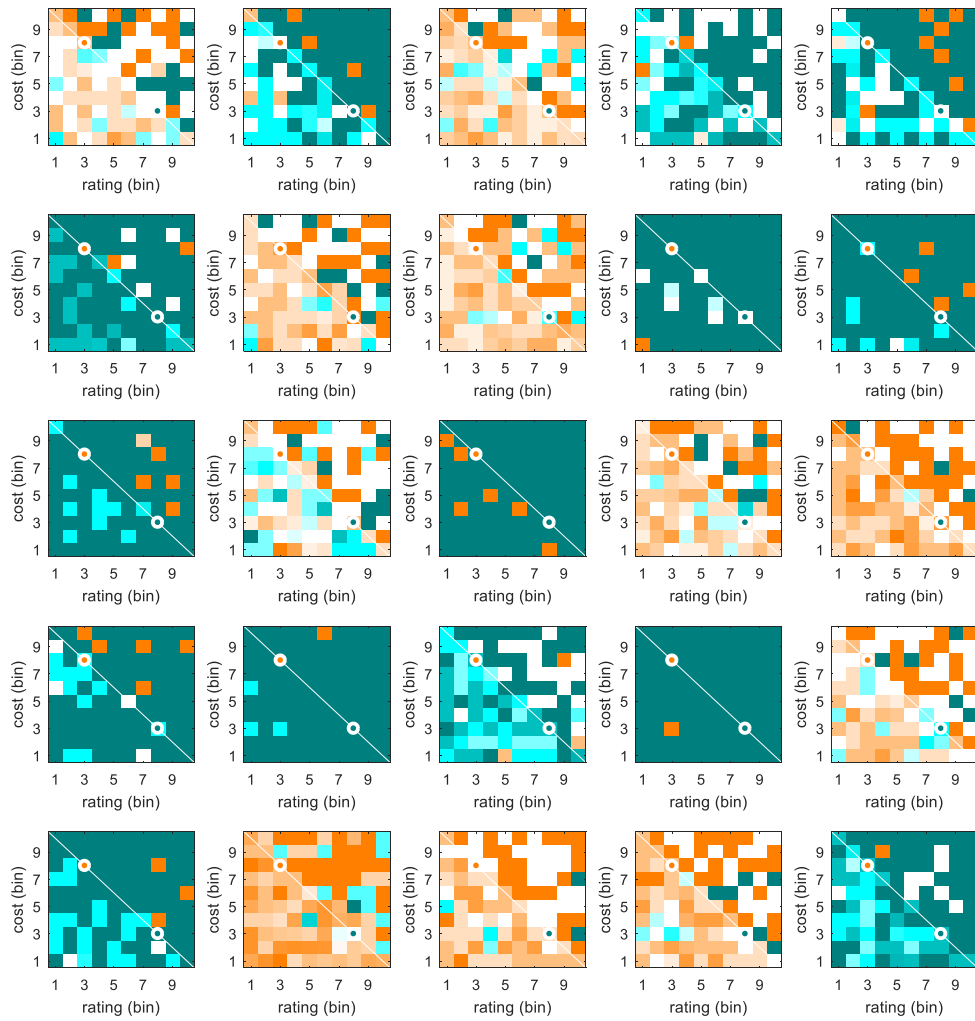
| Regressors    | Cluster FDR corrected p | Number of voxels in cluster | Peak t-value | X   | Y   | Z   | Anatomical label                                                   |
|---------------|-------------------------|-----------------------------|--------------|-----|-----|-----|--------------------------------------------------------------------|
| $ X^i $       | 0.000                   | 155                         | 8.44         | -2  | -52 | 30  | Posterior cingulate                                                |
|               |                         |                             | 6.65         | -6  | -52 | 18  |                                                                    |
|               |                         |                             | 5.11         | 10  | -48 | 38  |                                                                    |
|               | 0.000                   | 168                         | 8.19         | -2  | 40  | -10 | Ventromedial prefrontal                                            |
|               |                         |                             | 7.01         | 6   | 52  | -10 |                                                                    |
|               |                         |                             | 6.91         | -14 | 60  | 14  |                                                                    |
|               | 0.000                   | 84                          | 7.99         | -46 | -60 | 26  | Angular gyrus                                                      |
|               |                         |                             | 4.72         | -38 | -76 | 42  |                                                                    |
|               |                         |                             | 4.71         | -58 | -56 | 22  |                                                                    |
|               | 0.000                   | 75                          | 6.98         | -18 | 24  | 42  | Superior frontal gyrus                                             |
|               |                         |                             | 5.50         | -14 | 44  | 42  |                                                                    |
|               | 0.003                   | 44                          | 6.16         | 54  | -64 | 26  | Temporoparietal junction                                           |
|               |                         |                             | 5.70         | 62  | -56 | 18  |                                                                    |
|               |                         |                             | 5.42         | 46  | -56 | 26  |                                                                    |
|               | 0.009                   | 35                          | 5.23         | 58  | -4  | -22 | Middle temporal gyrus                                              |
|               |                         |                             | 4.57         | 62  | -12 | -14 |                                                                    |
|               | 0.000                   | 156                         | 8.89         | 14  | -60 | 62  | Superior parietal lobule                                           |
|               |                         |                             | 8.45         | 54  | -32 | 50  |                                                                    |
|               |                         |                             | 8.17         | 26  | -48 | 46  |                                                                    |
|               | 0.000                   | 20                          | 8.39         | 30  | -4  | 50  | Supplementary motor area                                           |
|               |                         |                             | 5.81         | 18  | 0   | 58  |                                                                    |
|               | 0.000                   | 148                         | 8.11         | -2  | -20 | -2  | Midbrain                                                           |
|               |                         |                             | 8.01         | -14 | -16 | 14  |                                                                    |
|               |                         |                             | 7.72         | 10  | -16 | 6   |                                                                    |
|               | 0.000                   | 104                         | 8.00         | 50  | 8   | 26  | Insula                                                             |
|               |                         |                             | 7.88         | 34  | 24  | 2   |                                                                    |
|               |                         |                             | 7.39         | 50  | 20  | -6  |                                                                    |
|               | 0.000                   | 88                          | 7.76         | -6  | 12  | 42  | Anterior cingulate                                                 |
|               |                         |                             | 7.76         | 10  | 16  | 38  |                                                                    |
|               |                         |                             | 7.68         | 6   | 20  | 50  |                                                                    |
|               | 0.000                   | 63                          | 7.75         | -42 | 16  | 10  | Insula                                                             |
|               |                         |                             | 7.51         | -30 | 24  | 2   |                                                                    |
|               |                         |                             | 6.85         | -34 | -4  | -2  |                                                                    |
|               | 0.000                   | 25                          | 7.62         | -22 | -4  | 58  | Supplementary motor area                                           |
|               | 0.000                   | 35                          | 7.61         | -38 | -60 | -26 | Cerebellum                                                         |
|               |                         |                             | 7.03         | -30 | -60 | -22 |                                                                    |
|               |                         |                             | 6.53         | -34 | -52 | -34 |                                                                    |
|               | 0.000                   | 42                          | 7.46         | -38 | 0   | 30  | Supplementary motor area                                           |
|               |                         |                             | 7.18         | -42 | -4  | 38  |                                                                    |
|               | 0.000                   | 73                          | 7.32         | -18 | -64 | 54  | Spanning from Inferior parietal lobule to superior parietal lobule |
|               |                         |                             | 7.32         | -42 | -36 | 46  |                                                                    |
|               |                         |                             | 7.12         | -22 | -56 | 50  |                                                                    |
|               | 0.000                   | 14                          | 6.74         | -6  | -80 | -22 | Cerebellum                                                         |
|               |                         |                             | 6.45         | 2   | -76 | -22 |                                                                    |
| $ \bar{X}^j $ | 0.000                   | 128                         | 8.22         | 34  | -44 | 50  | Inferior parietal lobule                                           |
|               |                         |                             | 6.47         | 42  | -36 | 42  |                                                                    |

|                                        |       |     |      |     |     |    |                                                                    |
|----------------------------------------|-------|-----|------|-----|-----|----|--------------------------------------------------------------------|
|                                        | 0.000 | 167 | 5.67 | 26  | -64 | 34 | Inferior parietal lobule                                           |
|                                        |       |     | 7.12 | -46 | -32 | 46 |                                                                    |
|                                        |       |     | 6.87 | -26 | -76 | 30 |                                                                    |
|                                        |       |     | 6.54 | -30 | -40 | 46 |                                                                    |
|                                        | 0.045 | 23  | 5.23 | -30 | -4  | 54 | Supplementary motor area                                           |
|                                        |       |     | 4.18 | -22 | 4   | 58 |                                                                    |
|                                        | 0.021 | 28  | 5.09 | -38 | 16  | -2 | Left Insula                                                        |
|                                        |       |     | 4.44 | -30 | 28  | 10 |                                                                    |
|                                        | 0.033 | 25  | 4.39 | 30  | 24  | 10 | Right Insula                                                       |
|                                        |       |     |      |     |     |    |                                                                    |
| $\frac{ X^i - \bar{X} }{\times  X^i }$ | 0.000 | 176 | 7.24 | 26  | -60 | 50 | Superior parietal lobule                                           |
|                                        |       |     | 6.97 | 30  | -40 | 46 |                                                                    |
|                                        |       |     | 6.76 | 30  | -52 | 50 |                                                                    |
|                                        | 0.000 | 136 | 6.81 | 30  | 20  | 10 | Insula                                                             |
|                                        |       |     | 6.14 | 46  | 8   | 14 |                                                                    |
|                                        |       |     | 5.53 | 46  | 16  | 2  |                                                                    |
|                                        | 0.023 | 28  | 6.34 | 42  | 40  | 30 | dorsolateral prefrontal cortex                                     |
|                                        | 0.000 | 64  | 5.85 | -30 | 20  | 10 | Insula                                                             |
|                                        |       |     | 5.26 | -38 | 12  | -2 |                                                                    |
|                                        | 0.000 | 95  | 5.58 | -30 | -44 | 54 | Spanning from Inferior parietal lobule to Superior parietal lobule |
|                                        |       |     | 5.51 | -26 | -44 | 46 |                                                                    |
|                                        |       |     | 4.26 | -50 | -32 | 42 |                                                                    |
|                                        | 0.000 | 58  | 5.48 | 6   | 20  | 38 | Anterior cingulate                                                 |
|                                        |       |     | 5.45 | 6   | 12  | 54 |                                                                    |
|                                        |       |     | 5.39 | 6   | 16  | 46 |                                                                    |
|                                        | 0.007 | 36  | 4.98 | -34 | 0   | 30 | Supplementary motor area                                           |
|                                        |       |     | 4.41 | -46 | 0   | 42 |                                                                    |
|                                        | 0.049 | 23  | 4.87 | -2  | -20 | -2 | Midbrain                                                           |



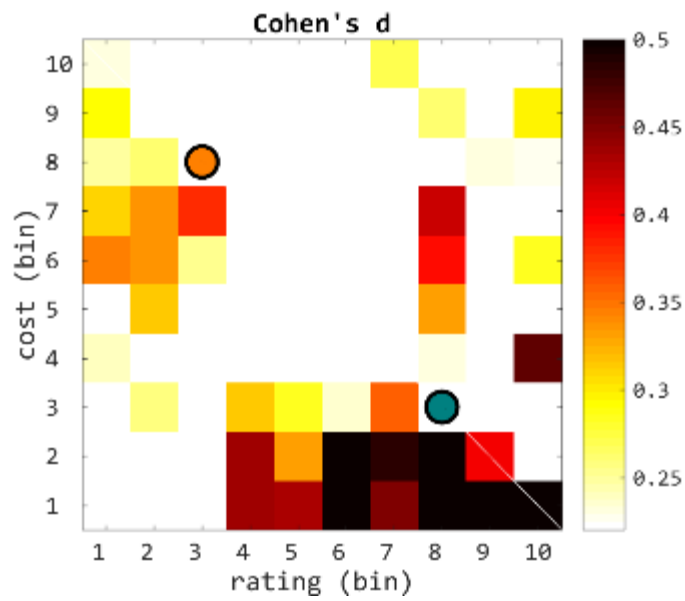
**Appendix Figure 5.1. Individual relative choice share.**

Each subjects' choice share is being correlated with the group choice share. 25 subjects who have the most similar choice share with the group average is being plotted here.



**Appendix Figure 5.2. Same as Appendix Figure 5.1**

Individual relative choice share of subjects who have little correlation on the choice pattern with the group-average.

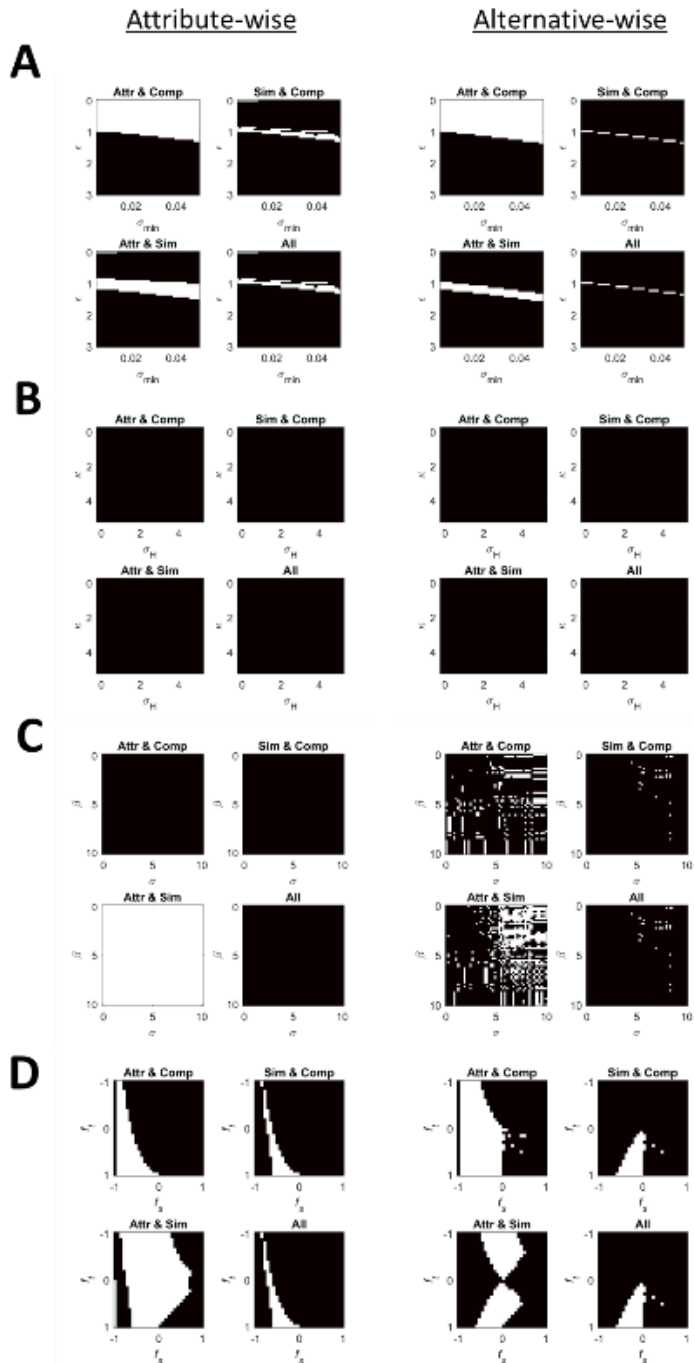


**Appendix Figure 5.3. Effect size on the relative choice share.**

Cohen's  $d$  is being computed on the relative choice share for each of the grids. Coloured grids represent the choice sets that achieve an effect size that has at least 95% statistical power given the sample size.

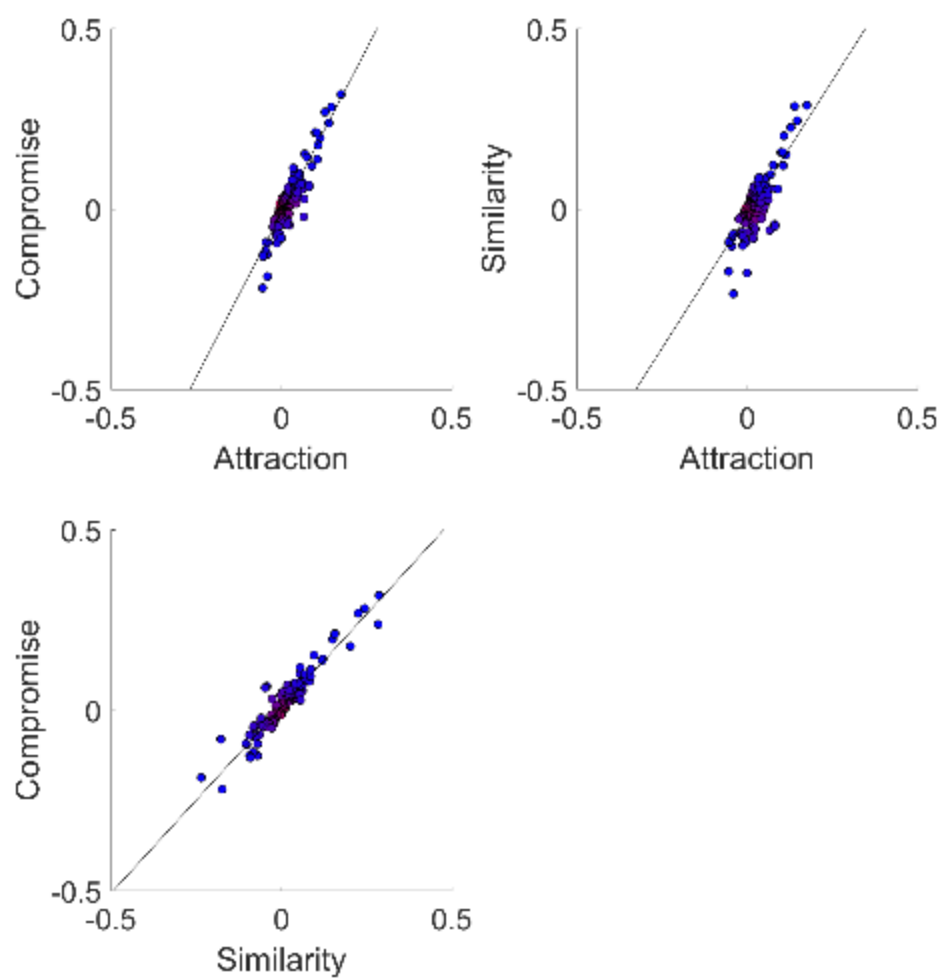
**Appendix Table 5.1. Parameters of each model for Figure 5.8.**

|                              | Models                 | Parameter search range          |                             |                        | Parameters             |                      |                   |
|------------------------------|------------------------|---------------------------------|-----------------------------|------------------------|------------------------|----------------------|-------------------|
| Attribute-wise Comparisons   | Probit                 | $\sigma_n = U(0.00001, 10)$     |                             |                        | $\sigma_n = 4.1$       |                      |                   |
|                              | Adaptive Gain          | $\sigma_{min} = U(0.005, 0.05)$ | $\sigma_{min} = U(0.01, 3)$ | $\sigma_n = U(0.1, 2)$ | $\sigma_{min} = 0.034$ | $\varepsilon = 0.99$ | $\sigma_n = 0.5$  |
|                              | Divisive Normalisation | $\kappa = U(0.01, 5)$           | $\sigma_H = U(0.01, 5)$     | $\sigma_n = 0.3$       | $\kappa = 5$           | $\sigma_H = 2.23$    | $\sigma_n = 0.3$  |
|                              | Range Adaptation       | $\beta = U(0.0025, 10)$         | $\sigma = U(0.0025, 10)$    | $\sigma_n = 5$         | $\beta = 0.0025$       | $\sigma = 0.0025$    | $\sigma_n = 5$    |
|                              | Range Normalisation    | $f_t = U(-0.99, 1)$             | $f_s = U(-0.99, 1)$         | $\sigma_n = 0.5$       | $f_t = -0.59$          | $f_s = -0.45$        | $\sigma_n = 0.5$  |
| Alternative-wise Comparisons | Probit                 | $\sigma_n = U(0.00001, 10)$     |                             |                        | N/A                    |                      |                   |
|                              | Adaptive Gain          | $\sigma_{min} = U(0.005, 0.05)$ | $\sigma_{min} = U(0.01, 3)$ | $\sigma_n = U(0.1, 2)$ | $\sigma_{min} = 0.08$  | $\varepsilon = 0.99$ | $\sigma_n = 0.58$ |
|                              | Divisive Normalisation | $\kappa = U(0.01, 5)$           | $\sigma_H = U(0.01, 5)$     | $\sigma_n = 0.3$       | $\kappa = 5$           | $\sigma_H = 2.23$    | $\sigma_n = 0.3$  |
|                              | Range Adaptation       | $\beta = U(0.0025, 10)$         | $\sigma = U(0.0025, 10)$    | $\sigma_n = 5$         | $\beta = 2.72$         | $\sigma = 6.5$       | $\sigma_n = 5$    |
|                              | Range Normalisation    | $f_t = U(-0.99, 1)$             | $f_s = U(-0.99, 1)$         | $\sigma_n = 0.5$       | $f_t = -0.59$          | $f_s = -0.45$        | $\sigma_n = 0.5$  |



**Appendix Figure 5.4. Parameters of each model that can predict combinations of decoy effects.**

“Attr” = Attraction; “Comp” = Compromise; “Similarity” = Similarity decoys. “All” refers to a prediction on all three decoy effects. White shaded area refers to the combination of parameters (x and y axes) that would predict specific combinations of decoy effects (title).



**Appendix Figure 5.5. Recovered correlation across decoy types using the adaptive gain model**

Same figure as Figure 4.16