THE INTERACTION BETWEEN TASK GOALS AND THE REPRESENTATION OF CHOICE OPTIONS IN DECISION-MAKING

Sebastian Bobadilla Suarez
Department of Experimental Psychology
University College London

Thesis submitted for the degree of
Doctor of Philosophy (PhD)
September 2017

Abstract

Most decision-making studies will focus on the value and uncertainty of each choice option but do not focus on the importance of the representation of the choice option itself. This thesis presents the effects that task goals have on creating the appropriate cognitive representation to achieve those goals and how these representations are dependent on the informational input within a given task. The overall hypothesis for this work is that cognitive representations reveal a trade-off between accommodating task goals and the format of the information sampled from the environment of the task. For example, ordering books in a stand in alphabetical order, to facilitate the task of retrieving a relevant one when necessary, reveals a material implication of such cognitive representations. As in many situations, the internal representations constructed by the agent embody the remaining degrees of freedom that map the input to successful task completion. The first two chapters in this work present how uncertain beliefs about ourselves and our preferences are either integrated or compared to fixed information about other agent's beliefs. The third chapter presents the direct manipulation of representations of choice options by changing both the stimuli and controlling for the decision strategy used by the decision-makers. The fourth chapter presents how choice options themselves are represented in the human brain. The findings related to 1) the adaptation of personal preferences and beliefs to the (fixed) preferences and beliefs of other agents, 2) observed reduction in decision strategy compliance contingent on stimulus format, and 3) the task-contingent results for similarities between brain states of choice options, support the general trade-off hypothesis. The conclusion that can be drawn is that the study of choice option representations is underdetermined unless both informational input and task goals are accounted for.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

Signature:

London, January 3rd, 2018

(Sebastian Bobadilla Suarez)

Acknowledgements

Dedicated to my family back home, especially my mom Paty,
in deep gratitude for their incredible support.

I would like to thank my supervisor, Brad Love, and the whole Love lab for their help, the interesting conversations and also for the great moments outside our workplace. Further thanks to Tali Sharot and all the people at the Affective Brain Lab who also helped me every step of the way. Thanks to my collaborators in all the projects pursued. Also, I would like to thank the people I met at The Alan Turing Institue during my final year who really broadened my horizons.

Lastly, I would like to thank all the people who ensured I had such a great time during my PhD.

The work presented here was funded by a scholarship from Consejo Nacional de Ciencia y Tecnologia (CONACYT) and an enrichment year stipend from The Alan Turing Institute.

Published and in prep. articles

Chapters 2, 3, 4, and 5 are partially or fully based on the following articles:

Chapter 2: Bobadilla-Suarez, S., Sunstein, C. R. & Sharot, T. (2017). The intrinsic value of control: The propensity to under-delegate in the face of potential gains and losses. *Journal of Risk and Uncertainty, 54*(3), 187-202

Chapter 3: De Martino, B., Bobadilla-Suarez, S., Noguchi, T., Sharot, T., & Love, B. C. (2017). Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. *Journal of Neuroscience, 37*(25), 6066-6074.

Chapter 4: Bobadilla-Suarez, S., & Love, B. C. (2017, May 29). Fast or Frugal, but Not Both: Decision Heuristics Under Time Pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000419

Chapter 5: Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A. & Love, B. C. (in prep). Neural measures of similarity.

I declare that I was fully involved through the whole process of investigation for these articles which includes matters regarding experimental design, analysis, and writing of the articles themselves.

Regarding the article used for Chapter 3, I was assigned as a second author given that Dr. Benedetto De Martino took the lead on the project as was previously agreed at the beginning of the investigation. Thus, Dr. Benedetto De Martino was first lead for the fMRI data analysis. Similarly, Prof. Bradley Love and Dr. Takao Noguchi were leads on the Bayesian update model.

Contents

List of figures

List of equations

List of tables

Chapter 1 The representation of choice options in decision-making

1.1 Introduction

Suppose you wish to multiply the numbers thirteen and forty-seven. Would you rather do such an operation using Arabic or Roman numerals? Within the decimal or hexadecimal system? The answer to this question may be explained by the desiderata of Western education but it can also be related to the goals of the task. For example, the binary number 1111111111111110011 can be represented as FFFF3 in the hexadecimal system. Such a shorthand representation may make it easier to communicate to another person (i.e., an IT professional) that a certain address in your computer's memory has been damaged. Other examples could apply to more everyday choices. If you want to choose an ice cream from the ice cream truck, the vendor might list the ingredients in layman's terms (like milk and chocolate), and not as a list of the constituent chemical components. It is clear from these examples that using an inappropriate representation, given the goal of the task, may lead to errors. The representation might lead you to choose an ice cream that is not really your favourite one or to communicate the wrong memory address to the IT professional. These cases illustrate that achieving a goal satisfactorily depends on the interaction between task goals, informational input, and the representation of choice options. Evidently, such a relationship can directly have an impact on the final choice outcome for the decision maker.

Decision-making studies largely study two properties of choice options: value and uncertainty (Kahneman & Tversky, 1982; Payzan-LeNestour, Dunne, Bossaerts, & O'Doherty, 2013; Rangel, Camerer, & Montague, 2008). Focusing only on value and uncertainty abstracts away other properties of choice options such as salience, valence and arousal, or conceptual knowledge and familiarity; features that are mainly addressed in other literatures like perceptual studies (Gottlieb, Kusunoki, & Goldberg, 1998; Reynolds & Desimone, 2003; Thompson & Bichot, 2005), studies on emotion (Gerber et al., 2008; Lane, Chua, & Dolan, 1999), and studies on memory (Henson, Rugg, Shallice, Josephs, & Dolan, 1999; Martin, 2007; Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997; Yonelinas, Otten, Shaw, & Rugg, 2005), respectively.

Furthermore, decision-making research has often been concerned with how information is processed from the environment and used to make a choice. Much focus has been put on

detailing the algorithms that people may use for making decisions depending on the goals of the decision-maker (Busemeyer & Johnson, 2004; Glimcher & Fehr, 2013). But more work is needed on detailing how those algorithms rely on the representation of choice options. An example of this is how frequency formats can avoid base rate neglect (Gigerenzer & Hoffrage, 1995), turning the interpretation of a judgment bias into a representational issue. Marr famously accentuated the difference between modelling the goal of an intelligent agent and describing the algorithms used to achieve that goal (Marr, 1982). Moreover, he emphasised the intimate relation between the algorithm and the representation of the information on which such an algorithm would operate on.

1.2 Dissertation objective

The aim for this present work is to characterise the factors than can affect the relation between representation of choice options, informational input, and task goals. A two-tiered approach is used; the first part of this work focuses on how expectations given by task constraints influence choice option representations (top-down view) and the second part focuses on how low-level stimulus features can influence them (bottom-up view). The top-down view of how task goals can invoke requirements on both algorithm (i.e., process) and representation is presented in chapters 2 and 3. Complementarily, the bottom-up perspective that stimuli format (i.e., informational input to the system) has a direct impact on the representation of choice options is addressed in chapters 4 and 5. From an experimental point of view, this can be achieved by careful consideration of both task and stimuli and how they interrelate. The introduction to the experiments outlined in subsequent chapters will commence by a review of general decision-making theories and the assumptions they impose onto algorithm and representation. This general review of decision-making theories aims to display that the focus of these theories relies mostly on the features of value and uncertainty while other features of choice options – related to stimulus format, context, domain, or task – are not considered or play a small role. To make sense of these theories, it will be necessary to discuss Marr's levels of abstraction for information processing systems.

1.3 Marr's three levels of abstraction

David Marr, in his book *Vision* (published posthumously in 1982), elegantly laid down the foundations for analysing any intelligent (i.e. information processing) system through three levels of theoretical abstraction:

1) Computational theory
   "What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?"
2) Representation and algorithm
   "How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?"
3) Hardware implementation
   "How can the representation and algorithm be realised physically?"

Marr's three levels of abstraction provide a two-fold perspective on issues regarding the representation of choice options. The first perspective clearly shows that the representation of choice options agrees with Marr's second level and is intimately related to the algorithmic transformations that an information processing system carries out in order to achieve a specific computational goal. In a decision-making setting, this means that the representation of choice options is difficult (if not impossible) to separate from the decision strategy. In this work, process and representation are two faces of the algorithmic level which give different views on the same phenomenon.[1] In this sense, process is dynamic; it implies sequential transformations, no matter if serial, parallel or distributed, but the notion of temporal order is important with regards to information processing. Complementarily, representation provides a more static view, with focus on the data and information structures at each time point, as opposed to the transformations between such structures.

The second perspective that Marr's framework offers is that the representation of choice options needs to be studied in the context of the task goal, addressed by Marr's first level, and grounded in plausible mechanisms of physical implementation (i.e., how do neuronal firing patterns implement the representation of choice options). The present work focuses on clarifying the relationship between task goal and the representation of choice options, as well as some stimulus-driven implementational aspects.

1.4 Complementary analytical axes

However influential Marr's three levels are in cognitive science, these levels often mix with other axes of analysis that frequently appear in studies in the field, such as:

[1] In fact, most times process and algorithm are treated as synonyms. In this work, process is distinguished from representation. On occasion, algorithm will be treated as synonymous with process but the emphasis will be denoted in brackets.

1) Goal directed (top-down) vs. stimulus-driven (bottom-up) processing
2) Normative vs. descriptive (or mechanistic) accounts of decision-making
3) Abstract (high level) to concrete (low level) choice option features

One could argue both for the differences or the similarities between these axes and how they emphasise different aspects of cognitive phenomena. For purposes of the present work, it is best to view the normative versus descriptive dichotomy (explained in the following sections) as equivalent to the computational versus algorithmic or representational levels, imbued with slightly different connotations (mostly related to the debate on rationality, explained below). However, the distinction between goal directed and stimulus-driven processing should be interpreted as orthogonal to Marr's levels of abstraction; the former relates to the agent's point of view whereas Marr's levels relate to the cognitive modeller's perspective. This can be clearly shown since goal directed processing can be studied through the physical implementation of the goal (i.e., how do neuronal firing patterns represent the goal?). Also, for purposes of this work, it is implied that top-down processes facilitate the study of abstract choice option features, such as sense of self, control, theory of mind, or personal preferences (chapters 2 and 3), commonly localised in more anterior brain regions (Craig & Craig, 2009; Frith & Frith, 2005; Hare, Camerer, & Rangel, 2009; Wagner, Maril, Bjork, & Schacter, 2001). Similarly, it is implied that bottom-up processes facilitate the study of concrete choice option features, such as numericity, colour-coding, serial ordering, and perceptual similarity (chapters 4 and 5), which are more likely to show effects in more posterior brain regions (Cavina-Pratesi, Kentridge, Heywood, & Milner, 2010; Harvey, Klein, Petridou, & Dumoulin, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Leonards, Sunaert, Van Hecke, & Orban, 2006).[2] These axes function as a way of structuring the approach to investigating the trade-off that choice option representations must handle between task goals and informational input, which includes matters regarding stimulus format.

1.5 Models and methods for each level of analysis

To understand what is meant by task goals it is necessary to describe the different approaches that decision-making theories employ when attempting to provide an explanation of their object of study. Such an understanding is best achieved by an inquiry into what is meant by the computational theory level of explanation which is addressed in the following section.

[2] Serial search is the exception here since it can invoke higher level processes (e.g., attentional control), associated with more anterior brain regions (Buschman & Miller, 2007).

Subsequent sections will present how Marr's computational level relates to the distinction between normative and descriptive accounts of decision-making before presenting theories posited at the algorithmic and representational level. Towards the end of this chapter, the methods used to study the implementational level will be introduced.

The aim of this chapter is to present the main cognitive models and methods used for each level of analysis relevant to this work: 1) at the computational (normative) level, theories such as Expected Utility Theory (EUT) (Von Neumann & Morgenstern, 2007) and Bayesian models of cognition (BMC) (Oaksford & Chater, 2001), 2) at the algorithmic and representational (descriptive) level, models such as Prospect Theory (Kahneman & Tversky, 1979), evidence accumulation models (e.g., Ratcliff, Smith, Brown, & McKoon, 2016), *fast and frugal* heuristics (Gigerenzer & Todd, 1999), and similarity models (e.g., Shepard, 1987; Tversky & Gati, 1982), and 3) at the implementation level, a brief overview of a technique used to study neuronal populations (functional resonance magnetic imaging, fMRI) and a relevant analysis technique known as representational similarity analysis (RSA) (Nili et al., 2014). This structuring is in alignment with the two-tiered approach for studying the representation of choice options. The top-down approach (chapters 2 and 3) relies on EUT and BMC as a means to study abstract features of choice options relevant to the social domain, especially pertaining to the sense of self, control, and personal – as opposed to social – preferences. The bottom-up approach (chapters 4 and 5) relies on *fast and frugal* heuristics and RSA (introduced in the following sections) to study more concrete features of choice options related to colour-coding, numericity, order of presentation (serial versus random), and similarity measures between stimuli.

The models and methods presented in the following sections serve two functions: as tools useful for conducting research on choice option representations and for displaying the fact that these models do not fully specify the role of such representations (which thus becomes responsibility of the experimentalist). The studies presented in the following chapters seek to specify such a role by detailing how the cognitive operations, needed at the algorithmic level, interact with specific aspects of the task goal or the stimulus format that facilitates or blocks that operation.

1.6 Computational level models of cognition and decision-making

When psychologists think about the computational level of decision-making and cognition, they will normally refer to a *normative* model that is optimal, in a sense, by some standard of rationality.[3] By definition, rational agents are expected to solve a task optimally as in Anderson's rational analysis (Anderson, 1991; Chater & Oaksford, 1999). More commonly, the *de facto* standard for rationality is heavily focused on the integration of evidence and uncertainty through Bayesian inference and estimation (Gershman, Horvitz, & Tenenbaum, 2015; Jones & Love, 2011; Mike Oaksford & Chater, 1998), although there are possibilities for Bayesian models to span multiple levels of abstraction (Griffiths, Lieder, & Goodman, 2015; Jones & Love, 2011). An example of using Bayesian inference at the purely computational level is applying Bayesian Decision Theory to provide an account for action selection and outcome evaluation (Körding & Wolpert, 2006). Similarly, the valuation process in decision-making is most commonly described by Expected Utility Theory (Von Neumann & Morgenstern, 2007), with extensions that allow for Bayesian computation of the expectations (explained in more detail below).

Rationality is important for cognitive models since it provides a benchmark to which agents can be compared to, either through consistency with a set of rules or through consistency through a chosen criterion (Pothos, Busemeyer, Shiffrin, & Yearsley, 2017). Thus, importance is given to the discussion of rationality below not because it is an interesting subject in its own right, but because it provides a firm foundation on which cognitive models can be compared to. As mentioned, Bayesian inference is generally accepted as the *de facto* standard for rationality and underlying the principles of Bayesian inference are the axioms of classical probability theory (CPT).

1.6.1 Classical probability theory (CPT)

Generally speaking, Bayesian statistics describes the degrees of belief an agent has over some parameter and provides the proper calculus for managing uncertainty (cf. Dutch book theorem in De Finetti, 2017; Ramsey, 1931). Frequentist statistics also provides a way of handling uncertainty which can differ both in methods and philosophy to the Bayesian

[3] However, computational models do not have to be normative. An environment could elicit computational goals that are not normative. This is especially the case when one considers the broader definition of normative outside the scope of rationality (cf. Canguilhem, 2012).

approach (Bayarri & Berger, 2004). Nonetheless, both approaches adhere to classical probability theory as usually formalised by the Kolmogorov axioms:

1) $P(E) \in \mathbb{R}, P(E) \geq 0 \qquad \forall E \in F$
2) $P(\Omega) = 1$
3) $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

These axioms define the probability *P* of some event *E*, where (Ω, *F*, *P*) is a probability space, with sample space Ω, event space *F*, and probability measure *P*. However, CPT can be seen as a special case of a more general framework called quantum probability theory (QPT) which imposes specific rules on how to assign probabilities to events and comprises a potential alternative for defining rationality (Pothos & Busemeyer, 2014; Pothos et al., 2017). QPT is an alternative mathematical formalism to CPT, independent of the physical substrate it was created to describe, that is based on complex-valued vector projections in a Hilbert space. Its status as the correct formalism for defining rationality has recently been proposed, given that human cognition is best described under the tenets of contextuality (see Pothos et al., 2017).

1.6.2 Why probability matters: sensitivity to uncertainty

Moreover, regardless of the particular axiomatisation used, humans are sensitive to uncertainty in their decision-making. Thus, numerous studies accept the Bayesian framework as the optimal way to handle uncertainty. Sensitivity to uncertainty can arise for many reasons such as handling signal-to-noise ratios that are intrinsic to neuronal firing, (Faisal, Selen & Wolpert, 2008). This is consistent with proposals that neurons can fire as if they were representing latent probability distributions (Hoyer & Hyvärinen, 2003). fMRI studies have shown that brain areas like the orbitofrontal cortex (OFC) and the striatum are activated by risk components of decision-making (Schultz et al., 2008). Sensitivity to uncertainty is also expressed through modulation of risk taking by a variety of factors including ageing (Rutledge, Smittenaar, et al., 2016) and dopamine levels (Rutledge, Skandali, Dayan, & Dolan, 2015). Furthermore, computing uncertainty makes sense from the point of view of an agent with incomplete information about the world. Bayesian models of perceptual decisions have been established (Körding & Wolpert, 2006) as well as for behaviours involving motor control (Wolpert & Landy, 2012), choices with ambiguous perceptual stimuli (Yuille & Kersten, 2006), or when integration across sensory modalities is necessary (Ernst & Banks, 2002). The success and proliferation of such models has led some to believe that the brain can be accurately

described as a Bayesian machine (Friston, 2010; Knill & Pouget, 2004). However, these models can suffer from excessive degrees of freedom if not properly constrained (Jones & Love, 2011).

1.6.3 Bayes' theorem

One thing that makes Bayesian inference – as a computational level theory – appealing for modelling human cognition is its strong use of prior information, naturally associated to our concepts of prior experience and inductive biases (Lake, Ullman, Tenenbaum, & Gershman, 2016). From a purely statistical point of view, Bayesian inference is also appealing because it uses uncertainty as a source of information about the parameters of a statistical model, which in turn results in a natural regularisation of said model (i.e., naturally avoids overfitting) (Barber, 2012). For completeness, the essence of Bayesian calculations, Bayes' rule, is presented here:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B independently. $P(A)$ is known as the prior.
- $P(A|B)$ is the conditional probability of A given B, also known as the posterior.
- $P(B|A)$ is the conditional probability of B given A, also known as the likelihood.

If referring to statistical models, A would represent the parameters of the model and B would represent the observations. The equation basically demonstrates how to update knowledge (and uncertainty) about the model parameters after making empirical observations. Bayesian inference provides a very general machinery to model human cognition and is alluring because it uses uncertainty as effectively as possible. However, there is a myriad of computational theories in the field which are not all Bayesian. One of the studies in the present work assumes that agents are integrating preferences in a way that is consistent with Bayesian rules of inference (see Chapter 3). However, as one will see, choice option representations present issues that are irrespective of the theoretical vantage point one may take with respect to what is the "correct" theory of decision-making. This point is made clear in chapter 4 where decision strategies known as heuristics are directly tested in relation to their interaction with stimulus format.

1.6.4 Beyond Bayesian statistics

The Bayesian framework naturally adapts to the constraints of everyday decision-making under conditions of uncertainty. But then again, when it comes to choice option representations, one can think of a variety of different approaches that may be "optimal" – but only with respect to the criterion that they are optimizing (e.g., minimizing a cost or error function). For example, if you think of representation as a type of compression then perhaps a useful computational theory would be that of Minimum Description Length (Chater & Vitányi, 2003; Rissanen, 1978). Similarly, statistical learning theory (Friedman, Hastie, & Tibshirani, 2001; Vapnik, 2013) and probably approximately correct (PAC) learning (Valiant, 1984) can provide bounds (i.e. worst case and best-case scenarios) on effective learning algorithms. These fields actively pursue theoretical guarantees on the optimal solution of statistical problems – some problems for which solutions may not be known. However, not all the approaches considered are Bayesian.

In a way, machine learning (Bishop, 2007; Murphy, 2012) embodies a field that is less rigorous than statistical learning theory in attempting to establish formal proofs of why a certain learning algorithm works or is optimal in some sense. This field takes a much more practical approach to constructing algorithms that simply work (in terms of predictive accuracy for example). This engineering-like attitude contrasts with the explanatory goals in cognitive science that seek to understand the learning and decision-making algorithms that "simply work" for humans. This is also where descriptive accounts of decision-making differ from normative models, such as the previously mentioned Bayesian characterisations of cognition and decision-making.

1.7 Normative and descriptive theories of decision-making

If normative theories of decision-making are equivalent to postulating explanations of decision-making at Marr's computational level, then descriptive theories of decision-making are analogous to algorithmic and representational level explanations (Bell & Raiffa, 1988; Over, 2004). Algorithmic level theories of decision-making include contributions from cognitive psychology such as the biases and heuristics program (Kahneman, Slovic, & Tversky, 1974), gestalt-like representations (Reyna, 2008), or parallel constraint satisfaction (Glöckner & Betsch, 2012), to name but a few. Such theories may fit nicely into one of Marr's levels such as the algorithmic level, but not all theories do. There are possibilities of bridging between the

computational and algorithmic levels (Griffiths et al., 2015) as well as between the algorithmic and implementational levels (Love, 2015). Modern approaches see this as a special issue of information processing systems with resource constraints (Gershman et al., 2015). Such limitations were also a prime motivation for the emergence of the *fast and frugal* heuristics program (Gigerenzer, Todd, & The ABC Group, 1999), inspired by notions of bounded rationality (Simon, 1982). Furthermore, it is also possible to start from a purely computational level and derive algorithmic approximations to the implied goal. An example of this is describing category learning as a Dirichlet process that people may solve approximately by using a one particle filter approach (Sanborn, Griffiths, & Navarro, 2006). Such a model would contrast with other accounts of category learning that seek to directly capture psychological constructs such as attentional processes (Love, Medin, & Gureckis, 2004).

1.7.1 Expected Utility Theory (EUT)

Normative theories have practical use for evaluating both human and model performance. When a person makes a choice, the normative model can define if a choice is "correct" or not, through the perspective of said model. Many studies in the decision-making literature have found that humans make systematic errors when compared to certain standards of rationality (Green & Myerson, 2004; Rabin, 2000; Tversky, Slovic, & Kahneman, 1990). These findings suggest that suboptimal decision-making and irrationality are closely related. Standards of rational choice are characterised by normative accounts of decision-making such as Von Neumann's and Morgenstern's (1953) formulation of Expected Utility Theory. In their seminal work *Theory of Games and Economic Behavior*, these authors proposed four axioms on which to base expected utility theory that provided scholars of decision theory with a consistent set of predictions with respect to choices:

1) Completeness: For every A and B either $A \succcurlyeq B$ or $B \preccurlyeq A$.
2) Transitivity: For every A, B and C with $A \succcurlyeq B$ and $B \succcurlyeq C$ then $A \succcurlyeq C$.
3) Independence: Let A, B, and C be three lotteries with $A \succcurlyeq B$, and let $t \in (0,1]$; then $tA + (1\text{-}t)C \succcurlyeq tB + (1\text{-}t)C$.
4) Continuity: Let A, B and C be lotteries with $A \succcurlyeq B \succcurlyeq C$; then there exists a probability p such that B is equally good as $pA + (1\text{-}p)C$.

The binary operators $\succcurlyeq$ and $\preccurlyeq$ represent preferences for lotteries (i.e., choice options); greater preference or lower preference, respectively. These axioms provide the necessary and sufficient conditions under which subjective value can be defined as the expectation of

individual outcomes of a gamble. This theory assumes that the agent intends to maximise her or his utility. Utility is a function of value that respects these axioms. This normative theory describes the tenets of rational choice.

1.7.2 Violations of EUT

Violations of the axioms of EUT have been witnessed in decision-making studies and are an interesting starting point for a whole agenda of research in decision-making. Some patterns of human decision-making, such as risk aversion (Rabin, 2000), can be described by an adequate utility function. Certain violations of the axioms include, but are not restricted to, loss aversion (Tversky & Kahneman, 1991), hyperbolic temporal discounting (Green & Myerson, 2004), preference reversals (Tversky et al., 1990), and many more. The interplay between descriptive and normative theories can be quite fruitful and complementary at times. Questioning assumptions of normative theories has led to the development of alternative theories such as those that incorporate reference-dependence (Koszegi & Rabin, 2006; Tversky & Kahneman, 1992).

1.7.3 Rationality wars and *fast and frugal* heuristics

The view that humans showed systematic deviations from rationality was thoroughly documented by the heuristics and biases program of Tversky and Kahneman (1975). A heuristic is defined as a rule of thumb used for decision-making (Gigerenzer & Gaissmaier, 2011). It is a decision strategy that approximates an optimal or rational solution, usually by discarding some of the information present in the decision-making context. Sometimes heuristics can work well but sometimes they provide incorrect answers, thus the negative description as a cognitive bias. The evidence that humans possess cognitive biases is pervasive. Indeed, chapter 2 presents a finding that humans are willing to forgo rewards in order to retain control over their decision-making. The interpretation of whether this finding is irrational or not depends on if one only considers expected values as opposed to expected utilities (see chapter 2 for further details). However, the dim view on human rationality was challenged by Gigerenzer and colleagues with their research agenda on *fast and frugal* heuristics (1999). Their arguments were based on two findings:

1) Many biases are corrected if the right stimulus format is used; frequencies instead of probabilities (Gigerenzer & Hoffrage, 1995).

2) Heuristics are not irrational, they are ecologically adaptive.

The first point supports the notion that the representation of choice options can be strongly influenced by stimulus format. The second point was supported by findings that heuristics can present better generalisation performance than other more complex algorithms, such as linear regression (Czerlinski, Gigerenzer, & Goldstein, 1999). This point was surprising for the psychological literature but should not have been; statistically speaking, there is a trade-off between model capacity and the amount of data fed into the model (cf. bias-variance trade-off, Briscoe & Feldman, 2011). In any case, the defence of human rationality through the ecological validity of heuristics, as opposed to their view as a cognitive bias, created an ongoing debate in the literature cleverly coined as the "rationality wars" (Sturm, 2012).

The heuristics research agenda clearly operates on Marr's second level of analysis. This can be contrasted with the Bayesian models of cognition mentioned previously (Chater & Oaksford, 1999) which tend to operate on Marr's first level (computational level). The algorithmic nature of heuristics is relevant to evaluating choice option representations since they operate on the same level. Working with such models in the present work was deemed valuable because they embody suitable candidates for decision strategies that can be controlled for in an experimental setting, as performed in chapter 4.

1.7.4 Prospect theory

In the spirit of describing the principles that guide violations of rationality is Kahneman and Tversky's (1979) prospect theory. This theory sought to incorporate heuristics and biases, such as the isolation effect or the framing effect (Tversky & Kahneman, 1986), as well as reference dependence. Key features of the model are that it can express that people hurt more for losses than they enjoy gains (Tversky & Kahneman, 1991) and that probabilities are overweighed for small values and underweighted for large ones (Gonzalez & Wu, 1999; Prelec, 1998). Due to violations of first-order stochastic dominance, the original version of prospective theory was later revised (Tversky & Kahneman, 1992). This new version of prospect theory (termed cumulative prospect theory) introduced a probability weighting function derived from rank-dependent expected utility theory (Quiggin, 1982). The equations for the prospect theory model are presented in chapter 2. Prospect theory is used as a methodological tool for estimating the parameters describing loss aversion and risk aversion which is why this theory is reviewed here. However, it is obvious that one of the most renowned models of decision-

making, such as prospect theory, also has shortcomings in specifying features of choice options that go beyond the focus on value and uncertainty.

1.8 Process and representation in decision-making

So far this chapter has surveyed an important class of models surrounding the debate on rationality such as prospect theory and *fast and frugal* heuristics (Gigerenzer et al., 1999; Kahneman & Tversky, 1979) as a descriptive and algorithmic counterpoint to the normative and computational models such as expected utility theory and Bayesian models of cognition (Mike Oaksford & Chater, 1998; Von Neumann & Morgenstern, 2007). These models dominate much of the current decision-making literature and are used as methodological tools in the experimental work presented here. One thing that is clear from this survey is a remark stated at the beginning of this chapter; these models do not include rich theoretical constructs for choice option representations and focus mostly on the features of value and uncertainty. In the next subsection, a class of models that is distanced from the rationality debate is presented known as evidence accumulation models (sometimes also known as sampling models of cognition). These models are not made use of in the experimental work presented here. However, this is a class of decision-making models that should not be left out of any decent survey. Furthermore, although this class of models provides a richer, more mechanistic view on decision-making, these models also suffer from under-specification of how choice option representations affect decision outcomes. The effect of stimulus format and features related to specific task goals and domains on choice option representations is left to *ad hoc* analysis and experimental design decisions.

1.8.1 Evidence accumulation models

In contradistinction to EUT and BMC, this class of models proposes the exact mechanisms of deliberation involved in computing a decision, namely preference or value. These theories provide the basis for modelling the stochastic dynamics of value computations, which in turn implicitly model uncertainty in that value. Models in this spirit are based on the assumption that the process of making a decision involves accumulation of evidence in favour of each individual option such as Decision Field Theory (Busemeyer & Townsend, 1993), the Drift Diffusion model (Ratcliff, 1978), the Linear Ballistic Accumulator (Brown and Heathcote, 2008), or the Attentional Drift Diffusion model (Krajbich et al., 2010). The models are similar in some respects but differ crucially in others. For example, the Attentional Drift

Diffusion model establishes attention weights to the options via eye gaze fixations whereas the other models assume equal attention to all options. The models may assume an online account of valuation, meaning that the value estimate is constructed in real-time for each choice. With this line of reasoning, some authors propose that value is actually constructed through the process of its elicitation (Lichtenstein & Slovic, 2006) or that it is constructed as a product of choice itself (Riefer, Prior, Blair, Pavey, & Love, 2017; Sharot, Velasquez, & Dolan, 2010).

Evidence accumulation models can be expected to operate either sequentially or in parallel. The evidence for one option over another is accumulated until a certain decision threshold is reached. The threshold is important since it can have an effect on speed-accuracy trade-offs. The threshold is assumed to be constant but more recent approaches posit that the threshold may collapse over time (Ratcliff et al., 2016). Collapsing boundaries are adequate when timing constraints are relevant (chapter 4 touches upon the effect of timing constraints on the cognitive implementation of decision strategies). The models may be specified in terms of different accumulators (one per choice option) or in terms of a single accumulator. For example, the race model (Vickers, 1970) assumes only one accumulator. It was originally used for perceptual decisions but later used for the analysis of confidence and metacognition in a value-based decision-making study (De Martino, Fleming, Garrett, & Dolan, 2013). In general, evidence accumulation models provide a good descriptive account of decision-making. They not only consider the behavioural responses but also the form of the reaction time distributions as in the drift diffusion model (Ratcliff & McKoon, 2008).

As mentioned previously, these models provide good descriptions of how people compute value and uncertainty when making decisions. However, there is a gap in the literature that needs addressing; a gap that is identified as the influence of choice option representations. Studies seem to have focused more on process than on representation, even though the two are inseparable. To study the influence of choice option representations on decision-making, first it is necessary to study how such representations are modified by the decision-making context; by task goals and informational input. Decomposing these algorithmic models into different cognitive processes, which include representational issues, are a necessary step for further theoretical advancements.

1.8.2 A general process model of decision-making

When algorithmic models of decision-making are decomposed into constituent psychological constructs, these models are sometimes referred to as process models. Evidence accumulation models would generally be classified under this umbrella term. On occasion, these models try to link with the neural implementational level. Such links with the neural level are currently being studied through model-based cognitive neuroscience (Palmeri, Love, & Turner, 2017), with plentiful success relating theoretical models with functional neuroimaging (Pratte & Tong, 2017). Broadly speaking, these decision-making models can involve one or more of the steps shown in Figure 1.1: representation, valuation, action selection, outcome evaluation and learning.

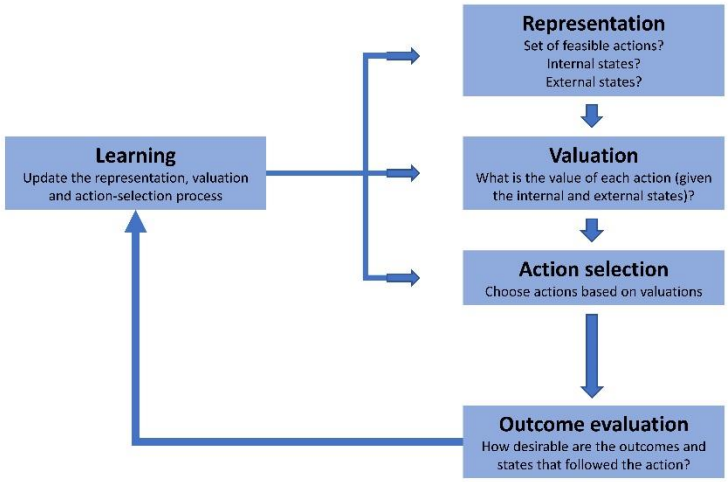

Figure 1.1 Key computational steps in a decision-making process (adjusted from Rangel, Camerer & Montague, 2008)

The figure shows the key components that should be considered when studying decision-making processes. The feedback diagram alternates between outcome evaluation and learning. In this framework, learning is decomposed into representation, valuation, and action selection.

Each step in this framework (Figure 1.1) is important in achieving adaptive actions. To represent a task goal accurately, information needs to be integrated between the sensory modalities in accordance with the internal state of the decision maker. It is hypothesised that introspective abilities are necessary for error-monitoring of value estimates and choice

outcomes (Yeung & Summerfield, 2012) in order to correct the internal representations of task-relevant information. Noise in the environment and in the nervous system (Glimcher, 2008) make it necessary for such a corrective mechanism to exist. Noise inherent to neuronal computations as well as uncertainty representing incomplete information about the environment is managed through action and outcome feedback loops, and improved through learning. Prediction errors emerge as a by-product of the difference between expected outcome and actual outcome. This difference, which is the magnitude of the prediction error, is gradually diminished through learning. Prediction errors can occur for different quantities such as expected rewards (Bayer & Glimcher, 2005; Rutledge, Dean, Caplin, & Glimcher, 2010; Rutledge, Skandali, Dayan, & Dolan, 2014), reward information (Bromberg-Martin & Hikosaka, 2011), or information state (Gläscher, Daw, Dayan, & O'Doherty, 2010).

The framework for decision-making processes in Figure 1.1 provides a fair blueprint that can guide research. It sensibly states the importance of representations, generally speaking. However, for purposes of the present work this scope is too wide since it involves representations regarding permissible actions and representations about the agent's own internal state such as satiation levels, heart rate, or emotional condition. The studies presented here focus on the narrower issue of internalizing information from the environment, specifically regarding choice options relevant to the decision. Although in chapter 2, the study does address representations of the self which become relevant given that one of the choice options is actually the agent (i.e., the agent has to choose between herself and another agent).

1.8.3 From process to representation: similarity models

As mentioned previously, process and representation give two slightly different views on the same information processing phenomenon. For purposes of this work, process focuses on a more dynamic, transformational view on information while representation will focus more on a static, morphological, view on information. Representational issues may address questions regarding the shape of the information or the structures it is composed of. Complementarily, the process will address the sequential transformations between such representations, as the models that have been presented earlier, attempt to explain. Cognitive models are underspecified if only one of these facets (representation or process) is taken into consideration (cf. the mimicry theorem in Anderson, 1978). Under this view, models of similarity in

psychology are the class of models that best characterise such representational issues, whether they be in relation to decision-making processes or not.

Computing similarity is a fundamental requirement for many cognitive processes such as learning, memory encoding and retrieval, discrimination, recognition, etc. Similarity is necessary for any type of abstraction. Thus, computing similarity is not only of interest to cognitive science but to other fields like machine learning (Chen, Garcia, Gupta, Rahimi, & Cazzanti, 2009; Hancock & Pelillo, 2011). Many cognitive models, like models of vision or categorisation, need to be grounded in some notion of similarity (Edelman, 1998; Goldstone, 1994). Proposals that distances (inversely related to similarities) can be represented as points in a multidimensional space are common (Coombs, 1958; Shepard, 1962). The metric that is used to compute distances can have an impact on the separability of stimuli. Previous research in human similarity judgments has shown that there are advantages of using a city-block metric as opposed to a Euclidean metric for certain types of stimuli (Shepard, 1987). Tasks such as spatial navigation implicitly rely on a Euclidean measure (Giocomo, Moser, & Moser, 2011). In other words, the choice of measure matters.

However, modelling similarity judgments within a metric space (as distances) imposes certain restrictions on the types of permissible judgments. Specifically, metrics or distance functions d on a set A are defined as a mapping of two variables to the set of positive real numbers, d: $A \times A \rightarrow [0,\infty)$, for all x, y, and $z \in A$ that respect the following axioms:

1) $d(x,y) \geq 0$ (non-negativity)
2) $d(x,y) = 0 \leftrightarrow x = y$ (identity of indiscernibles)
3) $d(x,y) = d(y,x)$ (symmetry)
4) $d(x,z) \leq d(x,y) + d(x,z)$ (triangle inequality)

Although these assumptions provide certain mathematical conveniences (Jäkel, Schölkopf, & Wichmann, 2008), they are sometimes relaxed by other approaches to better account for the empirical data observed for human similarity judgments. Other accounts include featural approaches (Tversky, 1977; Tversky & Gati, 1982; Tversky & Krantz, 1970), probabilistic approaches (Ennis, Palen, & Mullen, 1988; Tenenbaum & Griffiths, 2001), structural approaches (Gentner & Markman, 1997), quantum probability approaches (Pothos et al., 2015; Pothos, Busemeyer, & Trueblood, 2013; Pothos & Trueblood, 2015; Trueblood, Pothos, & Busemeyer, 2014) or transformational approaches (Hahn, Chater, & Richardson, 2003).

1.9 Studying the implementational level in decision-making

The previous sections have dealt with Marr's first two levels of abstraction for information processing systems: the computational level and the algorithmic/representational level. Models from each level have been selected as experimental tools for studying the representation of choice options in the following chapters. For the implementational level, models are not so much relevant to the experimental work presented here so much as the techniques and methods used to gather and interpret neural data. Admittedly, there has been much work done at this level in pinpointing the exact biological mechanisms that carry out brain computations, as in studies of spike-timing dependent plasticity (STDP) (Sjöström & Gerstner, 2010), long-term depression (LTD) (Ito, 1989), short-term depression (STD) (Zucker, 1989), or N-methyl-D-aspartate (NMDA) receptor channels (Lisman, Fellous, & Wang, 1998). At the neuronal population level, an important class of models is known as neural networks (i.e., connectionist) which can be classified either as artificial neural networks (Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Rumelhart, McClelland, & PDP Research Group, 1987) or biological neural networks (Dayan & Abbott, 2001). Although, one could argue that artificial neural networks are better suited for the algorithmic level, progress has been made in relating these models to brain biology (Scellier & Bengio, 2017). Although authors may classify these models at the algorithmic level, these types of discrepancies display that sometimes the same class of models can be reinterpreted at different levels. For example, the interactive word recognition model (McClelland & Rumelhart, 1981) – a connectionist model with interacting layers of representations – can be modified to perform Bayesian inference (Norris, 2013). Examples of other models that lie at the intersection between the algorithmic and computational levels are Bayesian explanations of inductive confirmation (Tentori, Crupi, & Russo, 2013) or probabilities with noise (Costello & Watts, 2014). Neural networks can illuminate the trade-off hypothesis between stimuli and goals given that they both seek to cluster similar stimuli together and to optimise a cost function at the same time (see Love, 2005 for a similar view). The scope of this area is vast but not relevant to this work. What is relevant here however are two methods that are closely related, one technological and the other analytical; functional magnetic resonance imaging and representational similarity analysis, respectively.

1.9.1 Brief overview of functional magnetic resonance imaging (fMRI)

When presented with a stimulus, a person's nervous system will transform its physical presentation into a neuronal representation. What is the nature of a neural representation of a stimulus? Technological and methodological advances in cognitive neuroscience give the tools that are necessary in addressing this problem. Insight is gained into the neural representations of stimuli through fMRI. This technology can measure blood oxygen level dependent (BOLD) signals that vary with task demands and presentations of stimuli (Huettel, Song, & McCarthy, 2004). Such a signal is believed to correlate with population levels of neuronal activity (Goense, Merkle, & Logothetis, 2012). Although its temporal resolution is confined to be within just a few seconds, its spatial resolution operates on a millimetric scale with voxels being the standard representational unit for the images; well suited for localizing different brain functions and computations.[4] This technique is used for both the top-down approach and the bottom-up approach in studying choice option representations (chapters 3 and 5, respectively). It is a technique that is well suited for many purposes such as: 1) validating cognitive models of decision-making and choice option representations, 2) localizing neuronal populations that correlate with cognitive operations in such models.

This technique gathers data in the form of voxels (mentioned above). A typical brain volume can be collected in matter of a few seconds and will be composed of approximately fifty thousand of these voluminous pixels for a typical experiment. The BOLD signal, which is intended to be recovered for each voxel from a series of brain volumes, varies on a time scale that is quite slow (around eight seconds to peak and fifteen seconds back to baseline). The typical analysis pipeline will seek to correct many artefacts in the data such as temporal drift, physiological noise like heart rate, and motions from the participant lying down in the fMRI scanner (Huettel, Song, & McCarthy, 2004). The canonical way of statistically analysing this data is through massive univariate general linear models (one for each voxel through time) (Monti, 2011). Other approaches include more global brain properties like independent component analysis (Calhoun, Liu, & Adali, 2009), principal component analysis (Smith, Hyvärinen, Varoquaux, Miller, & Beckmann, 2014), or multivariate pattern analyses (Norman, Polyn, Detre, & Haxby, 2006). For the work presented here, chapter 3 will make use of the

[4] Although, the fact that fMRI can detect signals on such a coarse spatiotemporal scale reveals interesting properties of the neural code (Guest & Love, 2017).

canonical tradition of massive univariate analysis and chapter 5 will use both the univariate approach as well as the multivariate approach. A special case of the multivariate approach is known as representational similarity analysis, which is closely linked to the analysis of the neural similarity measures between choice options in chapter 5. It is important to note that inferences drawn from fMRI data are only indirectly related to neural activity through the BOLD signal. Deviations from baseline BOLD signal can be seen as averaging activity over tens of thousands of neurons. Thus, inferences in this space can only make claims at the level of summary statistics for neuronal populations. Inferring latent spaces from such summary statistics (i.e., dips and peaks with respect to baseline BOLD signal) will thus be constrained by the nature of the data, as will be displayed in chapter 5.

1.9.2 Representational similarity analysis (RSA)

As mentioned, different analysis pipelines have been proposed for analysing fMRI data. Multivariate pattern analyses (MVPA) such as representational similarity analysis (RSA) or machine learning classifiers such as linear support vector machines (SVM) have found great success for this purpose (Joern Diedrichsen & Kriegeskorte, 2016; Haxby, Connolly, & Guntupalli, 2014; Kriegeskorte, Mur, & Bandettini, 2008) (see Chapter 5 for an example of how these techniques may be used for inference on fMRI data). Indeed, RSA is a great starting point for understanding the nature of neural representations. This method is based on computing representational similarity matrices (RDM). These matrices are symmetrical and are constructed from all pairwise distances between stimuli for a given experiment. Each cell in one of these matrices represents a similarity between two different brain states (i.e. voxel activations). In the case of fMRI, a cell in this matrix could represent the distance between voxel activations for seeing the image of a hammer and voxel activations for seeing the image of a basketball. These brain states can represent the way the brain has reacted to a stimulus presentation or any other experimental event of interest. Computing the similarity between two different brain states can grant direct insight into the nature of endogenous choice option representations.

1.10 Research question

The present research investigates how people represent choice options in the presence of task demands and constraints imposed by stimulus format. As displayed by this survey on decision-making models, the representational issue has been overshadowed by focus on the

algorithmic (i.e., process) aspects of decision-making. With regards to value and uncertainty, comparably less attention has been put on other features of choice options which can influence decision computations. Of course, there are infinitely many features one could wish to study; thus, features were strategically picked to expand upon. For the top-down approach, the studies focus on features relevant to the social domain, especially pertaining to the sense of self, control, and personal preferences. For the bottom-up approach, the studies focused on features that are stimulus specific such as order of presentation, numericity, colour-coding and similarity measures between stimuli. This subtle change in perspective, from processual to representational, can provide new and interesting insights for research in decision-making.

1.11 Dissertation outline

How do choice option representations handle the trade-off in requirements imposed by task goals and stimulus properties? At this point a general outline of the dissertation can be presented that aims to answer this question. The general strategy of the work is divided into a top-down approach (chapters 2 and 3), working with computational models such as EUT and BMC, and a bottom-up approach (chapters 4 and 5), working with algorithmic models such as *fast and frugal* heuristics and the analysis of similarity measures through RSA. Prospect theory (a descriptive model) is also employed in chapter 2 to obtain certain parameters of interest for the study (loss aversion and risk aversion parameters). Each approach starts with presenting a behavioural study (chapters 2 and 4) and then continues with an fMRI study (chapters 3 and 5). Adding an implementational perspective to each approach through fMRI provides added value and extra analytical degrees of freedom in addressing the trade-off hypothesis for choice option representations.

1.11.1 The top-down approach

To test the hypothesis that the representation of choice options is dependent on both informational input and task goals, the thesis starts by presenting a set of studies that tests how people integrate beliefs and preferences in the social domain. The social domain was chosen as a testbed because the integration of conflicting information from different sources (self versus other) naturally arises in everyday decision-making. These studies are presented in chapters 2 and 3. They rely on the assumption that the participants in the studies are trying to maximise their expected utility with respect to potential rewards in the task. Importantly, the tasks differ with respect to the choice options that are offered and they also differ in the

assumptions regarding the computation of uncertainty. The study in chapter 3 explicitly strives to model the integration of preferences within a Bayesian update framework, which implies that people represent preferences as distributions governed by parameters with their own uncertainty. In contrast, the study in chapter 2 only assumes point estimates of certain quantities are necessary for successful task completion.

1.11.1.1 The social domain

The study in chapter 2 starts with addressing a question that is pertinent to all decision-making; does the agent want to make a choice or would the agent prefer someone else to make a choice for them? This research question queries the role of agency in opposition to the faculty of delegating a choice. The choice options are actually the (potential) decision makers themselves – the delegator and the potential delegate. The study additionally tests whether people will be willing to incur a cost as a consequence of choosing themselves.

The study in chapter 3 is an fMRI study that directly investigates the integration of new information within the context of a value-based judgment. Specifically, the study tests how social information gets integrated into value judgments while considering the reliability of that information. The social information is presented as other people's preferences for a set of retail products. Participants are tasked to integrate this information with their own preferences. Thus, the integration of preferences of the participants are analysed in relation to the retail products, informing on the representation of these choice options. Furthermore, the study seeks to understand if such an integration of the information's reliability is consistent with Bayesian integration of uncertainty. The neural representation of this uncertainty is also discussed.

1.11.2 The bottom-up approach

The studies in chapters 2 and 3 make assumptions on the computational goals that participants will seek to achieve while letting participants freely choose the decision strategy to accomplish this. The assumed goals are that participants are trying to maximise expected utility and that they use uncertainty in this estimate to inform that maximisation process. The exact algorithm (i.e., decision strategy) that participants use is not known. This uncontrolled aspect restricts the extent of permissible conclusions regarding choice option representations, given the intricate relation between representation and algorithm. This caveat naturally leads the research agenda to setting up a study that explicitly controls for the decision strategy being

used in the experiment (chapter 4). This greater experimental control enables testing the interaction of decision strategies with other contextual variables of interest that inform on the nature of choice option representations such as stimuli format (colour-coding, numericity, and serial order) and timing constraints. The approach in chapter 5 is deemed bottom-up because of its focus on comparing similarity measures between stimuli sets (from two separate, previously published, datasets) and between individual stimuli.

1.11.2.1 Controlling for decision strategy

Chapter 4 presents a study that investigates the effect that stimuli formats have on a pair of decision strategies; specifically, two decision heuristics known as Tallying and Take-the-Best (see chapter 4 for their definitions). These simple heuristics are easily explained to participants; hence they are explicitly instructed to use them during the task. Thus, compliance with each heuristic is the dependent variable of interest. The results in this chapter show that different algorithms require different cognitive processes such that a stimulus format that is good for one algorithm is not adequate for the other.

1.11.2.2 Fundamental representational issues in the human brain

The study in chapter 5 seeks to deconstruct the very nature of the neural representations of choice options down to their core essence; that of similarity relations. It opens a new line of inquiry by questioning the appropriateness of various similarity measures (not necessarily distance metrics) for analysing fMRI data. Formal competition amongst measures is required to evaluate the brain's representational capacities. The findings here show that similarity is computed consistently across brain areas but is modulated by stimulus format and/or task.

Chapter 2 To delegate or not to delegate? Representation of the self as a choice option

2.1 Delegation study

This chapter presents a study that utilises a top-down approach for investigation of choice option representations. The social domain is picked as a convenient testbed provided that this domain is ubiquitous in our everyday decision-making. One pervasive choice that people need to make is whether to do a task themselves – and the deliberation that comes with it – or to delegate a task to someone else. We delegate choices all the time. The fact that cooperative societies can function at all is based on delegation of duties, providing access to pools of knowledge (cf. knowledge illusion in Sloman & Fernbach, 2017) and material resources that would be impossible to achieve at the individual level. In a sense, a delegation can be viewed as a second-order choice. Nonetheless, it is still a choice with two choice options: the decision-maker and the person to whom the decision would be (or could be) delegated to. Thus, for this decision, high-level and abstract features regarding the decision-maker (i.e., the self) and the other decision-maker play a key role in the representation of choice options. To be clear, delegation involves construction of the self and the potential delegate as a choice option. Representational features of these choice options such as preferences for control and overconfidence in one's own decision-making could influence the decision of whether to delegate a task or not.

As mentioned, people face a pervasive choice (e.g., in business, government, and daily life): whether to make a choice on their own or to delegate choice-making authority to someone else. With respect to investments, for example, one could rely on one's own judgment, or one could rely instead on a trusted agent. Employees might choose health care plans on their own or might ask employers to make the relevant choices. Patients and clients have the same dilemma in dealing with doctors and lawyers. Any principal can rely on or appoint an agent, who might have superior knowledge, might be immune from various biases, and might relieve the principal of the obligation to devote scarce time, and limited cognitive resources, to making difficult choices. On the other hand, an agent might have inferior knowledge, be ignorant of the principal's real concerns, have her own biases, or be influenced by her own self-interest.

Clearly, to make a truly informed decision regarding whether to delegate or not to delegate, the agent must have a clear representation of what each actor can offer. At the

minimum, the agent should have an intuition about their own performance compared to that of the other agent who could make a decision for them. For example, suppose you are going to be put on trial for a crime that you did not commit. Would you represent yourself or would you prefer that the state appointed attorney represent you instead? If you are one of the greatest criminal lawyers in the city then perhaps you are better off representing yourself. Otherwise, most people would choose to be represented by the state appointed attorney or to hire a private one instead. This leads to decomposing the representation of each agent according to key elements that can inform on making the best decision of whether to delegate or not and to whom.

In theory, the decision whether to choose, or instead to delegate, should be a fully rational one, based on some form of cost-benefit analysis (Friedman 1953; as expected by EUT introduced previously, Von Neumann and Morgenstern 2007). Choosers might begin by thinking in terms of expected value: would the payoff be higher with or without a delegation? They might also ask about the value of saving limited time and attention (McFadden 2001; Simon 1978). If the savings would be substantial, choosers might be willing to sacrifice something substantial in terms of expected value. It also matters whether choosing itself has benefits or costs, in the sense that choosers enjoy, or instead dislike, the time that they devote to choosing. For some people, it may be interesting or fun to think about the best investments or the right health care plan. For other people, those choices are unpleasant and tiring, a kind of hedonic tax, and it is a great relief if someone else can make the choice for them. In this chapter, it is assumed that people are trying to maximise expected utility for delegations (i.e., this is the task goal). Furthermore, the biases that shape this utility function with respect to features of the choice option representations are investigated. The nature of these biases in choice option representations is tightly linked with the nature of the task goal.

With regards to everyday decision-making, choosers might consider whether the pleasure of a reward, and the pain of a loss, are amplified or reduced if they are personally responsible for the outcomes. Studies have shown that people value items they had selected themselves more than identical items that were selected for them (Brehm 1956; Egan et al. 2007, 2010; Lieberman et al. 2001; Sharot et al. 2009, 2010b). Neurologically, outcomes that were obtained by making an active choice are associated with greater activity in the striatum, a region that processes reward (Rao et al. 2008; Samejima et al. 2005; Sharot et al. 2009; Studer et al 2012), and with heightened dopamine release (Syed et al. 2015), which is a

neurotransmitter crucial for learning about the value of stimuli (Schultz et al. 1997). Thus, one could imagine a situation in which choosers would prefer (1) gaining $100 if that gain came from their own efforts to (2) gaining $110 if that gain came from someone else's efforts, as the subjective value of self-attained $100 may be greater than that of $110 that was attained via an agent.

Consistent with this speculation, it has been found that people prefer options that permit further choice over those that do not (Bown et al. 2003; Catania 1981; Suzuki 1997). Similarly, people are willing to pay to control their own payoffs, rather than delegate, when faced with potential rewards (Owens et al. 2014). Would people also choose to pay to retain control in the face of potential loss? It is unknown whether choice has positive utility in situations where a person needs to decide between two aversive outcomes, such as when deciding between medical treatments or when deciding whether a stock should be sold or held to minimise loss. On the one hand, people may want to delegate choices that involve a potential loss as to avoid a feeling of regret for selecting the wrong option, and thus they may prefer to accept the status quo rather than make errors of commission (Samuelson and Zeckhauser 1988). On the other hand, a sense of control has been shown to reduce stress and anxiety in the face of unwanted outcomes (Alloy and Clements 1992; Shapiro et al. 1996; Thompson 1999). For that reason, making a choice may reduce the aversive utility of a loss (Sharot et al. 2010a) leading people to prefer agency over delegation. The way people feel about losses, gains, and control, has a direct influence on their preferences and on the utility function that they seek to optimise. Maximizing that utility function, in the context of optimal delegation, is in effect the task goal. Such a maximisation can vary depending on individual preferences regarding losses, gains, control, and the way they represent uncertainty.

Here, this study tests if in the face of potential losses and gains, people will pay, or demand payment, to be choosers. On each trial, participants performed a simple choice between two shapes in order to maximise reward and minimise loss. On "gain trials," a correct choice would result in a monetary gain and an incorrect choice in no gain. On "loss trials," a correct choice would result in no loss and an incorrect choice in a monetary loss. After performing the task for an extended period of time on their own, participants were given an opportunity to delegate the decision-making to an advisor. The expected value of the advisor was disclosed on each trial and participants' perception of their own expected value was also elicited. This

allowed the examination of whether participants made “rational” delegation choices given their beliefs when faced with potential gains and with potential losses.

This chapter seeks to answer whether the task goal (i.e., maximizing expected utility) is affected by high-level features of choice option representations such as control preferences or overconfidence. The top-down approach used here facilitates the study of these high-level features. Furthermore, the study of whether these features hold for different framings of the task (either as losses or gains) is also of interest to how the task itself is represented (see Tversky & Kahneman, 1986).

2.2 Methods

2.2.1 Participants

Fifty-four participants (aged 18 – 61 years; mean age, 25.0 years; 33 females) were recruited via a University College London website. Participants gave informed consent and were compensated for their time. Four participants were excluded from the analyses because they failed to complete all parts of the study. The study was approved by the University College London Research Ethics Committee.

2.2.2 Stimuli

Stimuli included three hundred and sixty unique geometrical shapes varying in colour and orientation. Each shape appeared at most once throughout the study.

2.2.3 Procedure

The study included two parts. Part I was a learning task and Part II was a delegation task. Each part consisted of one gain block (60 trials) and one loss block (60 trials). Order of gain and loss blocks were counterbalanced across participants.

2.2.4 Part I: Learning task

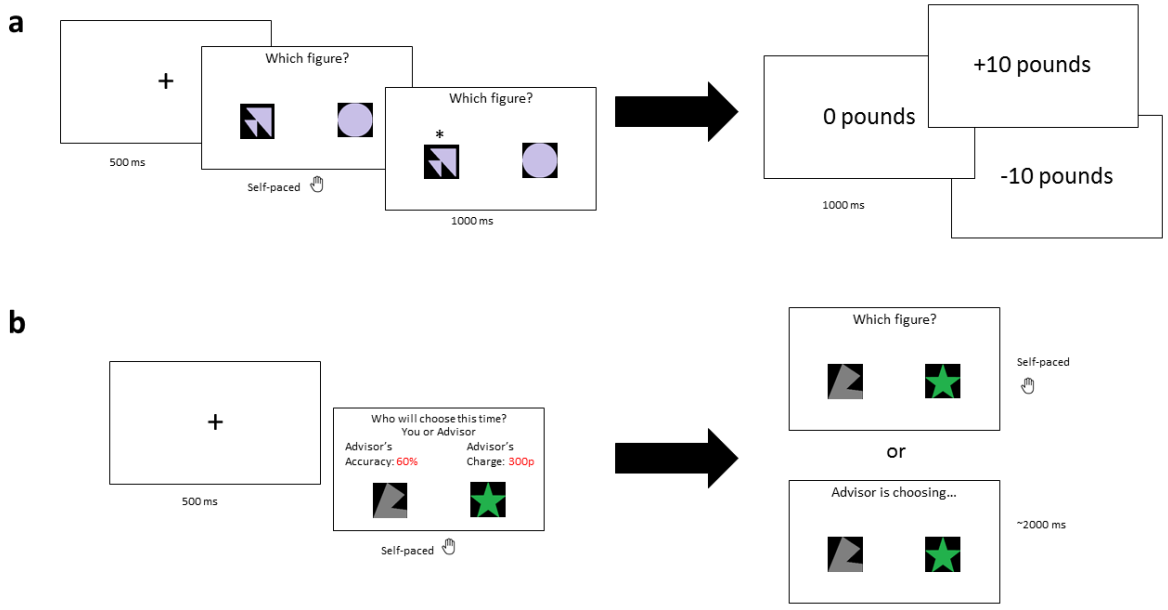

Figure 2.1. Task for the delegation study.

(a) In the first part of the experiment – the Learning Task - participants were asked to choose on each trial between two shapes. If they successfully selected the shape associated with the better outcome they received £10 in the gain block and £0 in the loss block. If they were unsuccessful and selected the shape associated with the worse outcome they received £0 in the gain block and lost £10 in the loss block. (b) In the second part of the experiment – the Delegation Task – participants first choose whether they would like to retain agency and make the choice themselves or delegate the choice to an "artificial advisor" (a computer algorithm). They were presented with the advisor's success rate, charge, and the pair of figures was shown at the time participants made the delegation choice. Outcomes were not revealed.

The goal of the first part of the study was to familiarise the participants with a simple decision-making task. On each trial, two shapes were presented and the participants' task was to choose the shape that would deliver a better outcome. Participants were told that their task was to discover the rules, if any, that made some shapes "better" than others. No underlying rules governed winning shapes in the task. Rather, outcomes were random, such that each participant received the desirable outcome on 50% of the trials in each block. This was done for two reasons: 1) so that all participants would be at the same level of expertise, and 2) to be able to provide a wide range of both superior and inferior advisors. Had learning been available, this level of expertise could have been correlated with delegation decisions but would have added unnecessary complexity to this initial experimental design.

The pairs of shapes were drawn at random from the stimuli set without repetition and remained on screen until the participant made their decision. Once the choice was made, an asterisk (*) was shown above the chosen shape for one second and then a new screen with the outcome was presented for one second (Figure 2.1). In the gain block the desired outcome was £10 and the undesired outcome was £0. In the loss block the desired outcome was £0 and the undesired outcome was -£10. Each shape appeared only once throughout the experiment.

2.2.5 Part II: Delegation task

The goal of Part II was to quantify participants' willingness to delegate decisions to an advisor in the same task experienced in Part I. On each trial, participants had to decide first whether to select between two novel shapes themselves or delegate the choice to an advisor. Participants were accurately informed that advisors were in fact computer algorithms (i.e. artificial agents), whose advice on each trial was determined prior to the start of the experiment. On each trial, there would be a different "artificial advisor." On each trial, the pair of shapes were presented before participants had to decide whether to delegate the decision so they had full information about the decision at hand.

On each trial, participants were presented with two pieces of information regarding the advisor: the advisor's mean success rate in the task (from 0% to 100%, with mean 71.83 %, s.d. = 23.9) and the advisor's fee in British Pounds (from £0 to £10, with mean £3.44, SD = 3.55). Participants could decide either to delegate the decision between the shapes or to make the decision themselves. If participants decided to choose themselves, participants were given unlimited time to make their decision. If participants decided to delegate, a new screen appeared for approximately 2 seconds with the pair of novel shapes and text saying, "Advisor is choosing…" Participants were clearly instructed that the advisors were in fact algorithms and that the "decisions" of those algorithms were taken from a pool of past decisions. In either case, participants would wait an additional one second before going on to the next trial. Outcomes (i.e. feedback) were not revealed, and thus participants were unable to learn about their own ability or to update their beliefs about the ability of the advisors. At the end of the study ten trials would be chosen at random from Part II and the average outcome for those trials would be added to the baseline compensation of £7.

2.2.6 Advisors

The participants paid the fee to the advisor only if they asked the advisor to choose for them and the advisor selected the correct shape. The expected value obtained by the participant by choosing to delegate ranged from £0 to £10 in the gain block with a mean of £5, and £0 to -£10 in the loss block with a mean of -£5. Because the participants' objective accuracy rate was 50%, their expected value was £5 (=50% * £10) on a gain block and -£5 on a loss block. Thus, a value maximiser should choose to delegate whenever the advisor would return a mean expected value higher than £5 on a gain block but not otherwise (and be indifferent when the advisor's expected value was exactly £5). On a loss block, a value maximiser should choose to delegate when the advisor would return a mean expected value that was higher than -£5 but not otherwise (and be indifferent when the advisor would return an expected value that was exactly -£5).

For example, on a gain block an advisor with 90% accuracy who charged £5 would return an expected value of ((accuracy*£10) – charge)) = ((90%*£10)-£5) = £4. On this trial, a value maximiser should not delegate. On half the trials, participants were offered advisors that would return an expected value greater than £5 (or greater than -£5 on loss trials) and the other half lower than £5 (or lower than -£5 on loss trials).

2.2.7 Self-perceived accuracy (SPA)

At the end of the study participants were asked to provide an estimate of how accurate they believed they were at choosing the correct shapes (from 1% to 99%). This estimate is referred to as self-perceived accuracy (SPA).

2.2.8 Additional questions

To test whether the loss block was perceived negatively and the gain block positively (i.e. losses and gains are perceived as such), participants were asked at the end of the study which block (loss or gain) made them happier; which they would select to repeat; and how much they would pay to repeat each block.

2.2.9 Gambling task

Participants completed an additional gambling task at the end of the experiment, which would help account for risk attitudes when estimating the control premium in the gain and loss domain (task adapted from Charpentier, De Neve, Li, Roiser, & Sharot, 2016). This is helpful because the probability of the agent selecting the correct stimuli in relation to the participant doing so is different and alters on each trial. Thus, risk preferences may influence delegation choices.

On each trial of the gambling task participants selected between a sure option and a 50-50 gamble with varying values of gains and losses. On half the trials, the sure option was £0, and the gamble option included gains ranging from £6 to £24, and losses ranging from £1 to £12. The difference in loss and gain ranges facilitates the estimation of loss aversion. Although, scaling effects have been seen to impact risk perception (Walasek & Stewart, 2015), this is not an issue for this study since the estimates will be used in a regression analysis. On half the trials, the sure option yielded a gain (range £1 to £12), and the gamble option included another gain (range £6 to £34) and £0. These trials allowed estimation of risk aversion (see Charpentier et al., 2016).

The task was divided into two blocks. This first block involved a staircase procedure that enabled estimation of the participant's equivalence point between a gain-only gamble and £0. These gamble-related equivalence points were then used to construct trials in the second block for improved parameter estimation (see Charpentier et al., 2016). The staircase procedure involved updating the difference between the value of the sure option and the value of the gamble option every 2 trials, increasing by £1 if the participant chose the gamble option, and decreasing by £1 if the participant chose the sure option. The equivalence points were then used to generate 120 trials in the second block centred upon the equivalence point. At the end of the study participants' reward for the whole experiment was revealed.

Three parameters were then estimated from the second block of trials in accordance with Prospect Theory equations (Kahneman & Tversky, 1979); the loss aversion parameter λ , which is the ratio of sensitivity to losses to sensitivity of gains; the logit sensitivity μ, which is the consistency of participants' choices across the task; and the risk aversion parameter ρ, which represented the diminishing sensitivity to changes in value as the absolute value

increases. This is usually <1 for risk-averse individuals, and >1 for risk-seeking individuals. It is also the curvature of the utility function, assumed to be the same in gain and loss domains. These parameters were used to calculate the probability of accepting a gamble as per the following softmax function:

$$\mathrm{P(gamble\ accepted)} = \frac{1}{1+ e^{-\mu(u(\mathrm{gamble})-u(\mathrm{sure\ option}))}} \quad (2.1)$$

where u(x) is the subjective utility of the respective options, estimated by:

$$u(x) = \begin{cases} x^{\rho} & x>0 \\ -\lambda(-x)^{\rho} & x<0 \end{cases} \quad (2.2)$$

Both the loss aversion parameter λ and the risk aversion parameter ρ would subsequently be used in the analysis of delegation indifference points (see below for more details on this analysis). These parameters would be included in the model as covariates.

2.2.10 Analysis of delegation rates

First, the percentage of trials in which participants chose to delegate was calculated. Then, the percentage of trials in which participants decided to delegate/retain agency out all of all trials in which delegation was optimal (i.e. trials when the advisor's expected value was above £5) and the percentage of trials in which they decided to delegate/retain agency when delegation was not optimal (i.e. trials when the advisor's expected value was below £5) were calculated. For loss trials, this would be above and below -£5.

The same analysis was conducted considering the participants' self-perceived accuracy (SPA). Namely, the percentage of trials in which participants chose to delegate/retain agency when this was optimal given a participants' SPA (i.e. the advisor's expected value was above ((SPA*£10)-charge), and when it was suboptimal (i.e. below ((SPA*£10)-charge) were calculated. For loss trials, this would be above and below ((SPA*-£10)-charge).

2.2.11 Analysis of indifference points

The indifference point of each participant was calculated as the expected value at which each participant would delegate with 50% probability. To this end, a mixed effects model for

each condition (Gains and Losses) was performed separately with the advisor's expected value as an independent variable, allowing for random effects and a random slope per participant. In other words, all participants had their own parameter estimates for the advisor's expected value (modelled as a random slope) and for an intercept (also modelled as a random effect) drawn from a common Gaussian distribution for all participants. This type of model is appropriate for repeated measures because the regression coefficients are given their own probability model (Gelman & Hill, 2007). They are also calculated more efficiently given that they are estimated hierarchically, both within and across participants (Bagiella, Sloan, & Heitjan, 2000; Clark & Linzer, 2015; Gelman & Hill, 2007).

The model predictions for each participant were used to estimate their indifference points respectively, controlling for trial number (to account for linear temporal effects such as fatigue or disengagement over time), loss aversion (λ) and risk aversion (ρ) in the model. Random effects were also included for trial number but not for the loss aversion and risk aversion parameters since these parameters were constant across observations within each participant. The latter two were estimated in a separate gambling task (see above). Median estimates of λ and ρ were 1.51 and 0.65, respectively.

Table 2.1 Parameter estimates for the mixed-effects model

Gains Model				
	β	s.e.	*z*	*p*
Fixed Effects				
(Intercept)	-8.44	1.26	-6.70	<.001
Expected Value	11.02	0.98	11.29	<.001
Trial Number	-.009	0.005	-1.86	0.063
Rho	0.04	0.08	0.54	0.589
Lambda	-0.02	1.68	-0.01	0.991
Losses Model				
	β	s.e.	*z*	*p*
Fixed Effects				
(Intercept)	-6.42	1.05	-6.12	<.001
Expected Value	9.77	0.83	11.83	<.001
Trial Number	-0.02	0.005	-3.06	0.002
Rho	0.09	0.07	1.27	0.2
Lambda	-1.27	1.46	-0.87	0.38

Coefficients (β), standard errors (s.e.), *z* statistics and *p* values for the mixed effects model in the Gains condition (top) and the Losses condition (bottom). The results reported here are the fixed effects.

Model comparisons were also made between a mixed effects model that had a random slope (per participant) for expected value and another model with random slopes (also per participant) for expected utility. Expected utility was calculated using the rho and lambda parameters from the gambling task according to Equation 2.2. In the Gains condition, the Bayes Information Criterion (BIC) was 1744.293 for the expected value model and 2113.857 for the expected utility model. In the Losses condition, the BIC was 1882.645 for the expected value model and 3403.532 for the expected utility model. These results clearly show that the estimated risk and loss aversion parameters did not present significant explanatory value when regressing on delegation decisions. The expected value models were also expanded by including a term that represents how many correct outcomes participants received during the second half of the second block in the practice phase. These models were intended to capture a potential effect of a fictitious learning signal (given that the learning task was unlearnable). Such models did not perform better than the models that included just a term for expected value (BIC: 1751.684 in Gains and 1890.048 in Losses).

2.2.12 Analysis of control premiums

Because participants' expected value was £5 in the gain block and -£5 in the loss block (their probability of choosing correctly was always 50%), a value maximiser's indifference point would be £5 for gain and -£5 for loss if she had an accurate perception of her ability. Thus the "control premium" - the amount participants are willing to forgo to retain agency – would be equal to her indifference point minus £5 (or -£5 for loss). However, participants' perception of their own ability is not completely accurate. Thus, if the participants' perception of their own accuracy (SPA) is taken into account, the control premium is equal to their indifference point minus the SPA.

2.2.13 Analysis of money forgone

To calculate the amount participants forgo in order to retain control, the average expected value for all of a participant's choices was calculated, relative to the amount she could have received if she had selected to delegate optimally. This measure is different from the control premium described above since it averages over all gains and losses for all actual choices throughout the experiment without estimating each participant's indifference point.

2.3 Results

2.3.1 Delegation

Participants had a strong preference to retain agency. While a value maximiser would delegate 50% of the time, participants' average delegation rate was significantly lower (mean delegation for Gains = 28.57%, one sample *t*-test compared to 50% $t_{49} = 8.84$, $p < 0.001$; mean delegation for Losses = 29.27%, one sample *t*-test compared to 50% $t_{49} = 8.36$, $p < 0.001$). Importantly, no differences were observed between Gains and Losses for this measure $t_{49} = 0.549$, $p = 0.586$, or any other measure reported below. A Bayes factor analysis (Morey & Rouder, 2011) was conducted to further test for null differences between the two conditions. This analysis outputs a ratio between the null hypothesis (i.e. conditions do not differ) and the alternative hypothesis (i.e. conditions differ). The analysis strongly supports the null hypothesis of no differences. It is 5.64 times more likely that the proportion of delegations was equal in the Gains and Losses conditions than that they were different.

2.3.2 Delegation "errors"

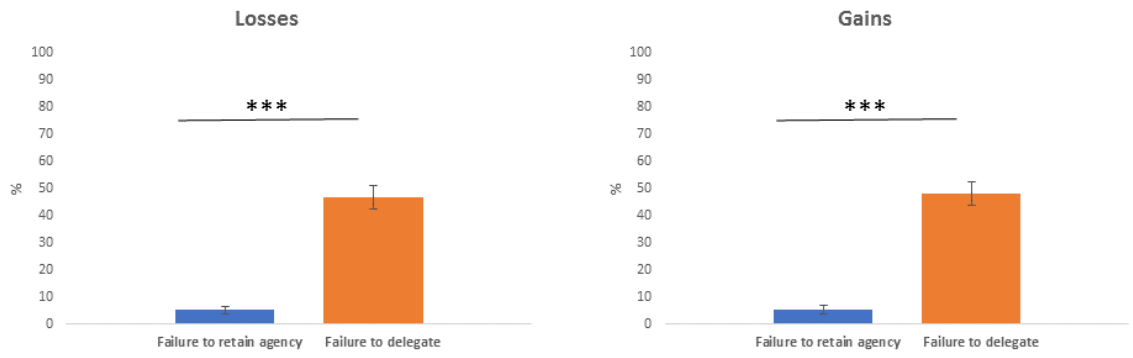

Figure 2.2. Delegation "errors".

Participants were more likely to fail to delegate when delegation was optimal (orange bars) than fail to retain agency when retaining agency was optimal (blue bars). This was true for both blocks; (a) Gain block, (b) Loss block. ***$p < 0.001$. Error bars are standard errors of the mean.

Participants were much more likely to retain agency when this was not the optimal decision (i.e., "failure to delegate") than to delegate when this was not the optimal decision (i.e., "failure to retain agency"). Specifically, out of all trials where delegation was optimal,

they chose to retain agency, failing to delegate, 47.33% in the gain block and 48.07% in the loss block (Figure 2.2). In contrast, out of all trials where agency was the optimal decision they chose to delegate, failing to retain agency, only 5.2% in the gain block and 5.13% in the loss block (Figure 2.2). As a result of these failures, participants earned £0.71 (s.d. = 0.495) less than they could if they selected optimally in the gain block and lost £0.68 (s.d. = 0.467) more than they should have in the loss block.

Note, however, that while participants were not maximizing their rewards, they were not delegating at random. Rather, they were more likely to delegate on trials when delegation was optimal (51.93%) than on trials when delegation was not optimal (5.20%) $t_{49} = 11.40$, $p < 0.001$) in both the gain block and in the loss block (53.4% and 5.13% respectively, $t_{49} = 11.77$, $p < 0.001$), indicating that they were sensitive to the expected utility of delegating to the advisors.

2.3.3 Self-perceived accuracy

Participants' perceptions of their own ability to select the shape associated with the better outcome was relatively accurate (average estimate was 56.68%) but differed from the true performance rate of 50% ($t_{49} = 3.07$, $p = 0.003$). Thus, participants' preference for agency might be explained by overconfidence in their ability to choose accurately. To account for this, optimal delegation was redefined, considering each participant's perception of their own accuracy (see methods section above). Even when taking this into account, participants were still more likely to retain agency on trials when delegation was in fact optimal (failure to delegate on gain block = 44.07%, loss block = 42.27%) than to delegate when retaining agency was in fact optimal (failure to retain agency on gain block = 8.73%, loss block = 7.87%). The frequency of the two types of errors was significantly different from each other (Gain: $t_{49} = 5.91$, $p < 0.001$; Loss: $t_{49} = 6.74$, $p < 0.001$).

2.3.4 Indifference point & control premium

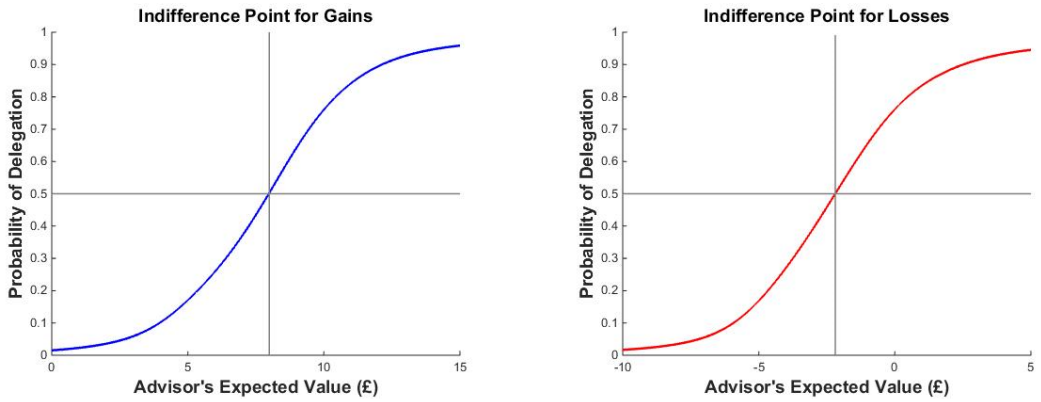

Figure 2.3. Indifference points.

The indifference point is the point when the probability of delegating is 50%. While a rational agent should be indifferent between retaining agency and delegating when an advisors' expected value is £5 in the gain block and -£5 in the loss block (participant's expected value is £5 in the gain block and -£5 in the loss block) the graphs clearly shows that in practice the indifference point is greater than £5 in the gain block and -£5 in the loss block. This suggests that participants assign positive utility to choice. The grey lines show the intersection between 50% probability of delegation and the indifference curve. The curves shown here are the model predictions for the group average.

When would participants be indifferent between making the decision themselves and letting the advisor make it for them? A rational agent should be indifferent between delegating the choice and retaining agency when the advisor's expected value is £5 in the gain block and -£5 in the loss block. This is because the participant's own expected value is £5 for gains and -£5 for losses. Thus, in those situations it should not matter who makes the decision.

This benchmark was compared to the indifference points that were observed given the participant's responses (except for one participant for whom an indifference point could not be estimated given the low variance in that participant's responses in the Gains condition). Participants' average indifference point was £8.15 (s.d. = 2.92) in the gain block and -£0.822 (s.d. = 7.23) in the loss block, which is significantly different from £5: gains: $t_{48} = 7.55$, $p <$

0.001, and -£5: loss: $t_{49} = 4.05$, $p < 0.001$. In other words, participants were willing to forgo £3.15 (i.e. = £8.15 -£5) (s.d. = 2.92) in the gain block and lose an extra £4.18 (i.e. = -£0.822-(-£5)) (s.d. = 7.23) in the loss block in order to retain agency (Figure 2.3).

This number is referred to as a control premium. Re-calculating the control premium, considering participants' self-perceived accuracy (SPA) by subtracting the SPA of each participant from their indifference point, results in the following control premium estimates; £2.50 (s.d. = 3.18) for gains and £3.53 (s.d. = 6.81) for losses. These control premiums are significantly different from the previous estimates, for both gains and loss: $t_{48} = 2.94$, $p = 0.005$. Even though controlling for SPA reduces the control premium estimate, the estimate is still much greater than zero, gains: $t_{48} = 5.50$, $p < 0.001$, and loss: $t_{49} = 3.63$, $p = 0.001$. Thus, the control premium seems to stem from an intrinsic value for control.

2.3.5 Perceived delegation accuracy

In addition to asking participants to report self-perceived accuracy of choosing the correct shapes, participants were asked to estimate how accurate they were at delegating. Participants' estimates were surprisingly accurate and did not differ significantly from how well they were in fact delegating. The latter was calculated as the percentage of trials on which they made optimal decisions (i.e. delegating when they should be and retaining agency when they should be). Comparing perceived delegation performance and actual delegation performance showed no difference for gains: $t_{49} = 0.04$, $p = 0.97$, or losses, $t_{49} = .30$, $p = 0.77$. Neither did perceived accuracy of delegation and actual accuracy of delegation differ when optimal delegations were defined taking into account SPA (gains, $t_{49} = 0.12$, $p = .90$, losses, $t_{49} = 0.69$, $p = 0.50$).

The two scores correlated with each other; in other words, participants' perceived accuracy of delegation correlated with actual accuracy of delegation defined based on SPA (loss: $r_{48} = 0.38$, $p = 0.007$, marginally for gains: $r_{48} = 0.27$, $p = 0.054$). SPA and perceived accuracy of delegation were also correlated ($r_{48} = 0.49$, $p < 0.001$). Together, these results suggest that participants knew how well they were delegating and were aware that by retaining control they were losing money – yet chose to do so nevertheless.

2.3.6 Loss and gain questions

To show that the participants made an explicit distinction between loss and gain blocks, they were asked (i) which block made them happier - 100% of participants responded that they were happier during the gain block than loss block; and (ii) which block they would prefer to do again – 96% selected the gain block. In addition, (iii) 92% of participants would pay more money to repeat the gain block over the loss block.

2.4 Discussion

The findings in this chapter show that high-level features like control preferences and overconfidence, studied through a top-down approach, are features of choice option representations that can have an effect on behavioural outcomes. Furthermore, it is clear that these features are only relevant given the way the task goal was constructed (i.e., as a choice between the self and an advisor). Focus was placed on features concerning representation of the self since the only information that was provided about the advisor was their accuracy and the cost of their services. The accuracy of the advisor represents uncertainty in this task and was presented as a point estimate (i.e., a property of the stimulus format). Providing richer information regarding uncertainty was restricted here to allow focus on self-representation (see chapter 3 for a relaxation of this restriction on stimulus formats for uncertainty).

The results demonstrate that participants are willing to forgo rewards for the opportunity to make their own choices and hence to control their own payoffs. Thus, placing a premium on control shapes the agent's utility function, having implications on the way they solve such a task goal. This preference was observed not only when faced with potential gains (in accord with Owens et al. 2014), but also when faced with potential loss. Moreover, the findings indicate that participants accurately assess the (sub)optimality of their delegation choices, suggesting that they are aware of the premium they are paying to maintain control.

The results could not be accounted for by participants' overconfidence in their ability to maximise rewards and minimise losses, as their beliefs regarding their own ability were elicited and accounted for. In fact, on average participants were slightly overconfident even though they were unlikely to hold prior beliefs about their ability from "real world" experience and were given plenty of experience with the task. In this case, self-representation as being better than they really are did not have as big an effect as the construction of the exact utility

function that considered the premium placed on control. Participants were also given complete information about potential advisors, which would allow them to make rational decisions. Thus, under-delegation could not be attributed to a systematic misperception of either the participant's expected utility nor the advisor's expected utility. Indeed, when questioned about their ability to delegate accurately, participants' representations of their delegation performance were surprisingly accurate.

The findings are in accord with a past study that identified a significant "control premium" in an experimental setting in which participants could bet that a partner, or instead they themselves, would answer quiz questions correctly (Owens et al. 2014). In light of participants' elicited beliefs, participants should have bet on themselves 56.4% of the time – but in fact, did so 64.9% of the time. This suggests that representation of themselves as overconfident only explains part of the effect but that the specific control-adjusted utility function explains the rest. It follows that participants were, on average, willing to give up 8% to 15% of their expected earnings to retain control. In other words, the preference for control could not simply be explained by overconfidence or subjective beliefs; control appeared intrinsically desirable. In a way, their overconfident self-representation was aligned with their specific utility function (i.e., the task goal).

Fehr et al. (2012) also find, in a fundamentally different design, that people will sacrifice their material interest to maintain authority. They conduct an "authority game," in which principals could choose to delegate decisions to an agent. Their central finding is that people will under-delegate, showing "a strong behavioural bias among principals to retain authority against their pecuniary interests and often to the disadvantage of both the principal and the agent." Their major explanation is that people do not like to be overruled, and they know that if they delegate, their agent might disregard their information, or their wishes, in order to make their own selection. Using a different experimental design, Bartling et al. (2013) similarly find that decision rights have intrinsic and not merely instrumental value. There is also a relationship here between the findings and the phenomenon of "reactance," which suggests that people rebel against choice-denying commands by defying them (Brehm and Brehm 2013).

The results support the past findings for the existence of a "control premium" and extend them by demonstrating a positive control premium not only in the domain of potential

gains, but also in the domain of potential loss. Furthermore, the results suggest that participants are aware that they are selecting to pay a premium for control. Future studies may wish to vary levels of expertise for the participants, include a measure of reliability or trust for the advisors, and perhaps include a variable that modulates effort that the participant may need to realize when he or she decides to retain control. Likewise, using this paradigm to study depressive populations could yield insights into the relationship of depression and locus of control; perhaps depressive populations would delegate more given that their locus of control is seen as driven by external forces (Benassi, Sweeney, & Dufour, 1988).

Why would people choose to under-delegate when they are seemingly aware of the material cost of under-delegation? Speculatively, this behaviour reflects a non-monetary intrinsic value for control, which is expressed via choice. The intrinsic value of choice may have emerged for several reasons. First, outcomes that we select ourselves often suit our preferences and needs more than those that have been selected for us (Beattie et al. 1994; Kray 2000; Polman 2010; Stone and Allgaier 2008; Waldfogel 1993). Thus, we have learned that environments in which we can exercise choice are usually (of course not always) more rewarding (Leotti et al. 2010; Patall et al. 2008; Rotter 1966). The frequent association between choice and reward may have led choice itself to be experienced as rewarding - something we seek and enjoy.

Second, a biologic system that provides higher intrinsic reward for things we have obtained ourselves compared to those that were simply chosen for us may be adaptive. If we learn that an action results in a reward, we can repeat that action in the future to gain more of the same. However, if we do not execute an action to obtain reward (or avoid harm), we lose the opportunity to acquire a "blueprint" of how to gain rewards (or avoid harm) again in the future. Thus, the value of outcomes we had obtained ourselves emerge both from their utility and from the information they contain for future outcomes. This means that features of self-representation, such as the intrinsic value of control, may reflect adaptive pressures from the environment.

In sum, in this controlled laboratory experiment the results show that people are willing to pay a control premium to make their own choices. With a somewhat different design, it is found that people will pay a control premium in the domain of gains and a similar premium in the domain of losses. This finding runs counter to the idea that people prefer to delegate

decisions involving unwanted outcomes in order to avoid regret (Loomes and Sugden 1982) and instead supports the notion that choice may be preferred regardless of expected valence of the outcome (Cockburn et al. 2014; Leotti and Delgado 2011, 2014; Sharot et al. 2009, 2010a). Moreover, the current study suggests that people are aware of the monetary premium they are paying to retain agency, but do so anyway, presumably for psychological benefit; impacting the construction of their utility function and the way they solve the task goal. Thus, in the normative sense, choosers can be losers, and knowingly so.

However, it is possible that the results are better explained by the fact that people have greater precision over their likelihood of picking the correct shape than over the likelihood of the advisor doing so. Aversion to ambiguity may thus contribute to the preference for control. This means that the exact representation of uncertainty can have an influence on which choice option (the self or the advisor) is opted for. Such a theme is expanded upon in the next chapter where uncertainty in the choice options is modeled under Bayesian statistical assumptions.

With this first study it is possible to see that defining a task goal imposes a cascade of constraints that interact with each other. Here the task goal was to maximise expected utility within a social domain. The specifics of the task involved a choice between two agents where the representations of these agents (choice options in this case) are relevant. Thus, the stimulus formats further constrained the study of choice option representations given the fact that the task goal was already defined. The stimulus format here consisted of verbal and numerical representations regarding the advisor and absent stimulus for the self (provided that this stimulus has already been internalised by the participating agent). In the next chapter, the study will show how modifying assumptions about the task goal (with respect to computation of uncertainty) requires relaxing stimulus formats constraints.

Chapter 3 Representation of choice options with respect to personal and social preferences: an fMRI study

3.1 Preference integration study

Whereas in the previous chapter the stimuli constraining representations of the self are absent but inferred through task goal assumptions on the maximisation of expected utility, here the stimulus format enables elicitation of features regarding the self in the form of retail product ratings (participants here are shown products from the Amazon retail website). Personal preferences are taken to be a feature of the self but in relation to a retail product. The representation of the retail product is informed through personal preferences (i.e., the value adjugated to the product). What is of interest here is how social ratings of retail products can exert forces on the representation of value. The research question is about how integration between personal preferences and preferences of others are integrated, specifically: 1) does this integration between preferences of the self and of others consider uncertainty? 2) is uncertainty updated in a roughly Bayesian normative manner? and 3) is there any indication that this type of process is implemented in the human brain? This chapter addresses these questions through a top-down view that uses a Bayesian computational model as a different perspective to how uncertainty is treated in maximizing expected utility. This perspective requires modification of the stimulus format with respect to the previous chapter since distributional information now needs to be conveyed to the participants.

We may not like to admit it, but our own opinions are greatly influenced by those of other people. When we book a holiday, buy a new electronic device or choose a film to watch we often rely on the opinions of other people expressed in the forms of reviews. Taking other people's judgments into account can be a sensible strategy for a social species. Humans have similar needs and therefore often share preferences with others in their socio-demographic group. The effect of social influence on judgments (i.e. social conformity) has been a topic of intense investigation (Cialdini & Goldstein, 2004), and in more recent years the field of cognitive neuroscience has begun to dissect the circuitry underpinning social conformity (Behrens, Hunt, & Rushworth, 2009; Berns, Capra, Moore, & Noussair, 2010; Campbell-Meiklejohn, Bach, Roepstorff, Dolan, & Frith, 2010; De Martino et al., 2013; Izuma & Adolphs, 2013; Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009). Social conformity is part of a more general process that considers social information as having strong

importance for the well-being of the individual (Rutledge, De Berker, Espenhahn, Dayan, & Dolan, 2016). In this study, social conformity is expressed as a function of integrating personal preferences with respect to a norm (i.e., social preferences). Thus, preferences themselves, personal and social, are the relevant elements of choice option representations selected for scrutiny in this investigation.

The social information we receive, much like our own beliefs, varies in its reliability or uncertainty. For example, should one purchase headphones on Amazon's website with a 4-star average based on hundreds of reviews or a competing product with a 5-star average based on only a few people's opinions? In such circumstances, people should be sensitive to both the opinions of others but also to their prevalence. The way people represent uncertainty in personal and social preferences with respect to a retail product will have an effect on the final choice outcome.

Hence, this study is not concerned as much with the representation of the retail product (this is an issue that is more directly studied in chapter 5 through the neural representation of choice options). The aim of the current study is to investigate whether the human brain integrates social information according to its reliability and how this in turn affects valuation and confidence judgments. More specifically, this study evaluates whether people integrate their initial beliefs and those of others in a Bayesian fashion such that the combination is weighted by the uncertainty of each source of information. For example, according to the Bayesian view, people should update their beliefs most toward the social consensus when they are initially uncertain about the value of the headphones and there are a large number of Amazon reviewers. The task goal here is not just maximisation of expected utility but optimal integration of uncertainty between personal and social preferences weighted by (perceived) reliability. The struggle between maintaining personal preferences or integrating new ones from other people is just another example of a trade-off that needs to be dealt with by the representations of choice options.

As mentioned previously, Bayesian inference is a normative framework for how prior beliefs are updated in the light of new information (Jill X. O'Reilly, Jbabdi, & Behrens, 2012; Vilares & Kording, 2011). One empirical signature of Bayesian integration is that the relative uncertainties of an individual's prior beliefs and some external source of information should govern how the information is combined. The Bayesian approach has been successful in

providing a compact description of how beliefs are updated during perceptual decision-making, multisensory integration (Angelaki, Gu, & DeAngelis, 2009), motor control (Ernst & Banks, 2002; Knill & Pouget, 2004; Körding & Wolpert, 2004; Summerfield & Koechlin, 2008) and also higher level cognitive abilities such as memory, language, and inductive reasoning (Chater, Tenenbaum, & Yuille, 2006). However, it is still unknown whether prior beliefs and social information are integrated in a Bayesian fashion that weights the information sources by their uncertainty. How this process would be implemented in the brain is also an open question.

This study tests whether people integrate social information with their prior beliefs in a Bayesian fashion and examine how the integration process is implemented in the brain. The main focus of the neural analysis is medial prefrontal cortex: more specifically the ventromedial (mPFC/vmPFC) and dorsomedial medial (dACC/dmPFC) sub-regions. The first region (mPFC/vmPFC) has a well-established role in representing value estimates (Clithero & Rangel, 2013; Levy & Glimcher, 2012) and more recently, it has been proposed that the same region tracks the reliability in these estimates (Barron, Garvert, & Behrens, 2015; De Martino et al., 2013; Donoso, Collins, & Koechlin, 2014; Lebreton, Abitbol, Daunizeau, & Pessiglione, 2015; Rolls, Grabenhorst, & Deco, 2010). The second region (dACC/dmPFC) was chosen because of its central role in social cognition (Amodio & Frith, 2006; Gallagher & Frith, 2003; Kolling, Behrens, Wittmann, & Rushworth, 2016) and more specifically in mediating social influence over value computation (Campbell-Meiklejohn et al., 2012; De Martino et al., 2013; Hampton, Bossaerts, & O'Doherty, 2008; Nicolle et al., 2012; Suzuki et al., 2012). However, it is unclear how social information is integrated into value computation in prefrontal cortex. It is not known whether signal in an area like dorsomedial prefrontal cortex can detect a conflict between the group consensus, triggering a compromise to the group evaluation, or if it is involved in a more complex Bayesian updating that takes into account variable levels of reliability in the social information as well as the level of confidence in the prior belief. Generally speaking, how does the neural representation of uncertainty (for choice options) handle the trade-off between personal preferences and social preferences?

3.2 Materials and methods

3.2.1 Participants

Twenty-two participants aged 18 to 35 (mean age = 24.82, s.d. = 4.10, 11 female) were recruited from University College London (UCL) psychology participant pool. One participant

was excluded because of a scanner technical problem. Another participant was excluded because of excessive head motion (>3° rotation on 4 occasions). Another two participants were excluded because of erratic product ratings (>3 skewness). A total of 18 participants were therefore included in the final analysis. The study was approved by the UCL Psychology Ethics Committee. Written informed consent was obtained from all participants and they were paid for participation.

3.2.2 Stimuli

Stimuli consisted of 210 pictures of products from the retail website Amazon (https://www.amazon.co.uk/) along with the product name. Each picture was presented once in each task (pre-scanning task and scanning task, see below) to participants in a randomised order. Four to five bullet points with descriptions of each product were provided in the pre-scanning task. These descriptions were based on the information available for the products on the Amazon website. During the task in the scanner, they were also presented with summary reviews of the products. This information was presented exactly as it is shown on the Amazon website: the mean of the reviews (1 to 5 stars), the number of reviewers, and a 5-bar histogram showing the distribution of ratings across reviewers (right hand side of Figure 3.1A).

3.2.3 Pre-scanning task

Participants were required to make a series of product ratings for 210 Amazon products. Participants were required to give their liking rating for each item (left hand side of Figure 3.1A) and their confidence in their liking rating. A fixation cross was presented for 500 ms. Participants then moved the slider located at the bottom of the screen to indicate their rating of the product. The location of the picture of the product and the respective bullet point descriptions were left-right counterbalanced across trials. The starting position of the slider was randomised on each trial. After deciding the product rating, the slider confirmed the selection by changing to the colour red for 500 ms. Once they provided the product rating, participants were asked to indicate their confidence in their decision on a continuous sliding scale with six ticks but no numbers, with text going from “Lower” to “Higher” confidence. After deciding on a confidence rating, the slider confirmed the selection by changing to the colour red for 1000 ms. There was no time limit for participants to rate a product or indicate their confidence rating. The 210 trials in which they did product and confidence ratings were divided into three blocks of 50 trials and one final block of 60 trials. The direction of the product rating scale and

the confidence scale were reversed after two blocks of trials. If a participant started the experiment with a left to right presentation of the scales (1 to 5 stars and “Lower” to “Higher” confidence, respectively), then after two blocks of trials (100 trials), participants would see the scales in right to left presentation (5 to 1 stars and “Higher” to “Lower” confidence, respectively). This is necessary to avoid visual and motor confounds during imaging in the scanning task, which is why it is preferable for participants to get accustomed to this procedure during the pre-scanning task. The direction of the scale for the first two blocks of trials was randomly chosen across participants. The pre-scanning session was conducted the same day of the scanning task.

3.2.4 Scanning task

The scanning task presented the same 210 products that participants rated in the pre-scanning task. In this task, participants did not see the product descriptions. Instead, they were presented with information on other people’s ratings retrieved from Amazon.co.uk. In particular, the scanning task showed the number of people that rated the product, the mean rating of the product (on a scale from one to five stars), and the distribution of ratings. An example screen shot is provided in Figure 3.1A (right hand side). Participants did not see their own rating from the pre-scanning task and were free to change their ratings in the light of other people’s ratings. Participants were incentivised in this task since they were told that a product would be selected at random at the end of the experiment and would be given to them at a later date as part of their compensation. They were told that the higher their rating for a product, the better the chances they would have in receiving that product. Products had a similar retail price range.

As in the pre-scanning task, a fixation cross was presented, participants decided on a product rating, and then the slider turned red for 500 ms before moving on to the confidence rating. The duration of the initial fixation cross was jittered. Unlike the pre-scanning task, participants were only allowed 7 seconds to rate a product and 4 seconds to report their confidence. Therefore, the timeline of the fMRI task was the following: fixation cross (jittered between 500ms and 1500 ms), item presentation + liking rating scale (7000ms), and confidence rating (4000ms).

3.2.5 Post-scanning choice task

At the end of the functional scans, and during the structural scan, participants made 49 forced choices between a pair of products that were both previously rated during the preceding scanning task. Each pair contained one product with a low rating (randomly sampled from the bottom tercile) and one with high rating (randomly sampled from the top tercile). Participants selected the item from the top tercile on 77.29% (s.d. = 11.07) of the forced choices.

3.2.6 Image acquisition

Scanning acquisition was performed using a 1.5 T Siemens TIM Avanto MRI Scanner with a 32-channel head coil used to acquire both T1-weighted structural images and T2*-weighted echoplanar images (64 x 64; 3 x 3 mm voxels; echo time, 50 ms; repetition time, 3132 ms; flip angle, 90 degrees; field of view, 192 mm) with blood oxygen level-dependent (BOLD) contrast. Each volume comprised 36 axial slices (2 mm thick). A specific sequence that improved the signal-noise ratio in orbitofrontal cortex, a region that usually suffer from signal drop-off (Deichmann, Gottfried, Hutton, & Turner, 2003), was used. To further minimise this problem, it was decided to acquire the imaging data in a 1.5 Tesla MRI scanner, which suffers from less-pronounced dropout in this region, and therefore can actually have greater BOLD sensitivity than higher field-strength scanners (Weiskopf, Hutton, Josephs, & Deichmann, 2006). Functional scans were acquired in four sessions, each comprising 228 volumes (~10 min). The first five volumes in each session were discarded to allow for T1 equilibration effects. At the end of the fourth functional scan, a 5.5 min T1-weighted MPRAGE structural scan was collected, which comprised 1mm thick axial slices parallel to the AC-PC plane.

3.2.7 fMRI data analysis

Image pre-processing was performed using Statistical Parametric Mapping 12 (SPM 12, Wellcome Trust Centre for Neuroimaging, http://www.fil.ion.ucl.ac.uk/spm/). Image analysis was performed using SPM 12. After discarding the first five dummy volumes, images were realigned to the sixth volume and unwarped using 7th degree B-spline interpolation. Field maps were reconstructed into a single-phase file and used to realign and unwarp EPI functional images. Structural images were reregistered to mean EPI images and segmented into grey and white matter. These segmentation parameters were then used to normalise and bias correct the

functional images. Normalisation was to a standard EPI template based on the Montreal Neurological Institute (MNI) reference brain using a nonlinear (7th degree B-spline) interpolation. Normalised images were smoothed using a Gaussian kernel of 8 mm full-width at half-maximum.

Two independent general linear models (GLMs) were performed. In GLM1, onset regressors were set at the beginning of trial presentation. Events were modelled by convolving a series of delta (stick) functions with the canonical HRF at the beginning of each item presentation. These onsets were modulated by two parametric regressors: (i) liking rating (R2); and (ii) post-choice confidence ratings (C2), which ranged from 0 to 500 on an arbitrary scale. In GLM2, onset regressors beginning at the presentation of the item was modulated by one parametric regressor: (i) KL trial-by-trial parameter estimate computed by fitting a descriptive Bayesian model to the behavioural data. Both GLMs included 6 movement regressors. In GLM2, two further participants had to be excluded since the KL parameter was zero in a number of instances: this resulted in the model not being estimable in SPM. Note that the parametric regressors for both GLMs were not-orthogonalised and regressors were allowed to compete to allocate the shared variance (Mumford et al., 2015). Contrast images for each regressor were tested for a significant deviation from 0 using one-sample *t*-tests. Activations were reported as significant if they survived family-wise error correction (FWE) for multiple comparisons across the whole brain at the cluster level. Note that the cluster forming threshold was set as $p < 0.001$ uncorrected to ensure an a well-behaved family-error control (Eklund, Nichols, & Knutsson, 2016; Flandin & Friston, 2016). For dmPFC isolated in the GLM2, small-volume correction using an 8-mm sphere centred on the coordinates ([-3,51,24]), taken from an independent study (Hampton et al., 2008), was employed. The rfxplot toolbox (http://rfxplot.sourceforge.net/) (Gläscher, 2009) was used to extract percentage signal change at each region of interest defined by 8-mm spheres around and used for the histogram plots. Note the signals are not statistically independent (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009) and these plots aren't not used for statistical inference (which was carried out in the SPM framework). It is shown solely for illustrative purposes (i.e. clarify the signal pattern in each cluster), this has been explicitly stated in the figure legends.

3.2.8 Behavioural data analysis

Hierarchical regression analyses were performed in R using lme4 package (Bates, Mächler, Bolker, & Walker, 2014). Participants' product (R1 and R2) and confidence (C1 and C2) responses were normalised (*z*-scored) separately for each participant for each of the four judgment types to correct for any potential differences in scale usage.

3.2.9 Bayesian update model

The model worked with the same *z*-scored data as used in the behavioural analyses and was fit to individual participants. First, the prior distribution (shown in blue in Figure 3.4A) was formalised as a Gaussian distribution. For each product *j*, the mean of this distribution for participant *i* was determined by the parameter, $\mu_{i,j}$. For the prior variance, each participant *i* had a variance parameter σ_i^2, plus a non-positive offset parameter σ_i' that was included for higher confidence trials. Thus, the prior distribution for participant *i* for product *j* is

$$\mathcal{N}(\mu_{i,j}, \sigma_i^2 + \sigma_i' I_{(C1_{i,j} > \text{median}(C_i))}) \tag{3.1}$$

where *I* is an indicator function returning 1 when confidence was rated above the median, and 0 otherwise. According to the Bayesian model, higher confidence should correspond to lower variance (i.e., greater precision). The use of the median split simplifies the model and reduces the number of assumptions needed to relate the model to the behavioural data.

The distribution of Amazon reviews for a product was also Gaussian (shown in Figure 3.4A in yellow). The mean was fixed to m_j, the observed mean of the amazon ratings for product *j*. Each participant *i* had a single parameter, τ_i^2, for the perceived variance (i.e., reliability) of the Amazon reviews in general, plus two parameters related to the number of Amazon reviews. Thus, the Amazon reviews for product j were parameterised to contain $v_i + v_i' I_{(n_j > \text{median}(n))}$ reviews, where n_j was the number of Amazon reviews as presented to participants during the experiment and v_i and v_i' are non-negative parameters. As with confidence in the prior, this median split of the parameters by the number of reviews mirrors the behavioural analyses. A posterior distribution (shown in green in Figure 3.4A) was then derived using Bayes theorem, and therefore, the model did not have a parameter specifically for a posterior distribution.

In summary, the model, which characterises the degree to which participants integrate information, accounts for 420 ratings (210 initial and 210 second ratings) from each participant with 210 parameters ($\mu_{i,j}$) for prior means, 2 parameters ($\sigma_i^2; \sigma_i'$) for prior variance, 2 parameters ($v_i; v_i'$) for the perceived number of Amazon reviews, and 1 parameter (τ_i^2) for the perceived variance in Amazon reviews. These parameters are necessary to include to account for individual differences in how the social ratings from Amazon are processed and interpreted. The parameter values were estimated independently for each participant, to maximise likelihood of both initial and second ratings. Estimated prior mean and derived posterior mean show strong positive correlations with initial and second ratings: across 18 participants, correlation coefficients range from 0.75 to 0.96 (mean: 0.90, 95% CI: [0.88, 0.93]) between prior mean and an initial rating, and from 0.85 to 0.96 (mean: 0.90, 95% CI: [0.89, 0.92]) between posterior mean and a second rating, which indicates a good fit.

Because the model was not fit to the confidence ratings, one avenue to evaluate the model is to compare the precision of its posterior to participants' second confidence ratings. Model precision should positively correlate with confidence. Correlation coefficients ranged from 0.14 to 0.40 (mean: 0.18, 95% CI: [0.12, 0.24], $t_{17} = 5.95$, $p < 0.001$). The main justification for the basic approach (i.e., integrating prior and likelihood information according to their uncertainties) comes from the behavioural results reported below.

Using the estimates from the model, the degree of resistance to Amazon reviews is computed for each participant for each product as follows:

$$\frac{\text{Prior precision}}{\text{Prior precision+Percived precision Amazon Rating}} \tag{3.2}$$

Here, prior precision is the inverse of prior variance, and perceived precision of Amazon rating is estimated Amazon *N* divided by estimated Amazon variance. Given Bayes' theorem, the above specification captures how heavily prior mean is weighted toward posterior mean.

Specifically, the degree of resistance to Amazon reviews is 1 when the Amazon rating is completely ignored and prior mean is the same as posterior mean. Also, the degree of resistance to Amazon reviews to is 0 when prior is completely discarded and Amazon mean is

the same as posterior mean. A larger value indicates that Amazon mean is more heavily weighted toward posterior mean than prior mean is.

This degree of resistance to Amazon reviews is mean-averaged for each participant across 210 product ratings.

3.3 Results

To address the questions of interest, a task was developed in which participants were presented with a series of products from the retail website Amazon (e.g. headphones, USB-pens, mugs). Participants were required to give their initial liking rating (R1) for each item and their confidence (C1) in their liking rating (Figure 3.1A). Both measures were collected before scanning. In the second part of the experiment, participants' neural activity (using fMRI) was recorded while they were shown each item again, this time together with reviews from other customers who had bought those products (n.b. these were the real reviews from the Amazon website). This information was presented as it is shown on the Amazon website: the mean of the reviews (1 to 5 stars), the number of reviewers, and a 5-bar histogram showing the distribution of ratings across reviewers (Figure 3.1A). At this second stage, another liking rating (R2) was elicited again followed by a new confidence rating (C2).

To foreshadow the results, people followed the basic tenets of Bayesian integration. A descriptive Bayesian model consistent with these behavioural results made it possible to conduct a trial-by-trial fMRI analysis to isolate brain regions that tracked the degree to which social information and its reliability affected participants’ beliefs.

3.3.1 Behavioural results

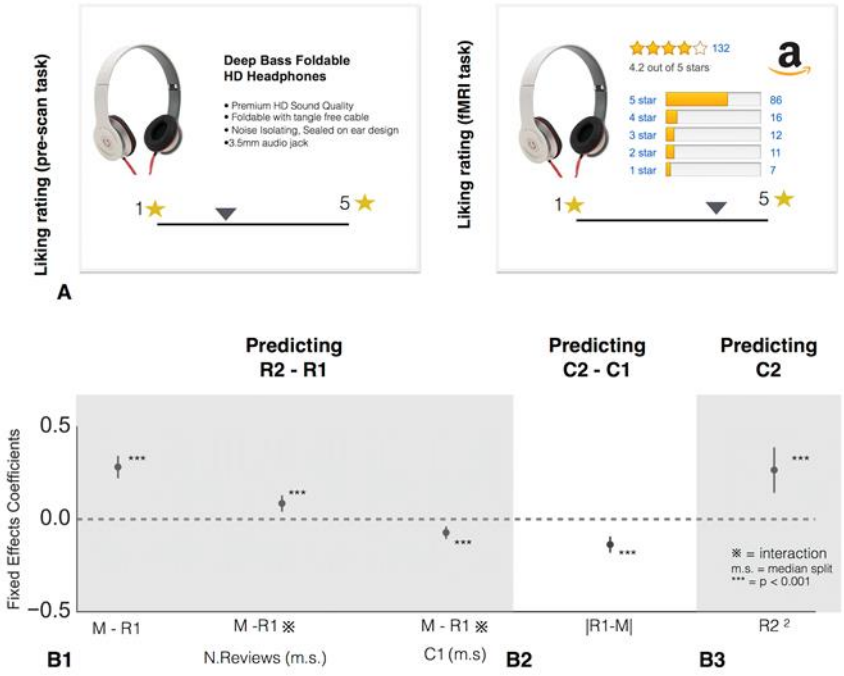

Figure 3.1. Task and fixed effects coefficients for the preference integration study.

(A)Task: In part 1 (before scanning) the participant is presented with a series of products from the retail website Amazon (e.g. headphones). The participant enters her liking rating R1 followed by her confidence rating C1 in her liking rating (not shown in the figure schematic above). In the part 2 (inside the scanner) she sees the same item again, this time together with real reviews from the Amazon website: the mean of the reviews (1 to 5 stars), the number of reviewers, and a 5-bar histogram showing the distribution of ratings across reviewers. At this stage, she is required to enter a new liking R2 and confidence rating C2. (B) All effects predicted by the Bayesian account are significant in the appropriate direction. Shown are fixed effects coefficients from hierarchical linear regression models predicting rating update (R2-R1), confidence update (C2-C1) and 2nd confidence rating (C2) for the following predictors: initial deviation from the group (M-R1), interaction between the initial deviation from the group and number of reviews (M-R1※ N. Reviews), interaction between the initial deviation from the group and 1st confidence rating (M-R1※ C1), absolute difference in a participant's initial product rating and the group consensus (|R1 - M|), quadratic function of product rating (R2). Error bars show 95% CIs., *** = $p < 0.001$ and m.s. = median split.

The central behavioural question was whether participants' initial product rating (R1) was combined with the Amazon group mean (M) in a Bayesian fashion to yield an updated

product rating (R2). The key property of Bayesian integration is weighting information by its reliability, which here corresponds to updating more toward the group consensus when initial confidence is low and the group is large. To evaluate whether people's judgments were consistent with Bayesian integration, a series of hierarchal regression analyses was conducted to assess which sources of information people considered when rating the products.

In particular, a hierarchical regression analysis was performed to isolate the factors that contributed to the update from the first to second product rating (i.e., R2 - R1). The first analysis considers whether people conform to the group mean, which in itself does not indicate Bayesian integration. It was found that participants' initial deviation from the group (i.e., M - R1) was a reliable positive predictor of participants' update $\chi^2_{(2)} = 1000.79$, $p < 0.001$) meaning that participants systematically updated their initial liking ratings in the direction of the group consensus (expressed here by the mean reviews). More complex regression models included additional terms that evaluated whether participants' judgments were consistent with aspects of Bayesian integration. In particular, interactions terms including confidence and the number of reviews were also assessed using median splits. Median splits were used because the psychological scaling of these quantities is unlikely to be linear. These scaling issues, which are topics of investigation in their own right (Kvam & Pleskac, 2016; Siegler & Opfer, 2003) are beyond the scope of this contribution.

Consistent with Bayesian updating, the magnitude of movement towards the group ratings was modulated by the level of confidence in their first rating, such that when the initial confidence was low participants were more strongly influenced by the group consensus (negative interaction between M - R1 and median split on initial confidence C1, $\chi^2_{(2)} = 15.62$, $p < 0.001$). This result is consistent with half of the Bayesian integration account, namely that participants' uncertainty in their own beliefs guides their judgments. Evaluating the other half of the Bayesian account, the update toward the group consensus (mean of the Amazon's reviews) was largest when that information was more reliable because the number of reviews was higher (positive interaction between M - R1 and median split of number of reviews; $\chi^2_{(2)} = 24.33$, $p < 0.001$) (Figure 3.1B). Finally, it was found that the full regression model, which is simultaneously taking into account both sources of uncertainty, was superior to regressions that were only sensitive to either confidence or number of reviews, ($\chi^2_{(2)} = 17.55$, $p < 0.001$ and ($\chi^2_{(2)} = 26.25$, $p < 0.001$), respectively. In summary, the change in rating from R1 to R2 was in line with Bayesian integration. The null model, that does not take into consideration

confidence or number of reviews, can be safely rejected. This justifies the use of the Bayesian update model as a descriptive characterization of the algorithmic processes that participants engaged in.

According to a Bayesian account of integration, confidence should be highest in the second rating when the initial rating and the group mean align. Indeed, the overall confidence decreased (i.e., C2 - C1) when the absolute difference in a participant's initial product rating and the group consensus (i.e., |R1 - M|) was high ($\chi^2_{(2)} = 36.79$, $p < 0.001$) and confidence elicited after the second rating (C2) was a quadratic function of product rating ($R2^2$), i.e. that confidence was highest for products at the ends of the rating scale ($\chi^2_{(2)} = 547.92$, $p < 0.001$).

Taken together, these analyses established that participants integrated their initial impression of a product and the group consensus by taking into account the uncertainty associated with each source of information (Figure 3.1B).

3.3.2 fMRI results

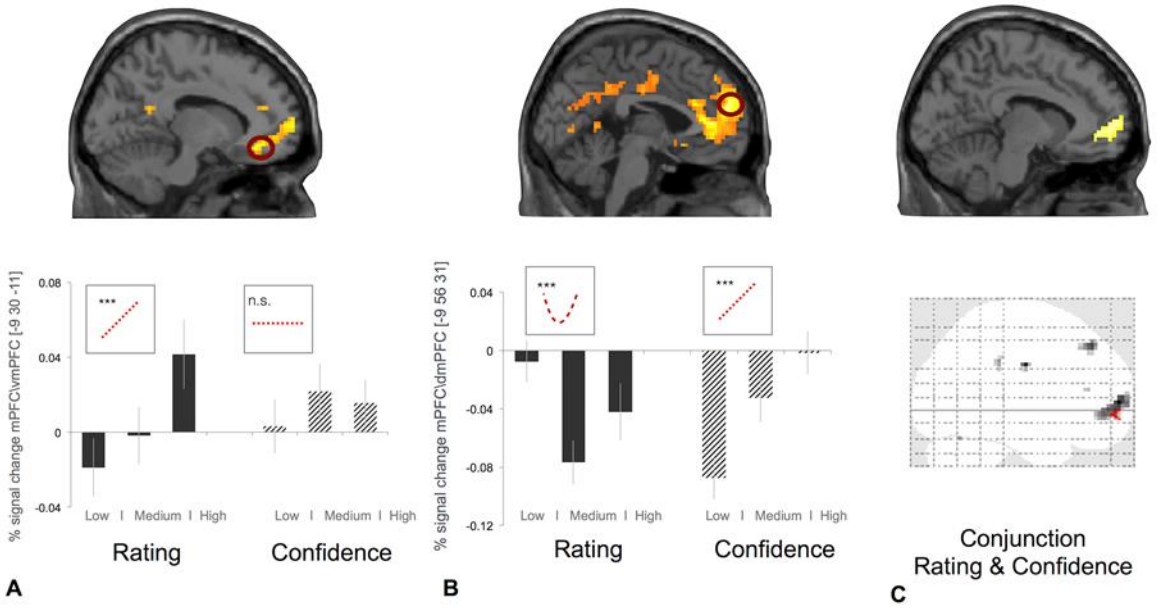

Figure 3.2. Linear and quadratic effects of ratings in mPFC.

(A) BOLD signal in mPFC/vmPFC correlates with monotonic increase in liking ratings (peak= [-9, 38, -11] mm, $z = 4.21$, $p < 0.05$, FWE corrected at cluster level). For illustration purposes only, percentage signal change in vmPFC (8-mm sphere centred at the peak of the main effect -9, 38, -11) for 3 levels or rating level and confidence (Low, Medium and High) are shown; a linear relation between % signal changes and rating level and a non-

significant (linear or quadratic) relation between % signal changes and confidence level (B) Activity in mPFC (extending in vmPFC and dmPFC) tracked monotonically the increases in confidence ratings (peak = [-9, 56, 31], $z = 4.55$, $p < 0.05$, FWE corrected at cluster level). For illustration purposes only, percentage signal change in mPFC/dmPFC (8-mm sphere centred at the peak of the main effect -9, 56, 31) for 3 levels or rating and confidence (Low, Medium and High) are shown; a linear relation between % signal change and confidence levels and a significant quadratic relation between % signal change and rating levels. The histogram plots are not used for statistical inference (which was carried out in the SPM framework); it is shown solely to illustrate the dynamic of the BOLD signal. Error bars represent s.e.m. SPM maps are thresholded at $p < 0.005$ uncorrected for display purposes. (C) Conjunction analysis for rating and confidence: activity in mPFC/vmPFC (peak activation at -12, 59, 4, $z = 3.61$, $p < 0.05$, small volume corrected at peak level using at 8-mm centred at [-2 52 -2] from (Lebreton et al., 2015)).

How the brain represents the value assigned to each item and the confidence in that value was tested. A general linear model (GLM1) was constructed in which each trial was modulated by two parametric regressors: liking rating R2 and confidence C2 (in the liking rating) both collected during the scanning (see Materials and methods above for more details). In line with previous work (for meta-analyses see (Clithero & Rangel, 2013)) the results show that activity in ventromedial prefrontal cortex (mPFC/vmPFC) responded linearly to increasing levels of subjective liking rate ($p < 0.05$, FWE corrected at cluster level – cluster forming threshold $p < 0.001$ see Materials and methods above for more details) (Figure 3.2A). In the same analysis, it was shown that medial prefrontal cortex also tracked subjective levels of confidence ($p < 0.05$, FWE corrected at cluster level - cluster forming threshold $p < 0.001$) (Figure 3.2B). To test whether liking rating and confidence in the liking rating were encoded in the same brain region, a conjunction analysis between liking rating and confidence was performed. This analysis isolated a functional cluster in mPFC/vmPFC (peak activation at -12, 59, 4, $z = 3.61$, $p < 0.05$, small volume corrected at peak level using at 8-mm centred at [-2 52 -2] from (Lebreton et al., 2015) - Figure 3.2C). This result is consistent with the recent finding that response in the same cluster in mPFC/vmPFC represents both a linear response to pleasantness rating and a quadratic explanation of pleasantness rating that in that study was used as a proxy for confidence (Lebreton et al., 2015).

Another test was performed on whether there existed a medial PFC gradient coding for confidence and value along the ventral-dorsal axis. A hierarchical linear regression model was fitted to contrast confidence vs. rating (C2-R2) extracted from 7 different locations (signal extracted by 8-mm sphere for each location) along the medial prefrontal cortex (Figure 3.3A). These locations were selected solely on an anatomical basis as opposed to by peak activity

from any preceding analysis. Across the group a significant gradient was found along the rating/confidence axis (slope = 0.02, $t_{20.95} = 9.17$, $p < 0.0001$). To confirm that the gradient was driven by both rating and confidence, two more regression analyses were performed which revealed a negative ventromedial gradient in BOLD activity in response to rating (slope = -0.01, $t_{27.06} = 4.74$, $p < 0.0001$) and a positive ventromedial gradient in BOLD activity in response to confidence (slope = 0.01, $t_{17.80} = 7.05$, $p < 0.0001$).

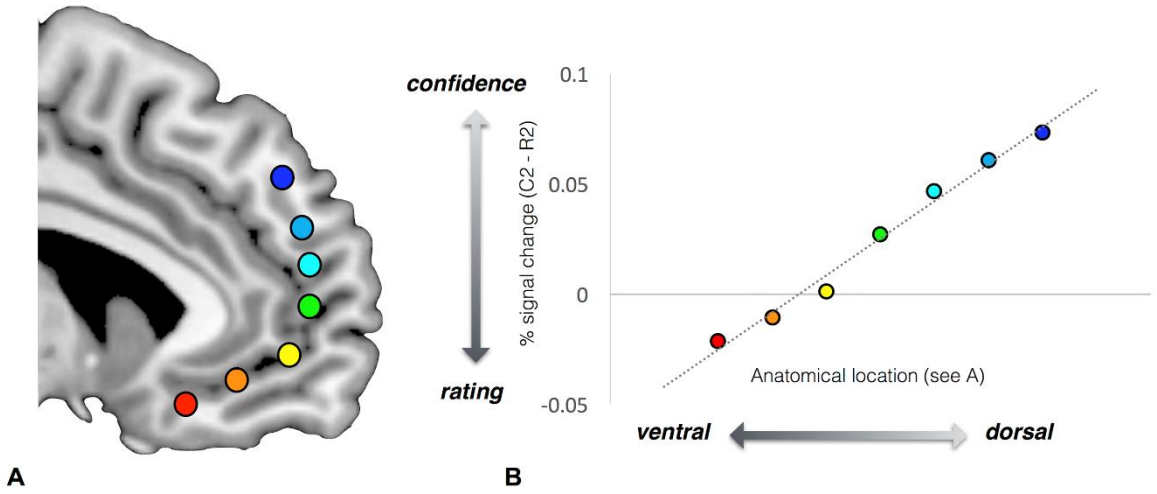

Figure 3.3. Spatial gradient analysis along the ventral-dorsal axis of mPFC (see coloured dots).

A contrast between the parametric response to rating and the parametric response to confidence (R2-C2). Data from seven anatomical locations (A) are mapped onto a line and the spatial regression slope is computed (B). Across participants there is a robust gradient along the medial lane of prefrontal cortex with response to rating expressed in the more ventral part and response to confidence represented in in the more dorsal part.

In order to quantify how social information shapes the value representation in prefrontal cortex, a Bayesian model was developed (a model schematic overview of the Bayesian model is shown in Figure 3.4A; see Materials and methods section above for full detail). The Bayesian model aimed to explain the value update with three steps: (1) an initial rating is drawn from a prior distribution, (2) this prior distribution is updated in the light of Amazon reviews to form a posterior distribution, and (3) a second rating drawn from the posterior distribution.

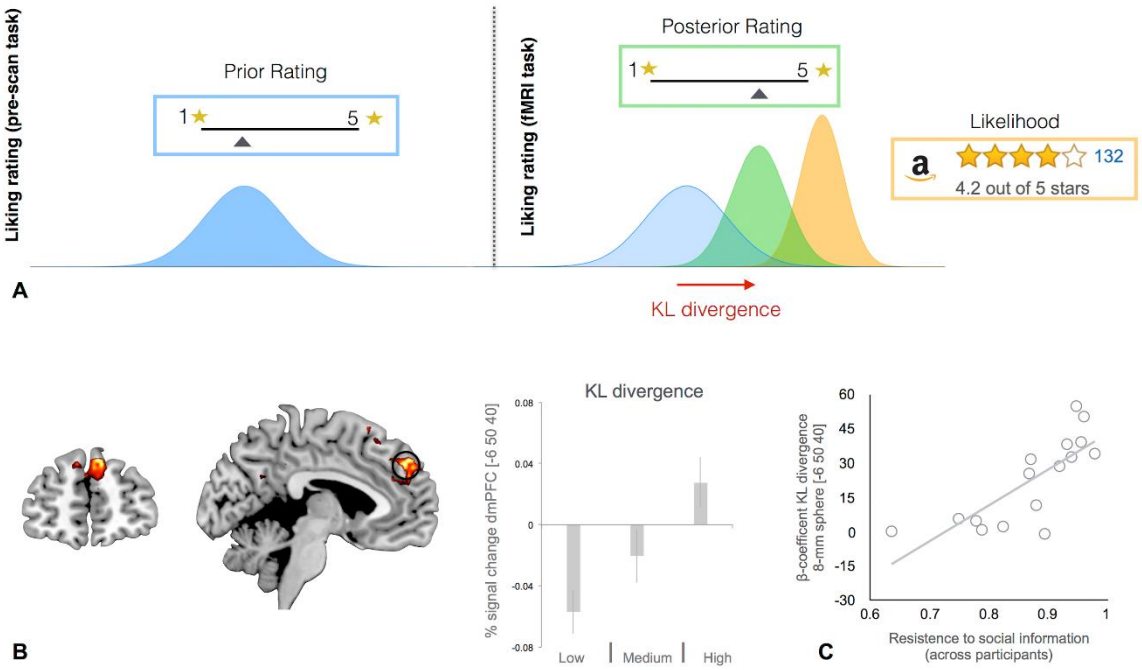

Figure 3.4. Bayesian updating and KL divergence in dmPFC.

(A) Schematic representation of the Bayesian update of liking ratings in response to social information communicated through reviews. KL divergence parameter index the impact of the reviews in shifting the liking rate from the first rating (made in the absence of review information) and the second rating (performed by the participants after seeing the Amazon reviews). (B) BOLD signal in dmPFC (peak = [-6, 50, 40]) correlates with increase in KL divergence (z = 3.66, p < 0.05, FWE small volume corrected). Percentage signal change for 3 levels (Low, Medium and High) of KL divergence. The histogram plot is not used for statistical inference (which was carried out in the SPM framework); it is shown solely to illustrate the dynamic of the BOLD signal. Error bars represent s.e.m. (C) Between-participant correlation between activity in dmPFC (8-mm ROI centred at -6, 50, 40) and the degree of resistance to social information (r = 0.77, p < 0.0005). This analysis shows people less influenced by the opinion expressed by others in the reviews have overall more activity in this area.

The Bayesian model allowed the calculation of how social information influenced participants' initial impressions of value. In the Bayesian framework, the Kullback-Leibler (KL) divergence can quantify the extent to which a prior distribution is updated to form a posterior distribution (Figure 3.4A). Thus, a larger KL divergence indicates a greater preference update. KL divergence will be critical in the fMRI analyses because it provides a combined measure of trial-by-trial update that takes into account both the uncertainty reflected by the participant's confidence rating and the number of reviews (i.e., group size). Letting p and q denote prior and posterior density function respectively, KL divergence is computed as

$$-\int p(x)\log q(x)d(x) + \int p(x)\log p(x)dx \quad (3.3)$$

In the Bayesian model, both prior and posterior distributions were Gaussian distributions. Therefore, the above equation reduces

$$\log\frac{\sigma_{post}}{\sigma_{prior}} + \frac{\sigma_{prior}^2+(\mu_{prior}-\mu_{post})^2}{2\sigma_{post}^2} - \frac{1}{2} \quad (3.4)$$

where μ_{prior} and μ_{post} are the prior and posterior means and σ_{prior}^2 and σ_{post}^2 prior and posterior variances.

The Bayesian model enables a key analysis, namely the identification of brain areas that track the magnitude of Bayesian value update in the presence of social information. A new general linear model (GLM2) was constructed using a parametric regressor that tracked the trial-by-trial KL divergence estimates, using the aforementioned model fits. KL divergence takes into account all aspects of belief change, such as the initial rating and confidence, and the mean and number of Amazon reviews. This analysis found a trial-by-trial response in dmPFC (see Figure 3.4B) to parametric increases in KL divergence ($p < 0.05$ small volume corrected centred on a priori hypothesised coordinates [-3,51,24] from (Hampton et al., 2008)). In other words, activity in this cluster indexes the size of update of a value judgment after the social information provided by Amazon review has been presented.

Another test was performed on whether this same region indexed how likely participants were to conform to the social consensus in general. A between-participant measure was constructed of how resistant participants were to the social information carried by the reviews. Specifically, the degree of resistance to Amazon reviews is 1 when Amazon rating is completely ignored and prior mean is the same as posterior mean. Also, the degree of resistance to Amazon reviews is 0 when prior is completely discarded and Amazon mean is the same as posterior mean. A larger value indicates that Amazon mean is more heavily weighted toward posterior mean than prior mean is (see Materials and methods section above for more details). The BOLD signal in this region of interest (8-mm sphere centred at the peak of the effect isolated from the independent within-participant analysis GLM2) was extracted and tested whether the activity in this region showed a positive modulation by the individual ability to resist to the social information (carried by the reviews showed by Amazon website) while constructing their value judgments. This analysis showed that activity in this cluster of dmPFC

(Figure 3.4B) was higher for those individuals who were less influenced by the information carried by the reviews of other people ($r = 0.77$, $p < 0.0005$). This between-participant analysis and the preceding trial-by-trial within-participant analysis provide complementary viewpoints on dmPFC's role in belief updating.

3.4 Discussion

In this study, it was shown that the degree by which value and confidence judgments is influenced by the opinion of others (expressed through online reviews) is modulated by both the reliability of the group's opinions and the individual's confidence in their own prior belief. It was found that people's updated judgments were consistent with a Bayesian integration account that updated more toward the group consensus when initial confidence was low and the group is large. The model was verified by eliciting liking and confidence judgments twice: the first time when each item was presented in isolation and a second time when it was presented together with the reviews collected from the Amazon website. At the behavioural level, the number of reviews significantly modulated the shift toward the group consensus (i.e. toward the mean of the Amazon's reviews). This shift was more substantial when the participants were less sure in their initial ratings (low level of confidence) and a large shift towards the group consensus was characterised by a drop in the overall level of confidence. These results showed that uncertainty in both the social information and participants' initial estimates (gauged through confidence reports) modulated the participants' behavioural responses. This result confirms the suspicion from the delegation study, that the exact form of uncertainty representation will have an impact on final choice outcome (i.e., judgment).

To help quantify the impact of the social information on the computation of value and confidence, a simple Bayesian model was constructed that captured the main aspects of the behavioural results. Although not fitted to the confidence data, the model correctly predicted confidence as evidenced by a positive correlation between the precision of its posterior distributions with confidence collected during the scanning phase. This finding is consistent with the idea that verbal reports of confidence closely match the formal concept of precision as defined in Bayesian probability (Meyniel, Schlunegger, & Dehaene, 2015; Meyniel, Sigman, & Mainen, 2015), although see also (Pouget, Drugowitsch, & Kepecs, 2016).

Analysis of the fMRI data showed that mPFC/vmPFC tracked both the subjective rating as well as the confidence level in that estimate. This work adds to recent studies that have

considered the role these areas play in representing confidence during value-based choice. For example, De Martino and colleagues have shown that activity in vmPFC correlates with both difference in value and confidence in a binary choice task (De Martino et al., 2013). The study presented here provides a strong test of this characterisation of the vmPFC because participants judged objects in isolation rather than in a binary choice task, which resulted in rating and confidence sharing a quadratic as opposed to linear relationship (i.e. confidence is highest for extreme ratings). This result was also supported by the behavioural analyses. Regression analyses that use orthogonalisation could also be used to disentangle the independent contributions of confidence and value to the mPFC signal (Mumford, Poline, & Poldrack, 2015). Such an analysis could reveal whether mPFC represents value over and above its relationship with confidence. The ventral-dorsal gradient found in this study suggests the effect is small but further studies will need to be optimized to answer this research question. Nevertheless, it was found that vmPFC tracked both the participants' confidence and liking ratings. These finding are in line with a recent study by Lebreton and colleagues that found that activity in mPFC/vmPFC correlates with both the linear and quadratic expansion of the pleasantness ratings that might reflect an automatic assessment of confidence (Lebreton et al., 2015).

The current study helped resolve the relationship between confidence and value representations in the PFC by finding a smooth gradient along the medial ventral-dorsal axis of PFC with liking ratings manifested more ventrally and confidence ratings more dorsally. A possible interpretation of this result is that there are two populations of neurons, distributed along the ventral-dorsal axis of medial prefrontal cortex, with the more ventral region coding for the mean value estimate and the more dorsal region coding for the reliability of these estimates (either measured directly by confidence ratings, or indirectly through the quadratic expansion of liking rating). A similar gradient has been found for values that are executed (represented more ventrally) and values that are modelled but not executed (represented more dorsally) (Nicolle et al., 2012). An intriguing possibility is the more dorsal part of the PFC is implicated in a high-order belief inference (Yoshida & Ishii, 2006) for monitoring the reliability of the behavioural strategy in which the agent is currently engaged (Donoso et al., 2014) as in value estimation in this current study. Such inferences may tap similar processes with those used to reason about other people's states, which is also hypothesised to involve the more dorsal regions of PFC (Denny, Kober, Wager, & Ochsner, 2012).

The modeling approach presented here quantified the degree of value update resulting from exposure to the social information carried by the reviews on a trial-by-trial basis. In this model, the Kullback-Leibler (KL) divergence indexes the shift from the prior to posterior belief when new evidence (i.e. likelihood) is available. The model-based fMRI analysis showed that activity in dorsomedial prefrontal cortex (dmPFC) positively correlated with KL divergence. It was also found that dmPFC responded at the trial-by-trial level to the size of update in value judgment from the prior judgments (made in absence of social information) to posterior judgments after the participants were exposed to other people's opinions (expressed at the aggregate level by the reviews). Recent work using a perceptual decision-making task also found that activity in dmPFC (though slightly more posterior to the peak of the main activation in the current study) co-varied with belief updating in response to new information (O'Reilly et al., 2013).

Although there are known effects of social conformity, it is possible that the findings presented here could be solely relevant to the domain of value-based decision-making. For example, if a deep neural network were the one that created the Amazon ratings, the results possibly would not have differed too much from the ones reported here. Nonetheless, earlier work implicates the dmPFC in theory of mind and in social cognition more generally (Amodio & Frith, 2006; Behrens et al., 2009), through enabling agents to take into account the judgments of others during value-based choice (Behrens et al., 2009; Behrens, Hunt, Woolrich, & Rushworth, 2008; Coricelli & Nagel, 2009; Hampton et al., 2008). Although these studies focused on the dmPFC, related studies have found a role for other brain regions in the social modulation of learning and hedonic experience. For example, the rostral cingulate cortex and striatum have been found to track the mismatch between the opinions of an individual and a group (Klucharev et al., 2009). This basic mismatch is analogous to deviating from the group in the current study absent weighting by the reliability of the individual and group information sources. A second fMRI study investigated how teenagers were influenced by popularity ratings in judging song tracks (Berns et al., 2010). Their analyses (using a masking procedure) focused on a network of regions (including insula) that were activated during hedonic experience (i.e. listening to the song track), which can be contrasted with the more abstract evaluation processes invoked by the task presented here.

From a computational perspective, internal models should be updated when new information (or a change in the task) makes the current model inadequate (Domenech &

Koechlin, 2015; Durstewitz, Vittoz, Floresco, & Seamans, 2010) (Durstewitz et al., 2010; Domenech and Koechlin, 2015). This shift usually pushes the agent towards more explorative behaviours (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Hayden, Pearson, & Platt, 2011; O'Reilly et al., 2013). Other studies have shown that activity in dmPFC tends to increase in those situations in which an agent has to abandon the current model (because it has become unreliable) and initiate exploration (O'Reilly et al., 2013; Tervo et al., 2014). One possibility is that the update is triggered by noradrenaline (Yu & Dayan, 2005) that signals a mismatch between the predictions of the current internal model and external feedback (McGuire, Nassar, Gold, & Kable, 2014; Payzan-LeNestour et al., 2013; Yu & Dayan, 2005). A recent study has provided experimental support for this idea by showing that noradrenaline mediates this switch by changing the noradrenergic inputs to the anterior cingulate cortex (Tervo et al., 2014).

The results presented here suggest that dmPFC involves a higher-order inference similar to that required when estimating the reliability in one's own appraisals of value - see also (Nicolle et al., 2012). It is possible that in most social interactions humans are required to represent others' preferences (an ability linked to theory of mind) and that this information is used to update their own preferences. An intriguing possibility is that the basic computation of dmPFC is to represent and manipulate multiple beliefs hence its prominent role in theory of mind.

Finally, while at the within-participants level dmPFC activity and KL-divergence positively correlated, at the between-participants level it was found that activity in dmPFC in response to KL divergence was more pronounced for people less amendable to conforming to the group consensus (i.e., adjusting their ratings toward the group's ratings). This result is consistent with dmPFC playing a role in monitoring differences between an individual's opinion and that of the group. Greater dmPFC involvement overall appears to indicate heightened sensitivity to divergence with the group, which may facilitate an individual maintaining their original opinion to a greater extent. In contrast, a person who readily conforms to the group consensus would not integrate personal beliefs with the group's as much as wholesale accept the group's opinion. In such a case, the dmPFC should not be very active overall, assuming its role is to monitor differences between belief representations. In reality, people should fall along a continuum of conformity, such that dmPFC activity tracks both trial-by-trial updates and the overall propensity to conform. These findings are also in line with two recent TMS studies that found that stimulating posterior medial frontal cortex modulates social

conformity (Klucharev, Munneke, Smidts, & Fernández, 2011) and choice-induced preference changes (Izuma et al., 2015).

In conclusion, the work here suggests that the update of value and confidence in response to social information involves an integration mechanism analogous to that used in perceptual decision-making. The evidence suggests that belief update follows Bayesian principles in which clear signatures of value, confidence, and belief update are reflected in prefrontal cortex activity.

These findings are enlightening with regards to how people integrate personal beliefs and preferences with those of other people. The integration process is only possible because there is some amount of shared preferences amongst people in the first place. For example, if the Amazon reviews showed only uniform preference distributions, the only information that participants would have is the number of reviews which should not shift their initial preference. Although, there is the possibility that participants are more likely to want to obtain that product if there are many reviews, even if the distribution is uniform. This possibility could not be tested with this design. However, it is clear that the form of the distribution, the way in which social preference uncertainty is represented, is important in the construction of a final preference judgment with respect to the retail product in question.

Another research variable that was not tested or controlled for in this study is how the participants made their judgments at the algorithmic level (Marr's second level). Although the focus was on the representational issue regarding preferences and beliefs, this was only informed by a model that was constructed at the computational level. Although it is possible to take these models as a serious algorithmic characterisation too (Love, 2015). The model was largely descriptive given its large number of free parameters, intended for capturing beliefs about products and focusing on how the integration between personal and social preferences works. The exact algorithmic transformations that were carried out by participants to perform their preference integration is not known. The same can be said about the delegation choices in chapter 2. Their choice option representations were informed by assumptions of the task goal made at the computational level (assuming a utility maximisation process for a control-adjusted and overconfidence-adjusted utility function). There are two ways to address this issue; to model the data with algorithmic assumptions or to instruct participants to use a specific algorithmic procedure (i.e., decision strategy) to compute their final choice. The latter option

is what was used for the study presented in the next chapter. As is presented in the next chapter, using instructed strategy use gives the benefit on focusing on issues that are much more specific to choice option representations such as the effects that stimuli formats may have on the implementations of such strategies. Restriction of experimental variables in this way provides more control to address how task goals and stimulus formats interact, displaying properties of the trade-off that choice option representations must handle between the two.

Chapter 4 Choice option representations characterised by the interaction between decision strategy and stimuli format

4.1 Studies on heuristic decision-making

Some decision strategies are slow and information demanding whereas others are "fast and frugal" (Gigerenzer & Gaissmaier, 2011), though for a dissenting opinion on the prevalence of heuristic use see Newell, Weston, and Shanks (2003), and Newell (2005). Consider a scenario in which a child suddenly crosses the street to get his ball. The driver has less than a second to evaluate the situation and decide whether to press hard on the brakes without swerving or press on the brakes and swerve onto the sidewalk. The former option risks hitting the child and the latter option risks hitting other pedestrians. The optimal decision depends on cues such as the speed of the vehicle, the distance from the car to other people, the car's stopping distance, the number of people on the sidewalk, the driver's ability, etc.

What is the best way to integrate all this information quickly? Representing and integrating all this information in an optimal – perhaps Bayesian – manner may be impossible or too time-consuming. The previous chapter presented a study where personal and social preferences could be described by a Bayesian update framework. The integration only needed to be carried out on a very restricted one-dimensional problem since preferences were represented as distributions in one dimension. In high dimensions (as in the example above), Bayesian inference can quickly become intractable and algorithmic approximations are necessary for achieving the required task goals. This study addresses potential candidates for such algorithms and proposes that the best way to study them, in terms of experimental control, is through instructed use. Instructed strategy use (i.e., explaining to participants which algorithm they should use) is a way of controlling representations at the algorithmic level. This enables the study presented here to address how decision algorithms can be hindered or enhanced based on stimulus format, enhancing the difficulties encountered by choice option representations when handling the trade-off between stimuli and task goals.

Alternatives to the computationally intensive decision strategies are commonly referred to as "fast and frugal" heuristics (Todd & Gigerenzer, 2000). Heuristics are fast in that they can be applied quickly and frugal in the sense that they use less information to make a decision than more complex procedures that selectively weigh all information sources. Despite the fact

that heuristics use less information from the environment, in practice they can perform very well, often surpassing regression approaches for certain decision problems (Jean Czerlinski, Gigerenzer, & Goldstein, 1999). Heuristics have been described as "efficient cognitive processes, conscious or unconscious, that ignore part of the information" (Gigerenzer & Gaissmaier, 2011) and as employing "a minimum of time, knowledge, and computation to make adaptive choices in real environments." (Todd & Gigerenzer, 2000). This characterisation of heuristics differs from earlier accounts that cast heuristics as imperfect approximations of rational decision procedures (Tversky & Kahneman, 1974).

When complex decision strategies, such as multiple linear regression, cannot be implemented because of resource constraints, such as lack of time, people might use alternative strategies like heuristics (Todd & Gigerenzer, 2000). In this study, two popular heuristics were tested to see if they differ from one another in cognitive processing requirements as reflected by timing constraints. Specifically, two popular heuristics were compared, Tallying (TAL) and Take-the-Best (TTB), which are introduced by way of example below. These heuristics follow from previous work, such as the lexicographic heuristics that only take into account the most discriminating attribute value (Fishburn, 1967; Tversky, 1969), the majority of confirming dimensions heuristic (Russo & Dosher, 1983), and the equal weights strategy (Dawes, 1979).

Suppose one wants to predict whether China or India will have higher gross domestic product (GDP) growth based on their productive capabilities, natural resource wealth, and the diversity in their exports, etc. The TAL heuristic chooses the country that bests the other across the most measures. TAL does not selectively weigh cues as linear regression does, but instead merely counts the number of cues favoring one alternative over the other (Gigerenzer & Gaissmaier, 2011). On the other hand, the TTB heuristic chooses based on the most predictive cue and only considers the next best cue when there is a tie. TTB implies that cues are rank ordered in terms of their predictive validity in determining the criterion (e.g., in predicting China or India). TTB sequentially searches until a discriminating cue is found and, thus, may reach a decision after only considering the first best cue. These two heuristics can both be effective in practice but can differ in their choices as shown in an example trial of the first experiment in Figure 4.1.

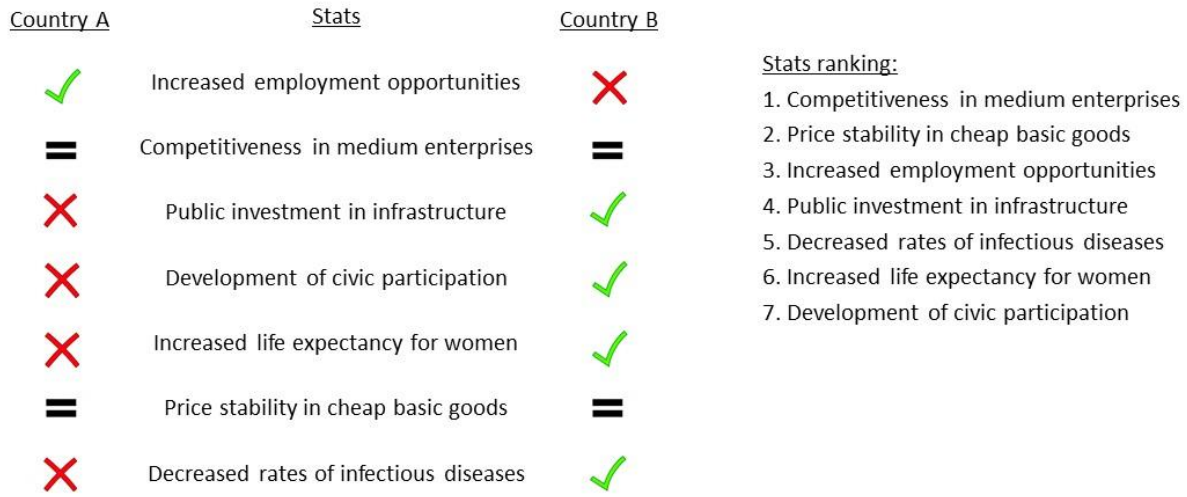

Figure 4.1. Example trial for the practice phase of Heuristic Study 1.

Participants were required to choose the country that would have higher gross domestic product (GDP) for the following year depending on the values of the economic statistics presented. Participants were assigned either to the Tallying (TAL) or the Take-the-Best (TTB) condition and asked to respond according to what the respective heuristic would predict. In this example trial, TAL would choose Country B since four cues have superior values (green checkmarks) whereas Country A is only superior on one cue. In contrast, TTB would choose Country A since the value for the best-discriminating cue, which in this case is the third most predictive cue (i.e., "Increased employment opportunities"), is superior to Country B.

The hypothesis is that these two heuristics differ in their cognitive processing requirements such that what is fast and what is frugal is contingent on the cognitive processes invoked by the environment. In Heuristic Study 1, it is predicted that TTB will fare worse under time pressure than TAL, whereas the opposite pattern is predicted in Heuristic Study 2. Although one might expect TTB to be fast given that it samples very little information (Bröder & Gaissmaier, 2007; Khader et al., 2011), the cognitive demands in Heuristic Study 1 should be high for TTB users because the stimulus format invites an effortful sequential search procedure. In contrast, the stimulus format in Heuristic Study 2 reduces this search burden while making it more difficult for TAL users to perform rapid summation operations.

The experimental procedures are intended to expand the scope of inquiry by deviating from the original formulation of TTB (Gigerenzer & Todd, 1999) in which decisions were made from memory for environments; instead the method here invites participants to make inference from givens in the environment. The distinction between inference from memory and inference from givens is interesting for the study of choice option representations but inference from memory makes it harder to study the stimuli formats as was intended here. Although memory demands are an important aspect of heuristic application, Heuristic Studies 1 and 2 focus more on the attentional demands of applying heuristics. Related previous efforts have noted that one subtle complexity of TTB is that it requires a hierarchy of cue validities (Dougherty, Franco-Watkins, & Thomas, 2008; Juslin & Persson, 2002) and that non-compensatory strategies such as TAL can be affected by cue salience (Platzer & Bröder, 2012).

When compared to algorithms that are more computationally intensive such as linear regression, TAL and TTB have various algorithmic aspects in common. Their most salient similarity is that they both disregard covariance structure among cues (Parpart, Jones, & Love, 2017). Both heuristics also disregard relative cue weight magnitudes.

Despite these similarities, these two heuristics may differ in their cognitive demands. For instance, TTB implements a search for the best discriminating cue which has been argued to take time (Bröder & Gaissmaier, 2007) and a certain level of cognitive control due to selective attention to the relevant cue, as well as inhibition of the irrelevant ones. Such sequential control processes are thought to be effortful, serial in nature, and time-consuming (Posner & Presti, 1987). In effect, TTB has the prescription that people will embark on a serial search, which in the first experiment is a visually guided search. Such visually guided searches are a common domain for research on top-down attentional control mechanisms (Mozer & Baldwin, 2008; Wolfe, Cave, & Franzel, 1989).

TAL's cognitive requirements may be quite different than those used for TTB. Furthermore, choice option representations may react differently to stimulus properties depending on the heuristic that is being used. It is argued that TAL requires people's ability to do quick summations over stimuli. Related research on numerosity has shown that people can be very fast at doing these types of operations without explicit counting (Feigenson, Dehaene, & Spelke, 2004). Given the right representation format of cues, a TAL decision problem could be reduced to a low-level perceptual categorisation problem (e.g., Palmer, Huk, & Shadlen,

2005). However, if the representation format is not suitable for such operations, as manipulated in Heuristic Study 2, the prediction is that TAL's performance should be reduced.

In Heuristic Study 1, it is predicted that compliance with the TAL heuristic would be higher than for TTB under time pressure conditions even though TAL considers more aspects of the stimulus. In Heuristic Study 2, this effect is reversed by eliminating search costs for TTB and altering the stimulus format in a manner that obstructs quick summation operations that favor TAL performance. Together, these two studies aim to establish that heuristics should be understood not only in terms of how they perform in various information environments, but also in terms of the cognitive processes they engage.

4.2 Heuristic Study 1

In Heuristic Study 1, the hypothesis tested is that the TTB heuristic will not always lead to rapid decisions because of the search costs and attentional control it can demand. For certain stimulus formats, it should be possible for TAL to be faster than TTB. For example, colour-coded stimulus values should allow for rapid perceptual integration of cue values, making TAL faster than TTB. Furthermore, TAL use should be unaffected by the randomisation of cue position across trials because TAL treats all cues identically and summates. In contrast, TTB's search requirements should be increased by this randomisation and should not benefit from the colour-coding. Thus, TAL should be faster than TTB under these conditions because of the basic resource requirements of each heuristic.

4.2.1 Methods

4.2.1.1 Participants

Participants (206 total, 95 female) were recruited on Amazon Mechanical Turk, an online study platform commonly used in psychological studies with good results (Crump, McDonnell, & Gureckis, 2013). Participants were restricted to the United States of America and assigned either to the TAL condition (107 total, 58 female) or the TTB condition (99 total, 39 female). They received $2.50 to complete a 40 min. (approx.) learning and decision-making task and the best participant was offered a $20 bonus in each condition. The average age of participants was 37.9 years (s.d. = 12.56). The study was approved by the local UCL ethics committee.

4.2.1.2 Design and materials

As a between-participants manipulation, participants were explicitly instructed to use either TAL or TTB in a two-alternative forced choice task (i.e., choosing which country will have higher GDP). Each participant first completed a practice phase (72 trials) followed by a test phase, consisting of two blocks of trials. Whether the test block was self-paced (72 trials) or speeded (72 trials) was counterbalanced across trials. Trial order was randomised separately for each participant.

Within participants, the same 72 trials were used across these three (practice, self-paced test, speeded test) experimental segments. Trials were designed such that one response was consistent with TAL and the other with TTB. In other words, the heuristics disagreed on every trial, which allowed for discriminating heuristic use. Perfect performance is achievable for both heuristics because compliance was measured with an instructed heuristic (compliance with heuristic is understood as % correct). Conversely, expected performance under random responding was at chance (50%). The presentation order of the cues was randomised on a trial-by-trial basis.

The 72 trials for each experimental segment consisted of three trial types in terms of three difficulty levels. The 72 trials were equally partitioned into the three trial types resulting in 24 trials for each difficulty level. Difficulty was defined differently for each heuristic. For TTB, difficulty level is referred to as Q1 (cue 1), Q2 (cue 2) and Q3 (cue 3), in order of ascending difficulty. Q1 trials represent trials where retrieving the value for the best cue was sufficient to respond in compliance with TTB. Similarly, Q2 trials require retrieving the value for the second-best cue and Q3 trials require retrieving the value for the third-best cue. For TAL, difficulty level is referred to as Δ3 (delta 3), Δ2 (delta 2), and Δ1 (delta 1), in order of ascending difficulty. The Δ3 trial types represented trials where one option was a better choice by a difference of three cue values, Δ2 trials represented trials where one option was a better choice by a difference of two cue values, and the Δ1 trial types represented trials where one option was a better choice by a difference of only one cue value. Controlling for difficulty level in each heuristic was also a way to verify that indeed participants were still trying to implement the respective heuristic in each experimental condition. For both the TAL and TTB condition, trials were randomly sampled from the same trial space (see Table B.1 and Table B.2 in Appendix B for more information on the exact sampling procedure).

The seven cues shown on each trial were economic statistics that would predict whether a developing country would achieve higher Gross Domestic Product (GDP) levels the following year when compared to another developing country (see Appendix B.1 for a list of the statistics). These statistics were artificially created and do not correspond to any real-world data in keeping with the focus on heuristic compliance as opposed to real-world performance. This domain was chosen in the hopes that it would be familiar in a general sense while also unlikely to invite prior knowledge to guide decisions (i.e., not strongly interfere with the instructed heuristic use).

4.2.1.3 Procedure

Participants were shown the seven cues on each trial and asked to choose which country would have higher GDP the following year. The two options were "Country A" or "Country B". Participants saw seven economic statistics with a value for each country on each trial (see Figure 4.1 for an example of stimulus presentation). Each statistic was framed as a comparison between two options (i.e., two developing countries) where a green checkmark was presented if superior to the other option, a red cross was presented if inferior to the other option, and a black equal's sign was presented for ties between countries on a given statistic. A list which ranked the statistics in order of importance, with randomised order between participants, was presented on every trial in the practice phase but not in the test phase. Before starting the practice phase, participants were provided with detailed instructions on how to use the corresponding heuristic for the condition they had been assigned to (either TAL or TTB).

In the practice phase, the participant would see the seven statistics with a value for each country as well as the list which ranked the statistics in order of importance. After making a response, immediate feedback was provided on the screen. If the participant responded in accordance with the heuristic, the screen would show "Good! You understood the rule!", in addition to an explanation of why the choice was correct according to the heuristic. If the participant did not respond in accordance with the heuristic, the screen would show "Bad. You did not understand the rule.", in addition to an explanation of why the choice was incorrect according to the heuristic. The statistics with the values for each country were left on screen during feedback but the list of ranked statistics was not. The list was not presented during feedback to discourage reliance on the list and incentivise better engagement with the task. Participants were not directly questioned about their knowledge of the cue validities. The

feedback was presented until the participant decided to move on to the next trial followed by a presentation of a blank white screen for 500 milliseconds.

Subsequently, the participant would enter either a block with time pressure or a block without time pressure of the test phase. The list which ranked the statistics in order of importance was removed for the test phase. For the block without time pressure, the participant was asked to answer in accordance with the heuristic that had been practiced on previously, only this time without feedback. The inter-trial interval was 1500 milliseconds with the first second showing "Thank You!" followed by a presentation of a blank white screen for 500 milliseconds. For the block with time pressure, the participant was also asked to answer in accordance with the heuristic practiced on previously and without feedback. Then, instructions stating that the participant only had 2000 milliseconds to respond on each trial were provided. The inter-trial interval was also 1500 milliseconds with the first second showing "Thank You!" followed by a presentation of a blank white screen for 500 milliseconds except when the participant reached the 2000 milliseconds deadline. If the participant reached the 2000 milliseconds deadline, the screen would explain why it was important for them to adhere to the imposed deadline. This screen was presented for 10000 milliseconds followed by a presentation of a blank white screen for 500 milliseconds.

4.2.2 Results

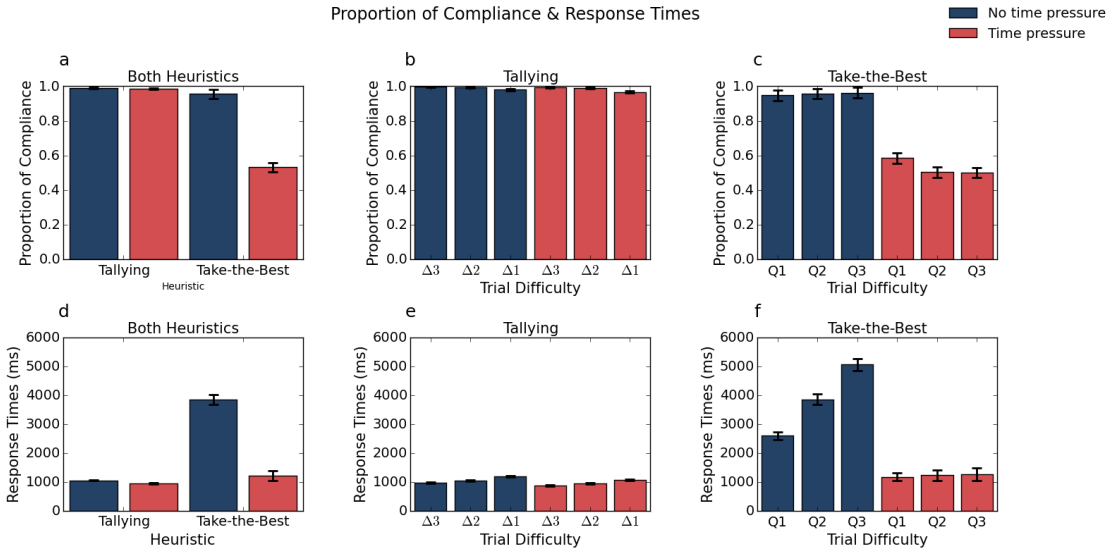

Figure 4.2 ***(previous page)*****. Main results after applying exclusion criteria for Heuristic Study 1 (*n* = 179).**

The figure shows the proportion of compliance with (a) both heuristics for blocks without time pressure (blue bars) and blocks with time pressure (green bars), (b) the proportion of compliance in the TAL condition for blocks with and without time pressure displayed by degrees of difficulty (Deltas), and (c) the proportion of compliance in the TTB condition for blocks with and without time pressure displayed by degrees of difficulty (Qs). Shows the response times for (d) both heuristics in both blocks with and without time pressure, (e) the response times in the TAL condition for blocks with and without time pressure displayed by degrees of difficulty (Deltas), and (f) the response times in the TTB condition for blocks with and without time pressure displayed by degrees of difficulty (Qs). For all panels, error bars are 95% within-participants confidence intervals.

4.2.2.1 Exclusion criteria

Exclusion criteria were based on the responses made on the second half of the practice phase (36 trials) and number of missed responses when under time pressure. Participants with performance under 90% in the second half of the practice phase or over 16 missed responses when under time pressure were considered outliers and excluded from all subsequent analysis (26 in the TTB condition, 1 in the TAL condition). Including the exclusion criteria from the practice phase, this resulted in a total of 26.26% of participants who were excluded from all further analyses in the TTB condition and 1% of participants excluded from the TAL condition. Analyses were also run without any exclusion[5] and this did not change any conclusions from the analyses shown below. All analyses that follow use people who passed exclusion (n = 179) to provide a more stringent evaluation of the predictions. All response time analyses were calculated with median response times. For a presentation of results from the practice phase, please refer to Appendix B.4.

4.2.2.2 Test phase

TTB participants had lower compliance than TAL, mostly because they were affected more by the time pressure manipulation (see panel a in Figure 4.2). Proportion of compliance was analysed using a 2x2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB), a within-participants factor of time pressure (present or absent), and a within-participants factor of trial difficulty (3 levels of trial difficulty). Main effects were observed for heuristic, $F_{1, 177} = 285.89$, $p < 0.001$, $\eta^2 = 0.62$, time pressure, $F_{1, 177} = 352.62$, p

[5] All analyses were also conducted with participants that had 100% performance in the second half of the practice phase (95 participants in the TAL condition and 25 in the TTB condition) and this did not change any conclusions presented here.

< 0.001, $\eta^2 = 0.67$, and trial difficulty, $F_{2, 354} = 13.85$, $p < 0.001$, $\eta^2 = 0.07$, as well as an interaction between heuristic and time pressure, $F_{1, 177} = 329.37$, $p < 0.001$, $\eta^2 = 0.65$, an interaction between heuristic and trial difficulty, $F_{2, 354} = 4.53$, $p = 0.011$, $\eta^2 = 0.03$, an interaction between time pressure and trial difficulty, $F_{2, 354} = 18.66$, $p < 0.001$, $\eta^2 = 0.10$, and the three-way interaction between heuristic, time pressure, and trial difficulty, $F_{2, 354} = 14.87$, $p < 0.001$, $\eta^2 = 0.08$. These results directly support the hypothesis by showing that TTB has lower compliance than TAL, especially under time pressure. The three-way interaction highlights the asymmetric effect of the time pressure manipulation on both heuristics' difficulty levels.

TTB participants responded more slowly than TAL participants and were markedly slower on difficulty trials (see panels d, e, and f in Figure 4.2). Response times were analysed using a 2x2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB), a within-participants factor of time pressure (present or absent) and with a within-participants factor of trial difficulty (3 levels of trial difficulty). Individual response times were calculated as median response times. Main effects were observed for heuristic, $F_{1, 177} = 382.41$, $p < 0.001$, $\eta^2 = 0.68$, time pressure, $F_{1, 177} = 370.92$, $p < 0.001$, $\eta^2 = 0.68$, and trial difficulty, $F_{2, 354} = 470.60$, $p < 0.001$, $\eta^2 = 0.73$, as well as an interaction between heuristic and time pressure, $F_{1, 177} = 313.50$, $p < 0.001$, $\eta^2 = 0.64$, an interaction between heuristic and trial difficulty, $F_{2, 354} = 248.39$, $p < 0.001$, $\eta^2 = 0.58$, an interaction between time pressure and trial difficulty, $F_{2, 354} = 285.34$, $p < 0.001$, $\eta^2 = 0.62$, and the three-way interaction between heuristic, time pressure, and trial difficulty, $F_{2, 354} = 265.66$, $p < 0.001$, $\eta^2 = 0.60$. The main effect of heuristic shows that TTB is a slower heuristic than TAL and directly supports the attentional control hypothesis. The interaction between heuristic and time pressure suggests that TAL can accommodate to time pressure more readily than TTB can. The three-way interaction shows that trial difficulty still shows differences between heuristics in both blocks, with and without time pressure. This gives reassurance that participants were still engaged and attempting to implement the respective heuristic even for blocks with time pressure.

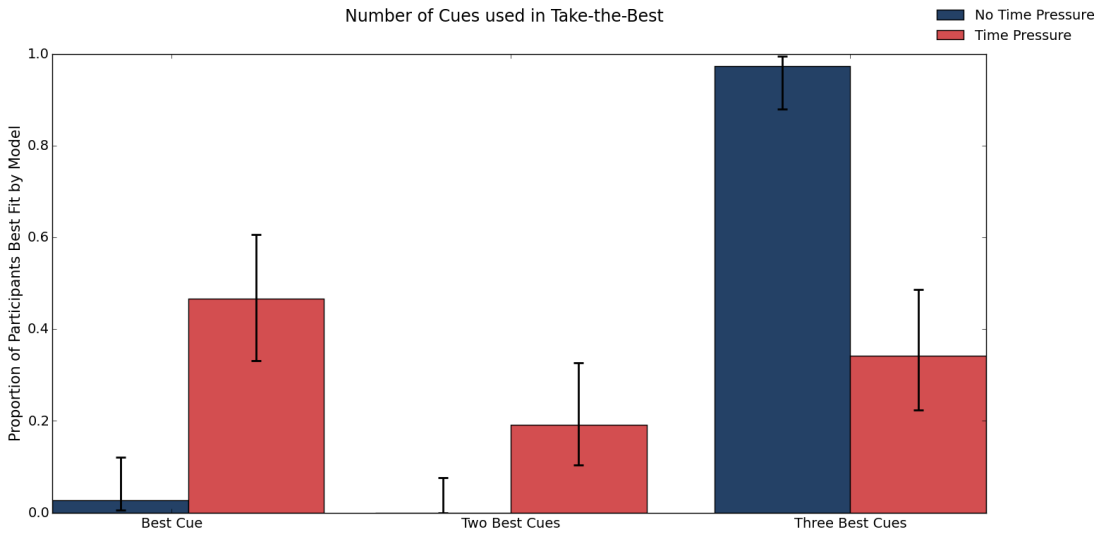

Figure 4.3. Model-based analyses reveal cue usage for Take-the-Best (TTB) for Heuristic Study 1.

The differences in how many cues were used in blocks without time pressure (blue bars) and with time pressure (red bars) for the TTB condition. The vertical axis shows the proportion of participants that were best described by one of the three models. Absent time pressure (blue bars), most participants' decisions were best fit by the full model using the best three cues, whereas with time pressure (red bars) some participants appeared to rely on fewer cues. Error bars are 95% confidence intervals.

4.2.2.3 Model-based analysis

For the TTB participants, three models were fit to determine how many cues participants tended to successfully incorporate under self-paced and speeded conditions. The models are reduced versions of TTB. Only three models were considered; TTB first-cue (only considers the first cue presented on all trials), TTB first-two-cues (only considers the first and second cue presented on all trials), and TTB first-three-cues (only considers the first, second and third cue presented on all trials). As shown in Figure 4.3, the number of cues successfully incorporated tended to be higher under self-paced conditions, suggesting that the TTB procedure broke down under time pressure. Fisher's exact test found that the model contingencies differed across self-paced and speeded conditions ($p < 0.001$).

Only models with three or fewer cues were considered because only the first three cues were needed to comply with TTB in the study design. Which model best fit each participant was determined by comparing the agreement in choices between human and model. For the one and two cues models, agreement was scored as 0.5 (i.e., the expectation for a random guess)

for trials that exceeded the cues encoded by the model because TTB guesses when there is no discriminating cue. All three models are parameter-free and, thus, are readily comparable. For the speeded blocks, the agreement in choices between human and model were significantly above chance (0.5) for the best cue model, $t_{72} = 2.81$, $p = 0.006$, but not for the two best cues model, $t_{72} = 1.68$, $p = 0.097$, or three best cues model, $t_{72} = 1.21$, $p = 0.229$. Additional model fits confirmed that participants were using the cues themselves, as opposed to their positions (see Figure B.1 of Appendix B).

4.2.3 Discussion

Heuristic Study 1 strongly supported the a priori attentional control hypothesis: TTB, however frugal, can be slow given its requirements with respect to attentional control and search costs. This can be seen by the lower proportion of compliance of the TTB heuristic under time pressure, whereas the TAL heuristic maintains the same proportion of compliance for both the self-paced and time pressure conditions. This effect could be due to both the trial-by-trial randomisation of the cue positions and the format of the values shown (e.g. red crosses and green checkmarks). These results highlight the importance of considering not only the structure of the environment, but also the cognitive demands of heuristics when pairing heuristics with tasks. Heuristic Study 2 attempts to reverse the TAL advantage observed in Heuristic Study 1 by using a stimulus format that favors TTB by reducing search costs (helpful for TTB) and the possibility of perceptual summation (harmful for TAL).

4.3 Heuristic Study 2

In Heuristic Study 1, TAL's advantage over TTB was striking – the less frugal (in terms of cues) heuristic was faster. This pattern of results was predicted by considering the cognitive processes each heuristic requires. In Heuristic Study 2, this result was built on by creating conditions that should favor TTB over TAL. In Heuristic Study 2, the search costs for TTB are reduced by ordering stimulus cues by their validity. Unlike Heuristic Study 1, the cues are no longer colour-coded, which prevents TAL from utilizing perceptual summation operations. Mirroring Heuristic Study 1, the format choice in Heuristic Study 2 should strongly favor TAL over TTB.

4.3.1 Method

4.3.1.1 Participants

Participants (194 total, 128 female) were also recruited on Amazon Mechanical Turk. Participants were restricted to the United States of America and assigned either to the TAL condition (97 total, 63 female) or the TTB condition (97 total, 65 female). They received $2.50 to complete a 40 min. (approx.) learning and decision-making task and the best participant was offered a $20 bonus in each condition. The average age of participants was 41.5 years (s.d. = 12.34). The study was approved by the local UCL ethics committee.

4.3.1.2 Design and materials

The design and materials for the experiment were similar to Heuristic Study 1 with some critical differences. Rather than order cues randomly on each trial, cues were always ordered by their validity. Although the stimuli consisted of the same economic statistics as in Heuristic Study 1, they were no longer colour-coded (see Figure 4.4). Instead, pairs of positive and negative adjectives (e.g., better and worse) were used (see Appendix B.1 for a list of the adjectives). The word "equal" was presented for ties.

Country A	Stats	Country B
lesser	Increased employment opportunities	greater
higher	Competitiveness in medium enterprises	lower
larger	Public investment in infrastructure	smaller
stronger	Development of civic participation	weaker
inferior	Increased life expectancy for women	superior
less stable	Price stability in cheap basic goods	more stable
better	Decreased rates of infectious diseases	worse

Figure 4.4 *(previous page)*. Example trial for the test phase of Heuristic Study 2.

Participants were required to choose the country that would have higher GDP for the following year depending on the values of the economic statistics presented. Participants were assigned either to the Tallying (TAL) or the Take-the-Best (TTB) condition and asked to respond according to what the respective heuristic would predict. In this example trial, TAL would choose Country A since four cues have superior values (positive adjectives) whereas Country B is only superior on three cues. In contrast, TTB would choose Country B since the value for the best-discriminating cue, which in this case is also the most predictive cue (i.e., "Increased employment opportunities"), is superior to Country A.

4.3.1.3 Procedure

As in Heuristic Study 1, participants were shown the seven cues on each trial and asked to choose which country ("Country A" or "Country B") would have higher GDP the following year. Participants saw seven economic statistics with a value for each country on each trial (see Figure 4.4 for an example of stimulus presentation).

4.3.2 Results

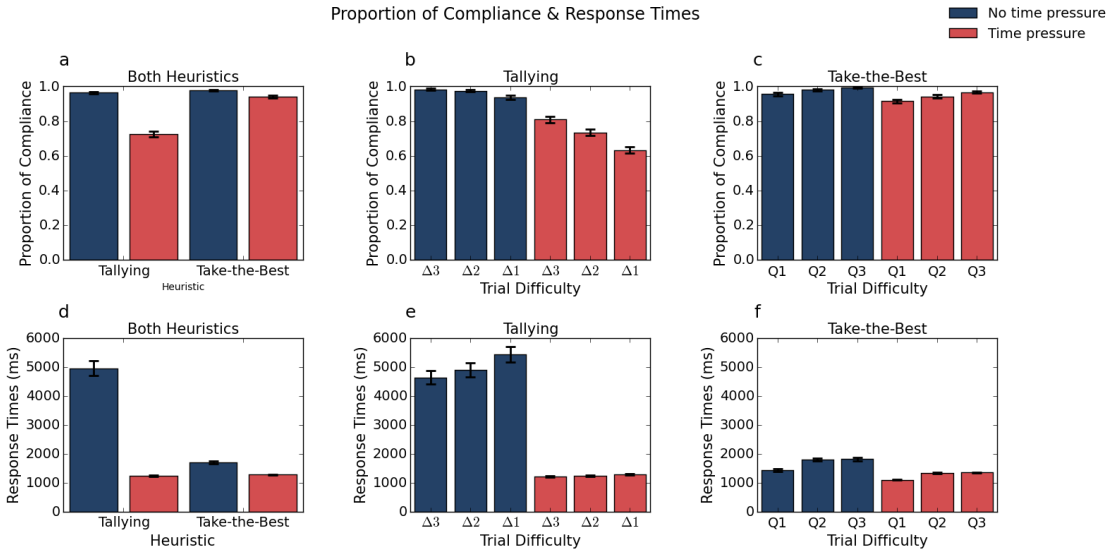

Figure 4.5. Main results after applying exclusion criteria for Heuristic Study 2 (*n* = 173).

The figure shows the proportion of compliance with (a) both heuristics for blocks without time pressure (blue bars) and blocks with time pressure (green bars), (b) the proportion of compliance in the TAL condition for blocks with and without time pressure displayed by degrees of difficulty (Deltas), and (c) the proportion of compliance in the TTB condition for blocks with and without time pressure displayed by degrees of difficulty (Qs). Shows the response times for (d) both heuristics in both blocks with and without time pressure, (e) the response times in the TAL condition for blocks with and without time pressure displayed by degrees of difficulty (Deltas), and (f)

the response times in the TTB condition for blocks with and without time pressure displayed by degrees of difficulty (Qs). For all panels, error bars are 95% within-participants confidence intervals.

4.3.2.1 Exclusion criteria

Exclusion criteria were the same as in Heuristic Study 1. Participants with performance under 90% in the second half of the practice phase or over 16 missed responses when under time pressure were considered outliers and excluded from all subsequent analysis (8 in the TTB condition, 13 in the TAL condition). Including the exclusion criteria from the practice phase, this resulted in a total of 8.25% of participants who were excluded from all further analyses in the TTB condition and 13.4% of participants excluded from the TAL condition. Analyses were also run without any exclusion[6] and this did not change any conclusions from the analyses shown below. All analyses that follow use people who passed exclusion (n = 173) to provide a more stringent evaluation of the predictions. All response time analyses were calculated with median response times.

4.3.2.2 Test phase

TAL participants had lower compliance than TTB, mostly because they were affected more by the time pressure manipulation (see panel a in Figure 4.5). Proportion of compliance was analysed using a 2x2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB), a within-participants factor of time pressure (present or absent), and a within-participants factor of trial difficulty (3 levels of trial difficulty). Main effects were observed for heuristic, $F_{1, 171} = 134.83$, $p < 0.001$, $\eta^2 = 0.44$, time pressure, $F_{1, 171} = 244.74$, $p < 0.001$, $\eta^2 = 0.59$, and trial difficulty, $F_{2, 342} = 21.51$, $p < 0.001$, $\eta^2 = 0.11$, as well as an interaction between heuristic and time pressure, $F_{1, 171} = 135.95$, $p < 0.001$, $\eta^2 = 0.44$, an interaction between heuristic and trial difficulty, $F_{2, 342} = 109.75$, $p = 0.011$, $\eta^2 = 0.39$, an interaction between time pressure and trial difficulty, $F_{2, 342} = 19.89$, $p < 0.001$, $\eta^2 = 0.10$, and the three-way interaction between heuristic, time pressure, and trial difficulty, $F_{2, 354} = 29.42$, $p < 0.001$, $\eta^2 = 0.15$. These results directly support the hypothesis by showing that when search costs are eliminated and the representation of stimuli values obstruct quick summations, then TAL has lower compliance than TTB, especially under time pressure. The three-way

[6] All analyses were also conducted with participants that had 100% performance in the second half of the practice phase (44 participants in the TAL condition and 48 in the TTB condition) and this did not change any conclusions presented here.

interaction highlights the asymmetric effect of the time pressure manipulation on both heuristics' difficulty levels.

TAL participants responded more slowly than TTB participants and were markedly slower on difficult trials (see panels d, e, and f in Figure 4.5). Response times were analysed using a 2x2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB), a within-participants factor of time pressure (present or absent) and with a within-participants factor of trial difficulty (3 levels of trial difficulty). Individual response times were calculated as median response times. Main effects were observed for heuristic, $F_{1, 171} = 171.41$, $p < 0.001$, $\eta^2 = 0.50$, time pressure, $F_{1, 171} = 291.79$, $p < 0.001$, $\eta^2 = 0.63$, and trial difficulty, $F_{2, 342} = 110.18$, $p < 0.001$, $\eta^2 = 0.39$, as well as an interaction between heuristic and time pressure, $F_{1, 171} = 186.14$, $p < 0.001$, $\eta^2 = 0.52$, an interaction between heuristic and trial difficulty, $F_{2, 342} = 15.80$, $p < 0.001$, $\eta^2 = 0.09$, an interaction between time pressure and trial difficulty, $F_{2, 342} = 40.50$, $p < 0.001$, $\eta^2 = 0.19$, and the three-way interaction between heuristic, time pressure, and trial difficulty, $F_{2, 342} = 22.11$, $p < 0.001$, $\eta^2 = 0.11$. The main effect of heuristic shows that TAL is a slower heuristic than TTB and directly supports the main hypothesis. For this experiment, the interaction between heuristic and time pressure suggests that TTB can accommodate to time pressure more readily than TAL can. The three-way interaction shows that trial difficulty finds differences between heuristics in both blocks, with and without time pressure, which provides reassurance that participants were engaged and attempting to implement the respective heuristic even for blocks with time pressure.

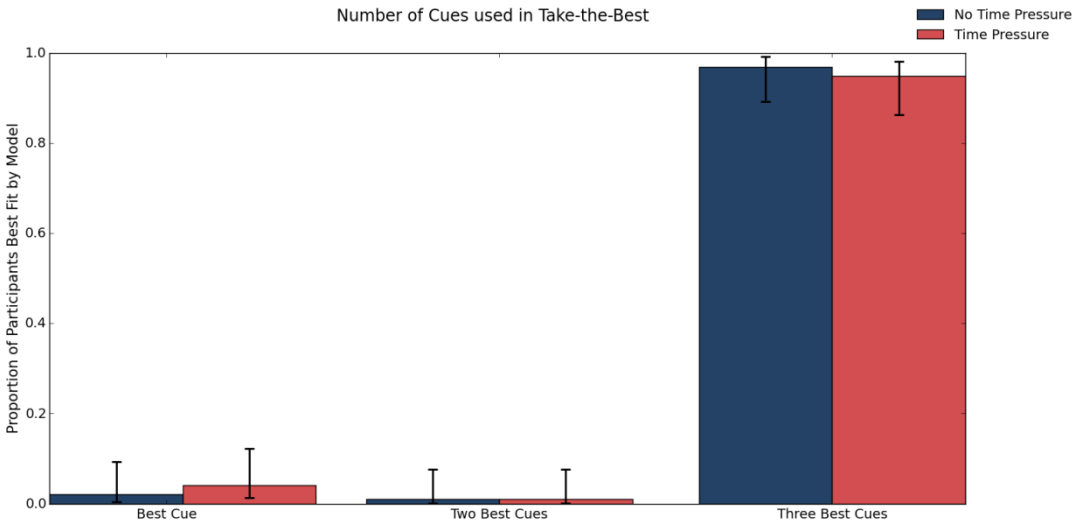

Figure 4.6 *(previous page)*. Model-based analyses reveal cue usage for Take-the-Best (TTB) for Heuristic Study 2.

The differences in how many cues were used in blocks without time pressure (blue bars) and with time pressure (red bars) for the TTB condition. The vertical axis shows the proportion of participants that were best described by one of the three models. Most participants' decisions were best fit by the full model using the best three cues for both speeded and self-paced blocks. Error bars are 95% confidence intervals.

4.3.2.3 Model-based analysis

As in Heuristic Study 1, three models were fit for the TTB condition to determine how many cues participants tended to successfully incorporate under self-paced and speeded conditions. As shown in Figure 4.6, the number of cues successfully incorporated was not different under self-paced conditions, suggesting that the TTB procedure was unaffected by time pressure in this experiment. Fisher's exact test found that the model contingencies did not differ across self-paced and speeded conditions (p = 0.840). Model agreement with human behaviour followed the same rationale as in Heuristic Study 1 (see above).

4.3.3 Discussion

In accord with the attentional control hypothesis, Heuristic Study 2 reversed the effect found in Heuristic Study 1 – TAL was more strongly affected by time pressure than TTB. TAL had lower compliance, particularly under time pressure. One interesting side note is that TTB compliance was better for difficult TTB trials, though these trials were also slower which indicates a speed-accuracy trade-off. Overall, the results from Heuristic Study 2 reinforce the conclusions from Heuristic Study 1, namely that heuristics are not monolithic and need to be decomposed into their constituent cognitive processes to understand and predict their performance under various task conditions.

4.4 Discussion for both heuristic studies

Heuristics are often contrasted collectively with other decision procedures, which could give the impression that heuristics form a uniform class. Instead, it is argued that heuristics are best understood in terms of their constituent cognitive processes. When decomposed into these processes, it becomes possible to predict when people will be successful in applying given the demands of the task. Furthermore, it is possible to predict which stimulus (choice option) representation, such as which serial order or which colour-coding, will be better suited to

achieve the task goals. In this case, the task goal was decision strategy compliance. Effectively, imposing instructed strategy use onto participants radically modifies the natural trade-off between stimuli and task goals handled by choice option representations. For each heuristic study, the trade-off was nullified by forcing the decision algorithm. Interrupting the trade-off did not allow its natural function, which was observed through detriments in performance dependent on stimulus format.

In Heuristic Study 1, stimulus format was chosen to suite the demands of TAL at the expense of TTB. The more frugal heuristic (i.e., requiring less information), TTB, was slower than the less frugal heuristic, TAL. This advantage was hypothesised to be due to the search costs incurred by TTB (cue order was randomised) and the ability of TAL to take advantage of fast perceptual summation operations (cue values were colour-coded). In Heuristic Study 2, these advantages and disadvantages were reversed by ordering cues by their validity and removing colour-coding. As predicted, an advantage of TTB over TAL was observed.

The results mirror previous results using different experimental procedures. As in Heuristic Study 1, previous work finds that the time requirements of TTB are dependent on the number of cues retrieved from memory (Bergert & Nosofsky, 2007; Bröder & Gaissmaier, 2007; Khader et al., 2011; Lee & Cummins, 2004; Newell & Shanks, 2003). These results dovetail with the notion that TTB's implementation involves greater complexity than its algorithmic description suggests (Dougherty et al., 2008; Juslin & Persson, 2002). The findings of Heuristic Study 1 also resemble results of another study where more cue information was processed quicker, as long as the added information would increase coherence among cues (Glöckner and Betsch, 2012). Conversely, participants difficulties with TAL in Heuristic Study 2 mirror previous research exploring when people choose to adopt a compensatory strategy (Bröder, 2003; Bröder & Schiffer, 2006; Platzer & Bröder, 2012; Platzer, Bröder, & Heck, 2014). A surprising result in Heuristic Study 2 was that proportion of compliance for TTB varied with difficulty in the opposite direction of that which was predicted. However, the trade-off observed between speed and accuracy limits the interpretation of this result (e.g. Ratcliff & McKoon, 2008).

The results support the view that heuristics should be unpacked in terms of the psychological processes they rely upon. One important area for continued research is understanding how the representational format of cues and their values affect performance

(Bröder & Schiffer, 2006; Gigerenzer & Hoffrage, 1995). Ideally, this research would expand to consider other heuristics and to conditions under which participants must choose which strategy to adopt for a given environment (Rieskamp & Otto, 2006). One possibility is that participants are adaptive and choose the heuristic procedure that minimises the cognitive demands given the stimulus format. The set of candidate solution procedures could extend beyond heuristic procedures to include gist representations (Peters et al., 2009; Reyna, 2008) and parallel constraint satisfaction approaches (Glockner & Betsch, 2008). Alternatively, there is also recent work suggesting that heuristics can be viewed as a special case of Bayesian inference (Parpart, Jones, & Love, 2017).

In contrast to findings suggesting that participants tend to default to strategies similar to TTB under time pressure (Ben Zur & Breznitz, 1981; Payne, Bettman, & Johnson, 1993), Heuristic Study 1 found that compliance with TTB was dramatically reduced by time pressure, whereas TAL was largely unaffected. As discussed, the choice of stimuli played an important role in determining these results, as observed when the stimuli were changed in Heuristic Study 2. Additionally, previous research has found that people tend to use strategies more like TTB when the decision is largely made from memory (Bröder & Schiffer, 2003). In contrast, people tend to use strategies more like TAL when the stimulus conveys key information about the options (Bröder & Schiffer, 2003). The findings, which follow from the predictions of the attentional hypothesis, suggest a possible explanation for previous results. Likewise, this account may help explain the costs of learning TTB.

An important limitation of these two studies is that they differ from past research in that participants were directly instructed on which strategy to use, either TAL or TTB, and how to use them. For the same reasons as another study (Khader et al., 2011), it was preferable to instruct participants to use these heuristics –instead of studying their spontaneous use– as a preventive measure for strategy switching. This detail provides the experimenters with more control over the task structure but it limits both comparison with other studies that do not employ instructed strategy use and the interpretability of these results for real-world applications. Arguably, the choice of which strategy to use would adapt based on the environment. For example, if TTB needs to engage in sequencing operations then perhaps it would not be the best strategy for that environment. This is our overarching point, that strategies must adapt to informational input and task goal constraints. A real-world example of this could be optimizing TTB for the practice of emergency room responders; if the cues are

preordered for them, and the number is small, then that strategy could prove effective. However, if sequencing operations are involved as tested in Heuristic Study 1, then such a strategy would be detrimental in such a domain.

In conclusion, more frugal heuristics will not necessarily be faster to implement than less frugal ones. Similarly, less frugal strategies can be fast given the right stimuli format. To understand how heuristics will perform across situations, the cognitive mechanisms that underlie heuristics need to be specified. However, the rules governing the specification of the "right" stimulus format are not known. This is a theme that has been built up throughout the course of the delegation study, the preference integration study, and these heuristic studies. The representations of choice options themselves clearly have a valuation component as shown in the delegation and preference integration study. But prior to the construction of the valuation component (and its uncertainty), the brain needs to construct a sensory and conceptual representation of the choice option. The heuristic studies presented here forced the decision algorithm that participants would use during the task. However, the trade-off hypothesis suggests that the human brain will reconfigure depending on the task goal or stimulus format that it is presented with. The following chapter presents an observational study addressing this theme.

Chapter 5 Choice option representations in the human brain: measures of neural similarity

5.1 Neural similarity study

What makes a brain state similar to another? Neural activity patterns never show the same pattern twice, not even for the same stimulus (Averbeck, Latham, & Pouget, 2006). This fact reveals properties about the brain and about the world itself; the neural code is probabilistic and signals from the environment are in constant change. Given its noisy biological substrate, the brain is required to abstract away commonalities from environmental signals by computing similarities between them.[7] Thus, a function of brain state similarity can be interpreted as the similarity measure used by the brain; a fundamental property that arises from the neural code.

Similarity is important to many cognitive processes and tasks, as formalised in computational models (e.g., Edelman, 1998; Goldstone, 1994). Understanding the brain basis of similarity can enlighten how the brain supports complex behaviours. Finding the brain's similarity measure will inform on its rules of abstraction, on whether the brain considers a stimulus to be identical to a previously encountered stimulus (recognition memory, e.g. Norman & O'reilly, 2003), whether it categorises another animal as posing a threat or not (categorisation, e.g., Mack, Preston, & Love, 2013), communicating the name of an object (semantic memory, Martin, 2007), how it can visually discriminate a line to be longer than another one (visual discrimination, cf. Müller-Lyer illusion, Howe & Purves, 2005) or even discriminating between similar tone frequencies (stimulus generalisation, Pavlov & Anrep, 2003). These tasks are so varied that the natural question to ask is if the brain uses one similarity measure across all tasks or uses a tailor-made version for each. Relatedly, given the hierarchical nature of the brain's processing stream, from low-level to high-level features (e.g., Bracci & de Beeck, 2016), it could also be the case that similarity measures differ between different brain regions for the same task and stimulus conditions.

Hence, there were two specific goals for this study. The first goal was to ascertain whether the similarity measure being used by the brain differs across regions of interest. The

[7] The fact that neurons are noisy leads some to speculate that this supports the argument that neural computations are intrinsically probabilistic, over and above the need to overcome biological noise (Knill & Pouget, 2004).

second goal was to investigate whether this measure differs across tasks and stimulus conditions (i.e., two previously published datasets). The aim was to address these open questions regarding the nature of neural similarity.

Understanding the nature of similarity is one of the central goals of cognitive science. In this study, the case is made that the best way to investigate the representation of choice options is through the study of their similarities. Decomposing the similarity measure used by the brain can provide information on functions regarding the representations of choice options. As displayed by the diversity of tasks mentioned above, similarity is one of the most fundamental operations in cognition (Tversky, 1977). It is the element that is common to all cognitive processes and, consequently, a necessary condition for all human endeavors. The foundation for this intuition can be found in the empiricist writings of David Hume (1955) and John Locke (1975), with statements asserting that similarity underlies inductive inference and that analogy is the "great rule of probability" (Weber & Osherson, 2010). In modern statistical terms, this relates to the grouping of *similar* events in a common sample space. Finding the right similarity measure is common to many research programs such as protein comparisons (Lipman & Pearson, 1985), DNA fingerprinting (Lynch, 1990), magnetic resonance image registration (Holden et al., 2000), pattern recognition (Chen, Garcia, Gupta, Rahimi, & Cazzanti, 2009; Hancock & Pelillo, 2011), distance metric learning (Xing, Jordan, Russell, & Ng, 2003) and the study of similarity judgments in psychology (e.g., Shepard, 1987; Tversky, 1977).

Historically speaking, various cognitive models – models of vision or of categorisation – have been grounded in notions of similarity (Edelman, 1998; Goldstone, 1994). Shepard elegantly hypothesised about the nature of psychological spaces when developing his method of multidimensional scaling; proposing that categories could be formally represented as points in a multidimensional space (Shepard, 1962), although previous proposals existed in the literature (Coombs, 1958). Thus, distances in this space would represent psychological notions of similarity under some functional transformation (i.e., exponential) (Shepard, 1987). Furthermore, endowing a psychological space with a specific metric depends on the separability of the stimuli attributes. For example, when stimuli differ on perceptually integral attributes such as lightness and saturation, then the Euclidean distance can be an adequate measure for describing psychological similarity. However, when the stimuli differ on

perceptually separable attributes such as size and shape, then another measure known as the city-block distance may fit the psychological data far better (Shepard, 1987).

Modelling similarity judgments within a metric space (as distances) imposes certain restrictions on the types of permissible judgments. Specifically, that they should be symmetric, non-negative, and respect the triangle inequality. Defining similarity judgments in this way provides certain mathematical conveniences for the cognitive modeler (Jäkel, Schölkopf, & Wichmann, 2008). Other approaches in psychological research have intended to relax these restrictions through alternative proposals such as featural approaches (Tversky, 1977; Tversky & Gati, 1982; Tversky & Krantz, 1970), probabilistic approaches (Ennis, Palen, & Mullen, 1988; Tenenbaum & Griffiths, 2001), structural approaches (Gentner & Markman, 1997), quantum probability approaches (Pothos et al., 2015, 2013; Pothos & Trueblood, 2015; Trueblood et al., 2014), or transformational approaches (Hahn, Chater, & Richardson, 2003). This is not to say that the metric approach does not have psychological or neural relevance; for example, grid cells are thought to naturally represent the local environment as a metric space (Giocomo, Moser, & Moser, 2011).

Previous studies have used different measures to relate pairs of brain states such as Pearson correlation or the Mahalanobis measure (Allefeld & Haynes, 2014; Haxby, Connolly, & Guntupalli, 2014; Kiani, Esteky, Mirpour, & Tanaka, 2007; Kriegeskorte et al., 2008). However, the issue remains with respect to what is the true measure that underlies neuroimaging data (i.e., in this case, fMRI data). In this sense, different similarity measures can be forced to compete as part of a model selection process. Each similarity measure brings with it a host of assumptions. These assumptions can provide insight into neural computations or at the very least help characterise the voxel activations. For example, Pearson correlation (a common measure for neural similarity, e.g., Nili et al., 2014) assumes that overall levels of voxel activity are normalised and that each voxel independently contributes to similarity, whereas Minkowski measures assume similarity involves distances in a metrical space instead of vector directions. Furthermore, the Mahalanobis measure expands on both Minkowski and Pearson by assuming that the distributional pattern of voxel activity is consequential. Which similarity measure best describes the brain's operation? Because some of the similarity measures considered are not distance metrics, henceforth the study will speak of similarity, which is defined as negative distance for the measures that are metrics. The measures discussed

here can be grouped into different families depending on their high-level properties such as scale invariance or whether their computations are sensitive to covariance (see Figure 5.1).

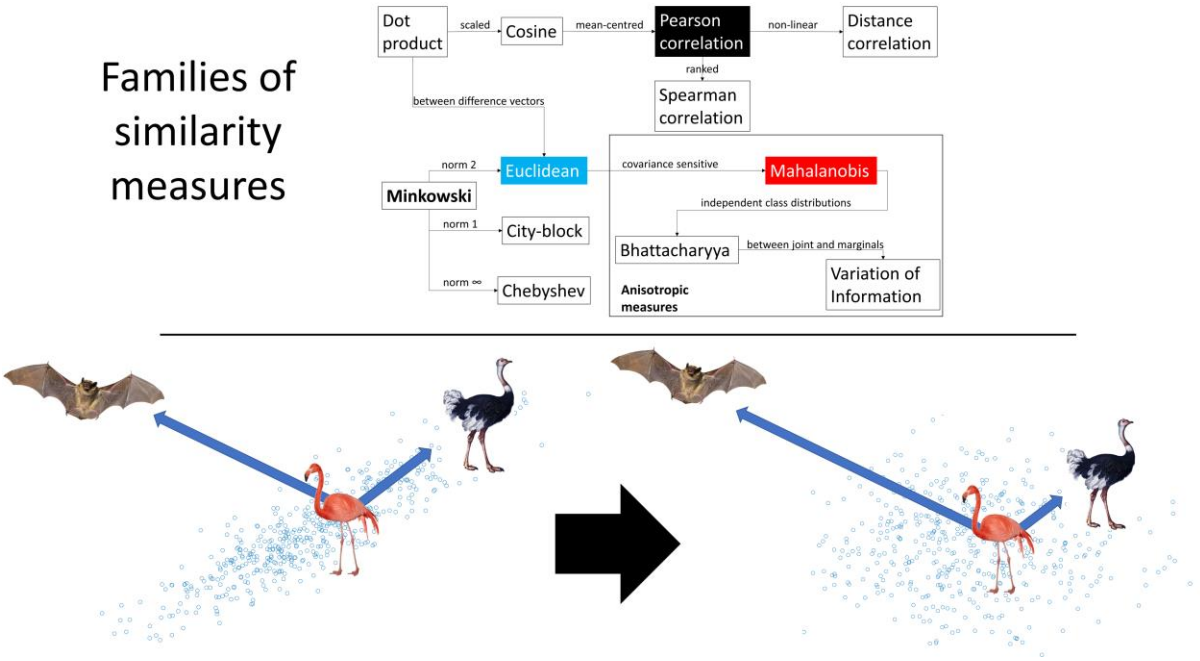

Figure 5.1 Properties of similarity measures.

Similarity measures are intricately related (top panel). Three families of measures can be identified: Minkowski measures, correlation-like measures, and anisotropic measures. Measures like Mahalanobis are anisotropic since the direction of measurement is important; this measure penalises stimuli that are far from the mean for directions of low variance. Thus, the Euclidean measure would say that the ostrich is more similar to the flamingo than the bat (left side, bottom panel). Mahalanobis would make this measurement in an uncorrelated space which has the effect that the ostrich is now even more similar to the flamingo and the bat is less so (right side, bottom panel). Thus, the scaling of the blue arrows differs between Euclidean and Mahalanobis, but the angle remains the same in this example. The angle is the focus of correlation-like measures which are invariant to changes in scaling.

The measures discussed here can be classified into two types of families; isotropic and anisotropic. Isotropic means that similarities are not influenced by the direction of the measurements. Conversely, anisotropic means the direction of measurement is consequential.[8] Isotropic measures can be further subdivided into Minkowski measures and correlation-like measures; examples of isotropic measures are Euclidean – which is a Minkowski measure with norm 2 – and Pearson correlation. These measures are intuitive in the sense that the direction

[8] Anisotropic measures should not be confused with asymmetric measures which give different values based on which stimulus is measured first (Nosofsky, 1992; Tversky, 1977).

in which the similarity measure is computed (for some chosen coordinate space) does not impact the measure. Minkowski measures are affected by scaling whereas correlation-like measures are scale invariant. Scale invariance is not a specific property of isotropic or anisotropic measures; it is a property that is only specific to some measures like Pearson correlation. Examples of anisotropic measures are Variation of Information, Mahalanobis, and Bhattacharyya measures. These measures consider the covariance between stimuli dimensions which means that the direction along which the measurement is made will impact the measurement. An example of the difference between families of similarity measures can be seen in Figure 5.1. The density of stimuli exemplars affects anisotropic measures whereas isotropic measures are not affected. The impact of exemplar density on categorisation of stimuli has been suggested before (Ashby & Perrin, 1988; Corter, 1987; Krumhansl, 1978). Furthermore, for the anisotropic measures considered here, the way the stimuli density affects the measurement is based on assumptions of Gaussian noise. Other noise distributions are possible (i.e., computing Variation of Information for non-Gaussian distributions).

For this study, the tradition of grounding similarity in confusability was followed; when two things are similar they are easily confused. Complementarily, when two things are dissimilar they are easily discriminated. This idea has roots in early psychological literature on stimulus generalisation (Pavlov & Anrep, 2003) and discrimination learning (Spence, 1952). Confusability of different brain states can be measured with classification algorithms such as linear support vector machines (SVM). For example, a sparrow may be more likely to be misclassified as a robin than a truck. Another consideration was to examine which similarity measure best characterises this confusability data for each brain region and task considered.

The method for picking the best similarity measure can be divided into two steps:

1) Find a gold standard approximation to the brain's similarity measure using a pairwise accuracy matrix from the best-performing decoder (i.e., best classifier).
2) Evaluate candidate similarity measures to discover which measure best describes the pairwise accuracy matrix found in step 1.

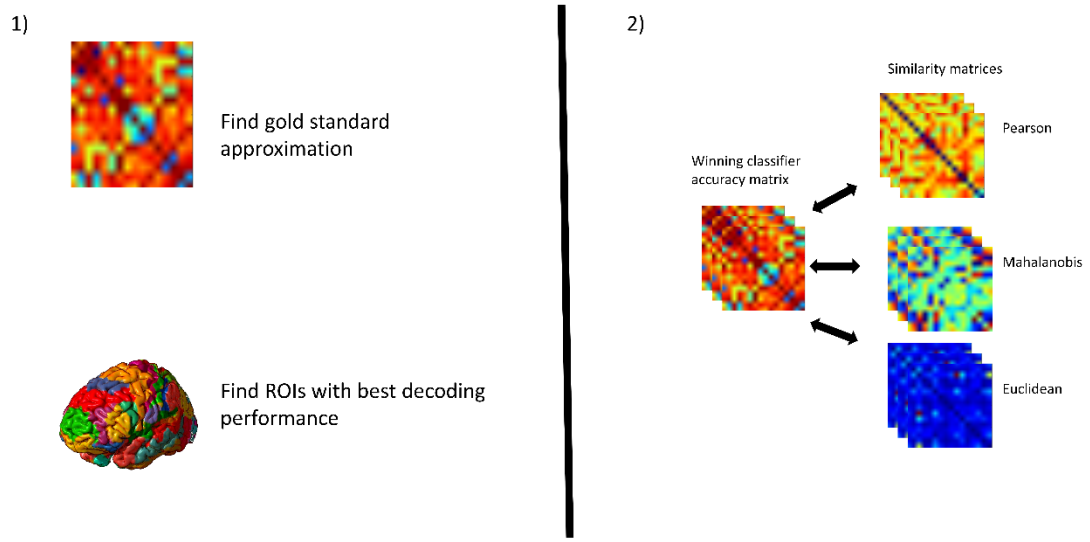

Figure 5.2 Cross-validated optimisation procedure.

The procedure starts with finding the best-performing classifier for each dataset (1), measured by accuracy of the classifier for each dataset (whole-brain). The winning classifier is then optimised to find the optimal number of features (i.e., voxels) for each ROI. The number of ROIs evaluated is reduced to a subset of top ROIs with the best classifier performance measured by mean accuracy. Feature selection is optimised for the similarity measures for the top ROIs (2) (see Methods section below for more details). Finally, the optimised accuracy matrix (for the wining classifier) is (Spearman) correlated with the similarity matrices on a held-out test set. This procedure was performed on each ROI independently. More than one matrix represents that this procedure was done for each run with leave-one-out cross-validation (LOOCV).

The first step consisted of choosing a classifier with the best performance in terms of accuracy (whole-brain classification) in decoding individual stimuli for each dataset. After finding the best-performing classifier, feature selection was optimised for the winning classifier for each ROI in each dataset. Subsequently, a subset of top ROIs ranked by classifier performance (i.e., mean accuracy) was obtained. This step is necessary since it ensures that the subsequent evaluation of similarity measures is based on true signal.

The second step consisted of optimizing feature selection for each similarity measure (for each top ROI in each dataset). Finally, Spearman correlations between each similarity measure's similarity matrix and the winning classifier's accuracy matrix were computed. The

accuracy matrix from the best-performing classifier constitutes a reference model to which all similarity measures can be compared to.

The initial analyses were done throughout the whole brain for 110 regions of interest (ROI) (excluding areas like cerebral white matter and the lateral ventricles, see Appendix C.1 for a list of the 110 regions)[9] from the Oxford-Harvard Brain Atlas (provided with FSL, Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) for two previously published datasets; one study that presented participants with geometric shapes (GS) (Mack, Preston, & Love, 2013) and another study that presented participants with natural images (NI) (Bracci & de Beeck, 2016). These studies are referred to as the GS study and the NI study respectively. A key point to mention is that the GS study consisted of stimuli that varied on four highly separable and independent binary dimensions (colour: red or green, shape: circle or triangle, size: large or small, and position: right or left), whereas the NI study consisted of naturalistic stimuli (photographs) that varied in unique ways and were not easily decomposable into predefined features (see Appendix C.2). This difference in stimuli sets between studies harbours the potential for finding differences in similarity measure profiles between studies.

The 12 ROIs that are reported in the Results section were selected as the union of the top ten regions across both studies, ranked by classifier accuracy within each study (see Figure 5.2 and Methods section below). This step is crucial because good decoding signal is necessary for obtaining informative similarity relations between all pairs of stimuli. Paradoxically, noise can slightly benefit the analysis presented here given that 100% decoding accuracy would result in an identity matrix for the accuracy matrix which would not reveal the similarity relations that are of interest.

All similarity measures were optimised to fit these accuracy matrices, using Spearman correlation as the optimisation criterion (Figure 5.2). Spearman correlation is best suited because it can avoid scaling issues and does not assume linearity. The similarity measures evaluated included the well-known Minkowski distances (which contain Euclidean and city-block), other measures that consider covariance structure such as Mahalanobis and Bhattacharyya, both Spearman and Pearson correlation, cosine distance and the dot product.

[9] More areas could have been excluded based on a priori hypotheses of where similarity signals would arise. However, including areas where no signal was expected served as a sanity check for the method (see Methods) and still retained the possibility that similarity signals could have been found in otherwise unexpected brain regions.

Mahalanobis and Bhattacharyya measures had diagonal regularisation (d), Ledoit-Wolf shrinkage (r) or no regularisation of their covariance matrices (see Methods section below). Two other similarity measures called distance correlation (not to be confused with Pearson correlation or Spearman correlation) and variation of information were also evaluated. Measures whose average Spearman correlation was more than three median absolute deviations away from the group average are not reported in the results; they would only skew the results obtained from the other measures (see Appendix C.3). These measures were chosen because they represent a wide variety of possible similarity measures and they can be divided into two families of measures; isotropic and anisotropic. Feature selection optimisation was done within each ROI for all similarity measures and for the best-performing classifier accuracies in each participant's native space (see Methods section below).

The aim of this study was to address open questions regarding the nature of neural similarity; if it differs across tasks and brain regions. The results focus both on the individual performance of measures and on the aggregate performance of all measures – denoted as similarity profiles – for different ROIs across both studies.

5.2 Methods

5.2.1 Datasets

The analyses are based on two previous fMRI studies: a study that presented simple geometric shapes (GS) to participants (Mack et al., 2013) and a study that presented natural images (NI) to participants (Bracci & de Beeck, 2016). The GS study consisted of a visual categorisation task with 20 participants and the NI study of a 1-back size judgment task with 14 participants. Descriptions of the tasks and acquisition parameters can be consulted in Appendix C.2. For further information, the reader should consult the source citation directly.

5.2.2 fMRI preprocessing

The original raw (NIfTI formatted) files from both studies were preprocessed and analysed using FSL 4.1 (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012). Functional images were realigned to the first volume in the time series to correct for motion, co-registered to the T2-weighted structural volume, high-pass filtered (128s), and detrended to

remove linear trends within each run. All analyses were performed in the native space of each participant.

5.2.3 Trial-by-trial estimates

For both studies, the method suggested by Mumford et al. (Mumford, Turner, Ashby, & Poldrack, 2012) known as LS-S (least squares separate) beta estimation was used to get a coefficient estimate for each individual presentation of each object. This method consists of calculating a general linear model for each object presentation with only two regressors; one regressor representing the effect of interest (the object presentation in question) and another regressor representing all other object presentations within the respective run. This procedure was done for each run separately to preserve as much statistical independence as possible between runs. Such a step is necessary for doing the multivoxel pattern analysis described below. After successfully estimating the object presentation coefficients within each run, these were then concatenated into a single 4D NIfTI formatted file. Furthermore, all runs were subsequently aligned to the last run within each study (e.g. the sixth run in the GS study or the sixteenth run in the NI study). The runs were then concatenated into a single 4D NIfTI formatted file for each participant within each study.

5.2.4 Initial region of interest (ROI) selection

The Harvard-Oxford cortical and subcortical structural atlases provided with FSL (Jenkinson et al., 2012) were used to parcellate the different anatomical regions for each participant (see Figure 5.2). A total of 110 regions of interest (see Appendix C.1) were used as masks that would be used in the multivoxel pattern analyses detailed below. The goal was to evaluate classifier accuracy across the whole brain (except for areas like cerebral white matter or the lateral ventricles). The masks were transformed from MNI space to each participant's native space. This masking by anatomical region can be considered the first part of a feature selection procedure. Feature selection was also done within each region of interest for each participant (see below).

5.2.5 Classification analysis

Pattern classification analyses were implemented using PyMVPA (Hanke et al., 2009), Scikit-Learn (Pedregosa et al., 2011), and custom Python code. Three classifiers were used for

the pattern classification: Gaussian naïve Bayes, k-nearest neighbor, and linear support vector machine (SVM). The output of one of these classifiers was to be chosen as the best representation of the underlying similarity matrix to which all other similarity measures would be compared to (see Neural similarity analysis below). The linear SVM was implemented with the Nu parametrisation (Schölkopf, Smola, Williamson, & Bartlett, 2000). This Nu parameter controls the fraction of data points inside the soft margin; the default value of 0.5 was used for all classifications. The k-nearest neighbor classifier was implemented using five neighbors. No hyperparameters required setting for the gaussian naïve Bayes classifier.

To pick the best-performing classifier, classification was conducted on the whole-brain (no parcellation into distinct ROIs) for each study independently. All classifiers were trained with leave-one-out *k*-fold cross-validation, where *k* was equal to the number of functional runs for each participant in each study (e.g. six runs in the GS study or sixteen runs in the NI study). To do feature selection on voxels, all voxels were ordered according to their *F* values computed from an ANOVA across all class (stimuli) labels. The top 300 voxels with the highest *F* values were retained based on classifier performance (i.e., accuracy) on the test run. For these classifiers, accuracy was computed across all classes (16 classes for the GS study and 54 classes for the NI study) with a majority vote rule across all computed decision boundaries (for classifiers where this is applicable like linear SVM). This means that random classification is equal to 6.25% for the GS study and 1.85% for the NI study for this whole-brain analysis. However, for all other classification analyses below, accuracy is computed as mean pairwise accuracy across all classes, which means that random classification is equal to 50%. The best-performing classifier was selected as the classifier with highest mean accuracy (mean across participants) in the GS and NI study, independently. Classifier accuracies (i.e., accuracy matrices) were multiplied by negative one for the neural similarity analysis explained below. This was done so that they would correlate positively with the similarity measures and facilitate presentation of results.

The following analysis was performed for each of the 110 ROIs that are described above. To train the classifiers leave-one-out *k*-fold cross-validation was also used. Within each fold, a (randomly) picked a validation run was used to tune the number of features (i.e., voxels) that would be selected for that fold. Thus, feature selection was done within each fold. To do this feature selection, all voxels were ordered according to their *F* values computed from an ANOVA across all class (stimuli) labels. This step aids classifier performance because it

preselects task relevant voxels (as opposed to item discriminative voxels). It is important to note that these ANOVAs were computed on the training runs but not on the validation run nor on the held-out test run, to avoid overfitting. The top *n* voxels with the highest *F* values were retained based on classifier performance (i.e., accuracy) on the validation run. Scipy's *minimize_scalar* function (Jones, Oliphant, Peterson, & others, n.d.) was used to optimise this validation run accuracy with respect to the top *n* voxels. After picking the top *n* voxels, the classifiers were trained on both the training runs and the validation run. Subsequently, the classifiers were tested on the held-out test run for that fold. This classification analysis was done for all possible pairwise classifications for each study (i.e., 120 pairwise classifications in the GS study and 1431 pairwise classifications in the NI study). From this analysis, the pairwise classification accuracies were retained for both the validation run and the test run for each fold.

5.2.6 Secondary ROI selection

The 110 ROIs were rank ordered by mean classifier accuracy (mean across participants) within each study. Subsequently, the union of the top ten ROIs was selected for the neural similarity analysis. This procedure was done to ensure that the ROIs used to evaluate the similarity measures was based on brain areas with adequate signal-to-noise ratio.

5.2.7 Neural similarity analysis

The goal of this analysis was to compare the representation of different similarity measures in the brain. The comparison criterion was chosen as Spearman correlation between all pairwise similarities and the classification accuracies mentioned above. This criterion was used since it avoids scaling issues. To achieve this, first all pairwise similarities (i.e., for all pairs of stimuli) were computed from the training runs defined in the classification analysis – not including the validation run. Incidentally, feature selection was also realised here. In the same fashion, as in the classification analysis, all voxels were ordered according to their *F* values computed from an ANOVA across all class (stimuli) labels. Then, the top *n* voxels with the highest *F* values were retained based on Spearman correlation of the similarities with the validation run accuracies of the classifier that were previously computed. After picking the top *n* voxels, the similarities were computed across training runs and validation run for those voxels. These similarities were then used to compute the final Spearman correlation with the

classifier test run accuracies. Conducting feature selection for the similarity measures is important because different measures leverage information differently.

5.2.8 Similarity measures

The following similarity measures were evaluated: dot product, cosine angle, city-block (Manhattan), Euclidean, three variants of Minkowski (with norms 5, 10 and 50), Chebyshev, Spearman correlation, Pearson correlation, three variants of Mahalanobis, three variants of Bhattacharyya, variation of information, and distance correlation. City-block, Euclidean, Minkowski, Chebyshev, Mahalanobis, Bhattacharyya and variation of information are proper distance metrices; to convert them to similarity measures they were multiplied by minus one. The three variants of Mahalanobis and Bhattacharyya were due to the way the sample covariance matrix was regularised; either no regularisation, Ledoit-Wolf shrinkage (implemented through Scikit-Learn, Ledoit & Wolf, 2004; Pedregosa et al., 2011) or diagonal regularisation. Diagonal regularisation was defined as the sample covariance matrix with all the off-diagonal elements set to zero. Note that city-block, Euclidean, and Chebyshev are also special cases of the Minkowski measure where the norms are set to one, two and infinity, respectively. To keep calculations consistent across all similarity measures, vector representations for each stimulus were defined as the mean vectors across trial presentations for that stimulus. Refer to Appendix C.3 for the equations for each similarity measure and Appendix C.4 for the covariance matrix regularisation procedures.

This analysis parallels the classification analysis in every way except that instead of optimizing model accuracy, here the optimisation criterion was model correlation (i.e., Spearman correlation) with the previously computed pairwise classifier accuracies.

5.3 Results

5.3.1 Classifier selection

The best-performing classifier was chosen out of three candidates; Gaussian naïve Bayes (GNB), k-nearest neighbor (KNN), and linear support vector machine (SVM). These classifiers were chosen because they are commonly used in data analysis, both inside and outside the field of neuroimaging, and they compute classification in very distinct ways (see Pereira, Mitchell, & Botvinick, 2009).

The linear SVM classifier was the clear winner across both studies, thus was chosen as our gold standard approximation to the brain's similarity measure. The performance of the linear SVM classifier compared to the other two classifiers is shown in Table 5.1.

Table 5.1 Linear SVM is best-performing classifier in both studies

	GS study		NI study	
	mean	s.d.	mean	s.d.
Linear SVM	20.49%	12.64%	23.51%	5.50%
GNB	15.00%	8.79%	10.24%	2.84%
KNN	14.51%	8.50%	8.49%	3.09%
Random classification	6.25%		1.85%	
	t	*p*	*t*	*p*
Linear SVM vs. GNB	5.22	< 0.001	14.33	< 0.001
Linear SVM vs. KNN	4.59	< 0.001	17.80	< 0.001
degrees of freedom	19		13	

Top panel shows mean accuracy and standard deviations (s.d.) (across participants) for each classifier. Bottom panel shows *t*-tests comparing the best-performing classifier (linear SVM) to the other two classifiers.

The linear SVM classifier was then optimised for each of the initial 110 ROIs. The ROIs were rank-ordered in terms of accuracy in each study and the union of the top 10 ROIs across both studies was: left and right intracalcarine cortex (CALC), left and right lateral occipital cortex (LO) inferior division, left and right lateral occipital cortex (LO) superior division, left and right lingual gyrus (LING), left and right occipital fusiform gyrus (OF), and left and right occipital pole (OP). This resulted in a secondary ROI selection of 12 ROIs with best (linear SVM) classifier accuracy.

Classifications were performed pairwise for this analysis and thus random classification is 50% for both studies (see Methods above). The mean accuracy for the linear SVM classifier in the 12 regions of interest was 59.47% (s.d. = 7.97%) in the GS study and 78.43% (s.d. = 7.41%) in the NI study. The best-performing classifier (linear SVM) was performing above 50% chance level in both studies; $t_{19} = 5.18$, $p < 0.001$, in the GS study and $t_{13} = 13.84$, $p <$

0.001, in the NI study (degrees of freedom are based on number of participants for each study). This provides reassurance that the ROIs that were selected indeed have information regarding stimuli presentation. Classification accuracy for the NI study was higher than in the GS study $t_{32} = 6.82$, $p < 0.001$, showing a potential difference in data quality due to the higher number of observations per stimuli in the NI study (see Methods section above).

5.3.2 Neural similarity

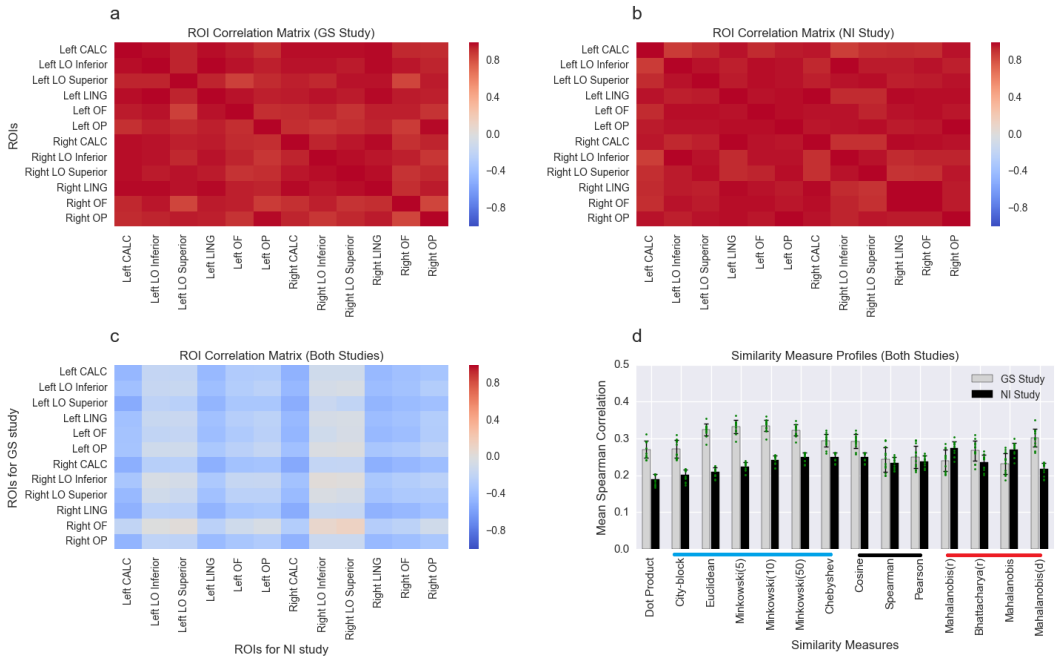

Figure 5.3 ROI correlation matrices and similarity measure profiles.

ROI correlation matrices for the (a) GS and (b) NI studies, demonstrating that the performance of similarity measures was Pearson correlated within task. ROI correlation matrix (c) demonstrating negative Pearson correlations for the similarity measures between studies. The 12 ROIs (see Methods above) were left and right intracalcarine cortex (CALC), left and right lateral occipital cortex (LO) inferior and superior divisions, left and right lingual gyrus (LING), left and right occipital fusiform gyrus (OF), and left and right occipital pole (OP). Mean Spearman correlations (d) for each similarity measure in the GS study (grey bars) and the NI study (black bars) is displayed. Measures under the solid horizontal bars are: Minkowski measures (blue bar), correlation-like (black bar), and anisotropic measures (red bar). The error bars are standard deviations. Green dots represent ROIs.

For each ROI, there was a vector of cross-validated mean Spearman correlations – averaged across runs, one mean per similarity measure (Figure 5.3, see Methods above). Pearson correlations of these vectors between ROIs are shown in Figure 5.3a-c. Mean Pearson correlation of the upper triangle for the correlation matrix was computed for each study. The

mean correlation (of similarity profiles) across the 12 ROIs was 0.95 (s.d. = 0.034) in the GS study (Figure 5.3a) and 0.96 (s.d. = 0.027) in the NI study (Figure 5.3b). Bartlett's test (1951), which shows the matrices are different from an identity matrix, was significant for the GS study, $\chi^2_{(66)} = 432.847$, $p < 0.001$, and the NI study, $\chi^2_{(66)} = 502.7494$, $p < 0.001$. Permutation tests (with 10,000 iterations), where the labels of the similarity measures were permuted, confirmed these results ($p < 0.001$). This provides evidence showing that the same similarity measure is being used for all ROIs within each study. The mean Pearson correlation for the correlation matrix that compares similarity profiles for ROIs across both studies (Figure 5.3c) was -0.27 (s.d. = 0.148). Jennrich's test (1970) showed that this matrix was different than a matrix of zeros, $\chi^2_{(66)} = 769.0349$, $p < 0.001$. Permutation tests (10,000 iterations) with shuffling of similarity label measures also confirmed these results ($p < 0.001$).

Furthermore, it is important to test whether the similarity measures differ in how well they characterise the gold standard. Do all measures do as well? To analyse the differences between the different similarity measures, a mixed effects model was performed for each study with Spearman correlations from the neural similarity analysis as the response variable. The models contained fixed effects of similarity measure, linear SVM accuracy, participant, and ROI. Linear SVM accuracy, participant, and ROI variables only serve to account for variance and get better estimates. The models also contained random effects of ROI (varying per participant) and of similarity measure (varying per ROI). Model comparisons were performed between the full model and a null model without any similarity measures. For the GS study, the effect of similarity measures was significant, $\chi^2_{(2)} = 1720.331$, $p < 0.001$. Likewise, for the NI study, the effect of similarity measures was significant, $\chi^2_{(2)} = 6770.249$, $p < 0.001$. These results show that neural computations are sensitive to differences in similarity measure representation (Figure 5.3d). A full model that included both studies was not possible due to convergence issues. However, the interaction between studies can be observed by the negative correlations of Figure 5.3c, tested above with Jennrich's test. *Post hoc* pairwise tests between similarity measures follow below as specific instances driving this interaction between studies.

Table 5.2 Comparison of similarity measures to Pearson correlation

GS study			NI study		
	z	*p*		*z*	*p*
Mahalanobis	(3.161)	0.02	dot product	(29.547)	< 0.001
			cosine	(22.803)	< 0.001
dot product	4.053	< 0.001	city-block	(10.411)	< 0.001
cosine	4.532	< 0.001	Mahalanobis(d)	(7.593)	< 0.001
Chebyshev	6.353	< 0.001	Euclidean	(5.170)	< 0.001
Minkowski(50)	6.624	< 0.001			
Mahalanobis(d)	8.825	< 0.001	Minkowski(10)	4.005	< 0.001
city-block	10.479	< 0.001	Chebyshev	4.733	< 0.001
Minkowski(10)	10.459	< 0.001	Minkowski(50)	4.920	< 0.001
Euclidean	12.145	< 0.001	Mahalanobis	10.304	< 0.001
Minkowski(5)	12.562	< 0.001	Mahalanobis(r)	11.301	< 0.001
Linear SVM accuracy	48.67	< 0.001	Linear SVM accuracy	85.54	< 0.001

Left panel shows significant *z* statistics for measures worse than Pearson correlation (in brackets) and better than Pearson correlation for the GS study. Right panel shows the same for the NI study. The effect of Linear SVM accuracy for the mixed effects model in each dataset is also included for completeness. *p* values are Bonferroni corrected.

To assess the performance of each individual measure, the Pearson measure was used as a baseline measure for comparison with all other similarity measures. These contrasts give *z* statistics with Bonferroni corrected *p*-values. The results in Table 5.2 provide evidence that some similarity measures are a superior description of our gold standard approximation (i.e., the best-performing classifier; linear SVM in this study) to the brain's similarity measure. The results for the GS study show that the Minkowski measures (including Euclidean and city-block) have the highest *z* statistics which is suggestive of the fact that stimuli in this study had a feature space where the dimensions were easily decomposable. The results for the NI study show that two of the Mahalanobis measures have the highest *z* statistics, amongst the measures that performed significantly better than the Pearson measure, which is suggestive of the fact that stimuli were naturalistic images (photographs) where the feature space is not easily decomposed as in the GS study.

Minkowski measures presented the best performance in capturing the gold standard approximation to the brain's similarity measure in the GS study, compared to Pearson correlation used as a baseline. The same is true for the Mahalonobis(r) measure (regularised with Ledoit-Wolf shrinkage) in the NI study. Given that the Minkowski measures had the

highest *z* statistics when compared to the Pearson measure in the GS study, one of the measures (Euclidean) was chosen for portrayal in Figure 5.4. For the NI study, the Mahalonobis(r) measure was chosen for the same reasons. All three measures (Euclidean, Mahalanobis(r), and Pearson correlation) are common selections for computing neural similarity (Nili et al., 2014). Thus, Figure 5.4 illustrates the ROIs where these two measures outperform Pearson correlation (computed with paired sample *t*-tests, $p < 0.05$, uncorrected, for visualisation purposes only).

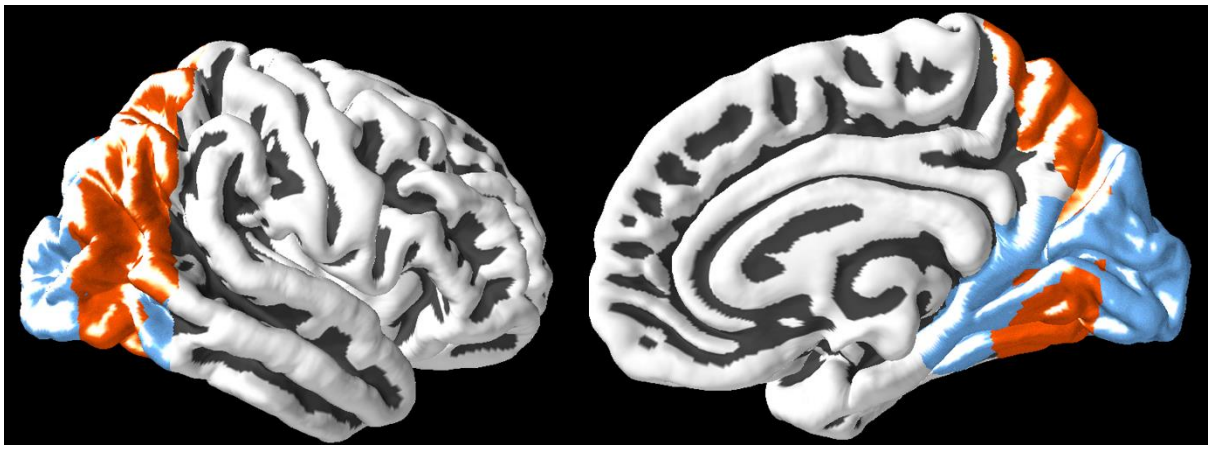

Figure 5.4 Similarity measures per ROI.

Presentation of lateral (on the left) and medial (on the right) views for the right hemisphere. Coloured areas (blue or red) show areas where the Euclidean measure significantly outperformed the Pearson measure for the GS study. Red colours only show areas where Mahalanobis(r) (regularised with Ledoit-Wolf shrinkage) outperformed the Pearson measure in the NI study. Significance defined as $p < 0.05$ for paired sample *t*-tests, uncorrected, only for demonstrative purposes.

5.4 Discussion

This study assessed what makes brain states functionally similar by evaluating a wide range of possible similarity measures. The aim of this assessment was to inform on how these measures constitute the representation of choice options in the brain. Using data from two previous studies, it was found that 1) measures of similarity differ in how well they gauge relationships between brain states, 2) different brain regions appear to use a common approach to coding state similarity, and 3) the operable measures of similarity vary across studies (i.e., task and stimulus set). These findings suggest that the representation of choice options with respect to neural similarity measures morph as a function of task demands or stimuli attributes.

Although significant differences were found between similarity measures, all Spearman correlations are between .15 and .35 for both studies (Figure 5.3d). While robust results can be achieved without formal evaluation of the implemented similarity measure, the results show that still important differences can be revealed that inform on underlying brain function. Although speculative, the fact that Minkowski measures performed best in the GS study, when stimuli were readily represented in a multidimensional space, is suggestive. One can readily contrast this with the fact that Mahalanobis provided a better fit for the data in the NI study. Perhaps the covariance structure between classes (i.e., pairs of stimuli) becomes relevant to neural computations when the stimuli are increased in complexity; assuming that geometric shapes are simpler than natural images. Such a case could reflect the fact that the stimuli in the GS study were orthogonal to each other whereas the natural images in the NI study were not designed in this way. This implies that a mapping going from the stimuli covariance to the task relevant covariance structure of voxel activations can be identified (Diedrichsen & Kriegeskorte, 2016). This claim needs further investigation but varying the covariance structure of the stimulus set could generalise to predictions for new studies (i.e., higher covariance between stimulus features results in higher covariance between voxels with respect to a Mahalanobis measurement of similarity). However, it is possible that differences in how measures of similarity performed across studies were due to differences in data quality (including issues concerning study design), cohort effects, or differences in fMRI equipment. The data quality issue is most relevant since it can directly impact the comparability between neural similarity measures. For instance, measures such as Mahalanobis or Bhattacharyya need to estimate inverse covariance matrices. These matrices grow with the square of the number of vector components which approaches both numerical and statistical unreliability when the number of components approaches the number of observations. This limiting factor was the reason for optimizing the number of top features (i.e., voxels) for the measures; so that all measures could compete on similar footing with respect to computational hurdles such as these. Relatedly, this was also the reason for regularizing the sample covariance matrix for the Mahalanobis and Bhattacharyya measures. This problem relates generally to the curse of dimensionality. Interestingly, adding features or dimensions indefinitely does not necessarily improve separability between classes since most measures, such as Minkowski measures, will be practically equivalent in very high dimensional spaces (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999). The complexity of the measures evaluated here interacts with the data quality. With less noise, more samples, and better-quality data overall, the results presented here could

change and reveal that other more complex measures (like distance correlation and variance of information) can be a better model fit to the gold standard approximation to the brain's similarity measure.

Although, it has been pointed out that comparing similarity correlations between regions has its obstacles given the influence of voxel selection, fMRI noise and non-specific voxel activation patterns (Diedrichsen, Ridgway, Friston, & Wiestler, 2011), the finding that similarity measure profiles were consistent across brain regions is indicative of two things. First, the fact that the measures were independently optimised for each ROI and yet showed such robust profile correlations provides reassurance against the results being a product of overfitting. Second, building on the previous claim, the ROIs that are task responsive seem to respond to the orthogonality of the stimuli in a regular manner that is task (or stimuli) specific. Of course, this finding needs to be delineated within the nature of the brain regions that is reported here. Both studies were comprised of simple tasks (1-back and categorisation) and they both presented visual stimuli in regular intervals. The ROIs that are reported are expected for experimental designs of this type; this study remains agnostic as to whether the effects are driven by differences in stimulus sets or tasks, although intuitively it seems that the former is a more likely candidate. Future research could attempt to disentangle effects of task and stimulus set, which was not possible with the two datasets that were analysed here. Perhaps different sensory modalities or different task demands would result in more varied similarity profiles for different brain regions. This remains an open question for further studies to investigate.

A great deal of effort has been put into understanding the nature of similarity in the behavioural realm, but the foundations in neuroscience are less established. For example, powerful analysis techniques that involve comparing the similarities of brain states, such as representational similarity analysis (RSA), typically assume that brain states are similar to the extent that they are Pearson correlated (Kriegeskorte, Mur, & Bandettini, 2008). As mentioned above, a common measure used for computing pairwise similarities in RSA is Pearson correlation (Haxby et al., 2001; Kiani, Esteky, Mirpour, & Tanaka, 2007; Kriegeskorte et al., 2008). Other possibilities include measures such as the Mahalanobis distance (Allefeld & Haynes, 2014; Nili et al., 2014), the linear discriminant *t* value (Kriegeskorte, Formisano, Sorger, & Goebel, 2007), or the linear discriminant contrast (LDC) – which can be thought of as a cross-validated version of the Mahalanobis distance (Nili et al., 2014).

The fact that the linear SVM classifier outperformed the KNN and GNB classifiers was not surprising. Linear SVMs have been reported to perform well on fMRI data and present superior accuracy when compared to other classifiers (Misaki, Kim, Bandettini, & Kriegeskorte, 2010; Pereira, Mitchell, & Botvinick, 2009). Furthermore, SVMs have the advantage that their solutions scale well with the dimensionality of the data, they provide good generalisation properties and they make no assumptions about the distribution of the data (cf. empirical risk minimisation in Vapnik, 2013).

In general, the study of similarity of spaces has wide application in and outside of neuroscience. For example, understanding which similarity measures perform best has applications in language models with word embeddings in a vector space where a measure like Euclidean distance is meaningful (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Yih, & Zweig, 2013). Perhaps such a representation is also possible with voxel spaces, analogous to old proposals of meaning being constructed from a state-space semantics (Churchland, 1993; Fodor & Lepore, 1992; Fodor & Lepore, 1999; Garzon, 2000). Also, seemingly unrelated cognitive phenomenon like reward prediction errors (Schultz et al., 1997) or the Weber-Fechner law (Fechner, 1966) can be interpreted as a way of computing similarity between brain states. Measures of similarity are sometimes quite specific to the domain and dependent on idiosyncrasies of the data such as work on olfaction (Soucy, Albeanu, Fantana, Murthy, & Meister, 2009) or with neuronal spikes (van Rossum, 2001). However, the measures presented here are quite domain general and can be interpreted with respect to their information usage (Lin, 1998). With bigger datasets, future studies might wish to use techniques that allow to infer a similarity measure, or a specific covariance matrix for the Mahalanobis measure, directly (Xing et al., 2003).

In conclusion, strong deductions based on only two datasets should be taken with caution. However, it is suggestive that these two datasets (from the GS and NI study) expressed different similarity measures based on differences on stimulus properties and task goals. Further studies would need to hold the task goal constant and change the stimuli, or vice versa, to obtain a complete picture of how each component can have an effect on choice option representations in the brain. This does nonetheless provide further evidence for the trade-off hypothesis; that choice option representations necessarily accommodate between informational input and task goals.

Chapter 6 General discussion

In this dissertation I evaluate the hypothesis that cognitive representations of choice options reveal a trade-off between accommodating task goals and the format of the information sampled from the environment of the task. This hypothesis proposes a specific theoretical vantage point as to what has causal effects on choice option representations. The alternatives to a trade-off between task goals and stimulus format are:

1) Stimuli format and task goals operate independently and do not have any causal effect on choice option representations.
2) The causal links are unidirectional: a) task goals fully determine choice option representations, or b) stimuli formats fully determine choice option representations.

The lessons learned through the work presented here is that indeed both task goal and stimuli format simultaneously have effects on the cognitive representations of choice options, supporting the hypothesis that a trade-off exists. Specifically, the trade-off seeks to emphasise that adequate stimulus formats can optimise satisfactory task completion, and vice-versa, that task goals can be modified to suit the stimulus formats that are available (e.g., when heuristic compliance is enhanced or impaired as seen in chapter 4). However, the findings from the experimental work serve to enlighten other aspects that were not originally contemplated. First, task and stimuli formats have interactions that are independent of the agent (e.g., as consequences of experimental design goals or of other external processes at work). This point is addressed below in the section on interactions between task and stimuli. Second, task goals and stimuli formats do not *fully* determine choice option representations; individual differences concerning sense of self or personal preferences can impact the representation as seen in chapter 2 and 3, respectively. This point is addressed below in the section on contextual factors and prior experience.

The evaluation of the trade-off hypothesis was enabled by a 2x3 theoretical framework; bottom-up/top-down approaches through the lens of Marr's three levels of abstraction (implementation, algorithm, and computation) (Figure 6.1).

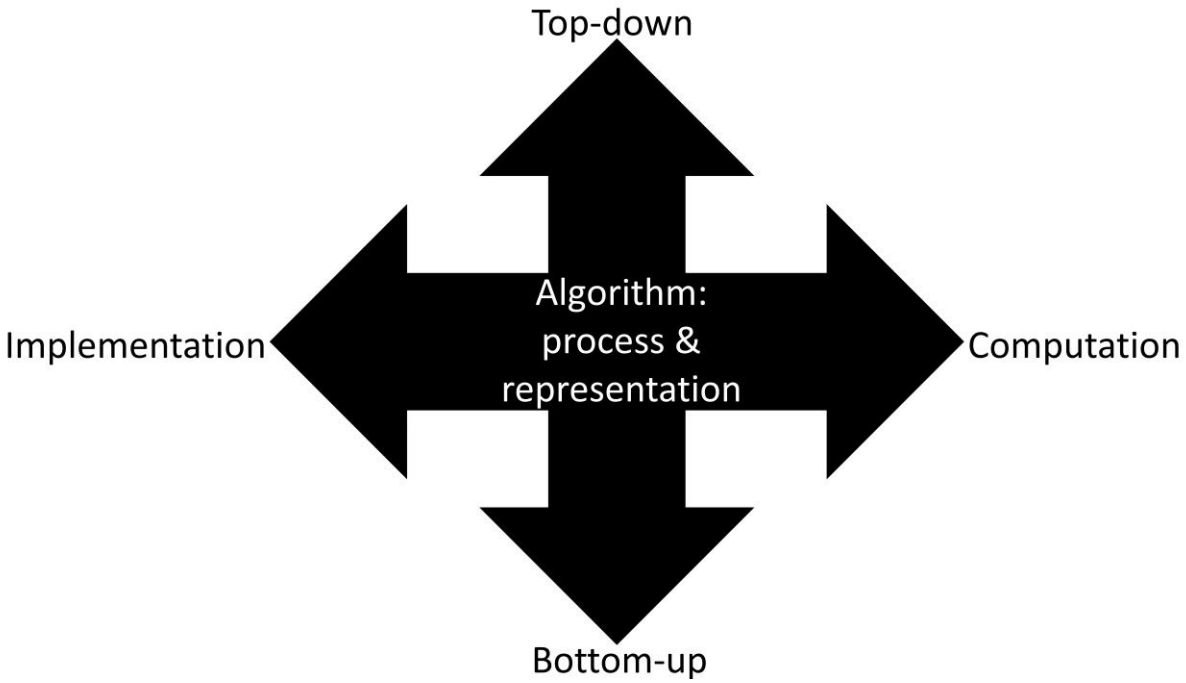

Figure 6.1 Two by three theoretical framework for studying choice option representations.

Top-down approaches (influences of task goals) and bottom-up approaches (influences of stimuli formats) interactively constrain choice option representations. Marr's levels constitute an orthogonal axis that provides computational and implementational constraints for studying the algorithmic level (decomposed into process and representation).

The top-down approach was used in chapters 2 and 3 since more emphasis was given to the effect of task goals on choice option representations; i.e., the behaviour, as a result of choice option representations, was primarily driven by the task goals. The bottom-up approach was used in chapters 4 and 5 since more emphasis was given to the effect of stimulus format on choice option representations; i.e., the observed behaviour, as a result of choice option representations, was primarily driven by stimulus format. Clearly, the findings eliminate the possibility that choice option representations operate independently of task goals and stimuli formats and further enlighten the forces that drive the trade-off between them. The dissertation as a whole shows that both task goals and stimuli formats can have effects on choice option representations. What is missing from this account is to show that task goals and stimuli formats do not have independent or isolated effects on choice option representations but operate simultaneously and interactively. To argue for this, it will be necessary to show that although a top-down or bottom-up approach can be taken experimentally, task completion is never truly independent of the stimuli used from the agent's point of view. This is a key issue that addresses

the interaction between the limits of experimental control and the nature of the cognitive representations of choice options. Hopefully, this digression will serve to shed light on the broader issue of the boundary conditions of choice option representations (i.e., when is a representation better suited to the stimuli and task at hand?).

As mentioned before, the hypothesis was motivated by the fact that most decision-making studies only focus on the value and uncertainty of each choice option but give less importance to the representation of the choice option itself. To be more precise, value and uncertainty constitute a key subset of features of choice option representations, but a subset nonetheless. Addressing choice option representations is difficult since it broadens the landscape of experimental variables to be considered. This explains the original focus of decision-making theories mostly on value and uncertainty, as displayed in the introduction of the present work. Then again, all experiments in psychology control for task goals and stimuli in one way or another. However, sometimes too much focus is given to task outcomes due to differences in experimental conditions and to the processes that led to those outcomes. Though they operate on the same level (Marr's second level), the argument made here is that representations present a slightly more nuanced view than processes. Processes describe sequential transformations on informational input whereas the representational view emphasises static questions pertaining to the morphology (shape) of the information at each time point. This work also seeks to make inferences on choice option representations due to differences in conditions (i.e., task or stimuli) or by comparison to benchmark models (e.g., EUT or BMC). After studying the interactions between task goals, stimuli formats, and choice option representations, a key learnt outcome of the work presented here is that cognitive models do not have to reside solely at one of Marr's levels but may have differential explanatory power at various levels (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010).

6.1 Interactions between task and stimuli

If the trade-off hypothesis is correct, then the empirical nature of this phenomenon imposes restrictions on experimental control. This means that if an experimental design seeks to focus more on controlling the subtleties or nuances of task goals, the adequate stimulus format follows, and vice versa. For example, in chapter 3, one of the goals of the study was to evaluate if humans follow the principles of Bayesian integration of value-based preferences. In effect, the way to address this required certain characteristics of the stimuli to be fixed (i.e.,

participants were shown distributions of social preferences in order to make this model evaluation feasible). Complementarily, in chapter 4, one of the goals was to show that serial ordering and colour-coding would be beneficial to the summation operations of the TAL heuristic. This experimental goal dictated that the agent's task goal should be constructed as compliance with this heuristic (or with TTB) presenting conflicts with stimulus format depending on experimental condition. This shows why defining one factor (task or stimuli) leads to a natural interaction with the other factor (stimuli or task, respectively).

Thus, predefining task and stimulus conditions predetermine the choice option representations that can be studied. This means that the empirical trade-off, that choice option representations must handle between task goals and stimulus format, translates into an experimental design trade-off; choosing a task goal impacts choice of stimuli and vice-versa. The observation that tasks interact with stimuli, independent of the agent, is true not just for experimental settings but for real-world processes as well. Normally, a task goal is set within a real-world domain, thus the stimulus set within that domain is intricately related to the task. For example, for an architect building a house, the task is naturally associated with available construction materials and the landscape properties for the construction site. Nowadays, an architect can negotiate between different stimulus formats to reach the end goal; blueprints, computer drafting and drawing software, scale models, or pencil and paper sketches. This example shows that not only the stimulus set is constrained by the task goal but the variability in stimulus formats is as well. The notion of compatibility between task and stimuli is related to the notion of regulatory fit of orientation towards a goal and the means to achieve that goal (Higgins, 2005).

Experimentally speaking, one seeks to minimise or account for the influence of external factors (external to the experimental context or research interests). At times, one seeks to relax this experimental control strategically in order to broaden the scope of contextual variables considered. The work presented here can be seen as starting with a broad scope of contextual variables (chapters 2 and 3) which then seeks to strip away complexity and add more experimental control (chapters 4 and 5). The former has the advantage of providing a more holistic view on choice option representations while the latter provides more precision for the conclusions that can be reached.

6.2 Contextual factors and prior experience influence choice option representations

The trade-off hypothesis states that task goals and stimuli formats interactively determine choice option representations. This statement proposes that task and stimuli format are the strongest causal determinants of such representations. However, this statement does not negate that there are other contextual factors, besides task goals and stimuli format, that can influence choice option representations. Contextual factors, especially pertaining to the participants' prior experiences, are shown to be germane to the findings of chapters 2 and 3.

The sense of self versus other is relevant to chapters 2 and 3 given that they were both experiments that operated within the social domain. The social domain was picked as a convenient domain for studying choice option representations because:

1) It is a natural domain for everyday decision-making
2) The task goals are easily explained to participants
3) High-level features of choice option representations are more readily studied in this domain
4) The top-down approach (emphasis on task goal constraints) is straightforwardly compatible with tasks in this domain

The social domain was not of interest in and of itself but only used as a vehicle for studying high-level features of choice option representations. However, after reviewing the findings of these two chapters, one can notice that there are factors that can affect choice option representations that are independent of task goals and stimuli format like sense of self and control. Thus, in chapter 2 it is found that participants are both overconfident in their performance (of choosing winning stimuli) and that they impose a premium on having control over their choices (intrinsic value of choice and control). These variables are outside of the experimental context and are related to the participants' prior experiences. The goal of the delegation task (i.e., choosing between yourself and an advisor) naturally led to describing the advisor with key pieces of information (i.e., accuracy and cost), driven by the assumptions of the computational model to be evaluated (i.e., EUT). This in turn meant that the self-as-a-choice-option retained representational flexibility in the task, resulting in the findings of overconfidence and the control premium for both gains and losses.

The delegation study is interesting because the choice options themselves are the potential decision-makers (i.e., self or advisor); requiring a representation of the self. Specifically, it requires a representation of the agent's performance for the task, perhaps influenced by the

agent's self-representation. Similarly, it requires a representation of the advisor. Although participants were instructed that the advisors were based on precomputed calculations from a computer algorithm, they may have imposed certain anthropomorphic characteristics onto the advisor or come with their own preconceived notions of what a computer algorithm entails (Duffy, 2003). This could imply that participants were dubious of the advisor's intentions (i.e, theory of mind for the advisor, Amodio & Frith, 2006; Behrens, Hunt, & Rushworth, 2009; Frith & Frith, 2005). Nonetheless, such complex representations for this task were not expected given that the participants were only given two pieces of information of the advisor on each trial: their accuracy in choosing a winning shape and the charge they would impose upon correct selection of the winning stimulus. Thus, the analysis focused on the representation of the self (as a choice option). Indeed, the construction of the self-as-a-choice-option for this task was influenced by subjective beliefs such as, 1) overestimation of their accuracy in the learning task (i.e., overconfidence), 2) awareness of their delegation choices being suboptimal, 3) the fact that choice itself has an intrinsic value. The findings are in accord with a past study that identified a significant "control premium" in an experimental setting in which participants could bet that a partner, or instead themselves, would answer quiz questions correctly (Owens et al. 2014).

Although this preference for control cannot be explained by overconfidence, it is important to note its role as part of the construction of the self nonetheless. The finding that humans possess a bias towards control is conceptually in accord with other experiments such as the authority game in Fehr et al. (2012), the evidence for the intrinsic value of decision rights in Bartling et al. (2013), and the phenomenon of reactance in Brehm & Brehm (2013). Note that the way delegation is framed in this task is more akin to the notion of expert advice taking (Harries, Yaniv, & Harvey, 2004; Harvey, Harries, & Fischer, 2000) than to the traditional notion of delegation between a boss and a subordinate (Bendor, Glazer, & Hammond, 2001). The nature of the social relationship surely influences the representation of an agent as a choice option; it is not the same delegating your taxes to an accounting expert than to delegate administrative paperwork to a subordinate. The functional roles differ as well as the contractual terms and relative hierarchy between agents.

The results for overconfidence agree with results that have been seen before in previous studies (Juslin, Winman, & Olsson, 2000; Klayman, Soll, González-Vallejo, & Barlas, 1999). The reasons for overconfidence are still not well understood but it is a bias that can be classed

together with other biases that hold a general positive outlook on events and attributes that are self-referential such as the optimism bias (Sharot, 2011; Sharot, Riccardi, Raio, & Phelps, 2007) or the better-than-average effect (Krueger & Mueller, 2002; Larrick, Burson, & Soll, 2007). There may be something special about computing confidence for self-referential attributes that results in overestimation of that quantity.

Representation of the self is also relevant to the findings in chapter 3. There, the choice options were more conventional; retail products from the UK Amazon website. But the preferences were clearly presented as pertaining to the self or to the social aggregate that bought and/or reviewed the product before. Prior experiences of the participant can influence the preference ratings and their consequent updating with social information. The participant could have varying degrees of beliefs regarding how reliable social preferences are in general, how reliable they are for online reviews, or for that website in particular. This characteristic of the self can be interpreted as the degree of stubbornness in personal preferences and reluctance to update towards the group mean (i.e., the degree of resistance to Amazon reviews as expressed in equation 3.2).

However, representation of the self is not the only contextual factor that could have an effect on the findings in the preference integration study. The representations of the retail products themselves are certainly influenced by prior experiences with retail products in general. Although, the stimulus format sought to mitigate and minimise this effect through the use of pictures of the items and bullet point descriptions. The effect of prior experience on preference integration was intended to be captured by the Bayesian update model which integrates confidence ratings from the initial preference.

Indeed, the KL divergence between prior and posterior ratings distributions, which was found to correlate with dmPFC activations, attests to the claim that the conflict between preferences of the self and preferences of others deserves a cognitive representation. The KL divergence signature for preference updating in dmPFC is in accord with a previous study that showed dmPFC activation for socially modelled value (Nicolle et al., 2012), indicating that this area may be useful for representing social preferences and helpful for processes that require theory of mind (Amodio and Frith, 2006; Behrens et al., 2009), or processes dealing with social conformity (Behrens et al., 2009; Berns et al., 2010; Campbell-Meiklejohn et al., 2010; Klucharev et al., 2011; Izuma and Adolphs, 2013; De Martino et al., 2013b). The KL

divergence was a result of a Bayesian model of belief updating, although Bayesian updating is not a necessary requirement for observing choice option representation trade-offs.

6.3 Levels of complexity and experimental control

Top-down approaches, such as the ones pursued in chapters 2 and 3, place emphasis on the way task goals constrain behavioural outcomes and cognitive representations. Thus, task goals, especially when formulated in an explicit way, consist of high-level abstractions. An obvious exception to this would be motor tasks that have become automatic (Bargh, 1994). Thus, top-down approaches are well suited for studying abstract, high-level concepts, like sense of self, control, and theory of mind. However, as argued above, contextual factors and prior experience can play a role in experiments that can be hard to control for but worth studying nonetheless. Stripping away this level of complexity can provide superior, more fine-tuned, experimental control.

Chapters 4 and 5 presented studies that were designed with a bottom-up approach in mind, that reduces experimental complexity and provides the possibility of gaining more control over variables like stimulus format. For example, the choice options used in the heuristic studies were countries, denoted generically as “Country A” and “Country B” for all choices. These were just placeholders that were used to denote the different options on each trial. The rationale was that using these placeholders, as opposed to names of real countries, would prevent people from using prior knowledge about these countries and interfering with the task goal (i.e., heuristic compliance). This was also the rationale of choosing a macro-economic domain; it is a domain that people are familiar with but do not necessarily have thorough experience with. Yet the economic statistics used in the task may still have suffered from such biases, which is why cue randomisation with respect to cue values is the best methodological approach to averaging over such possible confounds.

6.3.1 Instructed strategy use

Focusing on a bottom-up approach also facilitates the study of the algorithmic and representational level from a different perspective. Instructing participants to use a specific algorithm (i.e., decision strategy or heuristic) carries benefits over setting up a task goal and comparing behavioural outcomes with a benchmark model (e.g., BMC or EUT) after the fact. With counted exceptions (Khader et al., 2011; Khader, Pachur, & Jost, 2013), most heuristic

decision-making studies allow agents to enjoy the free choice of decision strategy during the task (Bergert & Nosofsky, 2007; Bröder & Gaissmaier, 2007; Lee & Cummins, 2004; Newell & Shanks, 2003). It is only through normative or descriptive accounts of decision-making that these decision strategies are later recovered, usually through some form of model fitting. The motivation for instructed strategy use in Khader et al. (2011) was to avoid strategy switching, especially since they were interested in studying the cognitive properties of a specific decision heuristic. This was the same motivation for the experimental designs presented in Chapter 4; to have more experimental control and directly study the properties of the decision strategies of interest (in this case TAL and TTB). The experimental approach in Chapter 4 differed greatly from the one presented in the previous chapters. In chapters 2 and 3, a normative model is assumed based on expected utility theory and perhaps specific assumptions regarding either the frequentist or Bayesian computation of uncertainty. Such an approach sought to recover properties of the representation of choice options based on the computational assumptions made by these models. In contrast, chapter 4 sought a descriptive processual account of decision-making. By holding the decision strategy constant, the study of the trade-off of interest – between task goals and stimuli format in the construction of choice option representations – was more precise.

The choice of heuristics as instructed decision strategies can be seen as conflicting with the choice of a Bayesian model of preference integration in chapter 3. Quite the contrary, the contrast between an optimal model of decision-making (Bayesian) or an approximation to that optimum (heuristics) only highlights the importance of choice option representations for both accounts. Sidestepping the issue as to which class of models better describes human decision-making, it is clear that the trade-off that choice option representations must handle between task and stimuli is important for both classes. However, at the same time it is clear that one class of models will predict different choice option representations than the other – as when Bayesian models will require a full distributional representation of uncertainty, or at least a way to sample from it. In chapter 4, the fact that heuristics are seen as a simpler algorithmic alternative to the high computational demands of Bayesian computation is addressed. However, this argument is turned on its head when it is shown that heuristics themselves may constitute hidden computational and cognitive complexity (e.g., the search costs of TTB when compared to TAL).

The findings of Heuristic Study 1 in chapter 4 revealed the complexity of a supposedly simple algorithm like TTB as other studies have shown before (Bergert & Nosofsky, 2007; Bröder & Gaissmaier, 2007; Dougherty et al., 2008; Juslin & Persson, 2002; Khader et al., 2011; Lee & Cummins, 2004; Newell & Shanks, 2003). The algorithmic complexity of a decision strategy may not be as important as its cognitive implementation. This line of thinking may result in counterintuitive findings, such that more cues can be faster to process as long as the added information increases coherence among cues (Glöckner & Betsch, 2008). The choice of a class of strategies, whether they be compensatory like TAL or non-compensatory like TTB, can be evaluated by analysing the findings of both experiments in chapter 4 as a whole. As mentioned previously, the reversal of effects between Heuristic Studies 1 and 2 agrees with previous research showing that the decision-making context influences the choice of a specific heuristic (Bröder, 2003; Bröder & Schiffer, 2006; Platzer & Bröder, 2012; Platzer, Bröder, & Heck, 2014).

However, the contingency of decision heuristic selection embodies the most important point of interest; showing the dependency of the representation of choice options on the interaction between task goals (i.e., complying with TTB or TAL) and stimulus properties (colour-coding and serialisation, respectively). The alignment between heuristic and stimulus properties is made clear if one focuses on the core operations required by each heuristic; serial search for TTB or summations for TAL. Stimulus properties that facilitate the cognitive implementation of one or the other will result in easier compliance with the respective heuristic. Note that the stimulus properties in chapter 4 differ from previous chapters in the sense that the choice options in chapter 4 (such as “Country A” and “Country B”) share the same properties within condition, whereas the previous chapters have different stimulus properties for each choice option (self and advisor in chapter 2 and personal versus social preference ratings of retail products in chapter 3). These notable differences in the generalisability of stimulus properties of choice options within task are part of the motivation for stepping back to address a more fundamental issue; how choice options themselves are represented in the brain.

6.3.2 Focus on representation over process

A manner in which experimental complexity can be reduced even further is by eliminating the focus on the algorithms that lead to choice altogether. Focusing solely on choice option representations, independently of decision strategy, is what the study in chapter 5

presented. Eliminating the focus on choice algorithms leads the discussion away from decision-making research into the realm of research on cognitive representations proper. The neural representation of choice options is at the intersection of the two research programs (decision-making and cognitive representations). The work presented here focused mainly on this intersection but it is worth mentioning that the study of cognitive representations has its own academic history. For example, the topics covered in chapters 2 and 3 such as sense of self, control, and theory of mind, have all been studied independently of issues related to decision-making (Baumeister, 2002; Frith & Frith, 2005; Harter, 1999). The same can be said of psychophysics studies that study the boundary conditions of cognitive representations (Gescheider, 2013) or old debates on the nature of mental imagery (Kosslyn, 1980; Pylyshyn, 1973).

Thus, the deeper issue of how choice options are represented neurally was addressed in chapter 5. There the findings showed the Minkowski measures performed best in the study composed of stimuli of simple geometric shapes. This finding makes sense if the stimuli were designed to be easily represented in a Euclidean feature space (i.e., there are obvious predefined axes of variation) as was the case with this study. In contrast, the Mahalanobis measure provided a better fit in the study with stimuli of natural images. Such images have no predefined feature space (for individual stimuli), which can lead to the importance of the covariance structure. Increased stimuli complexity could make the covariance structure between classes more relevant for brain computations of similarity. This idea is supported by the property of the Mahalanobis measure that makes it sensitive to covariance. The datasets also differed in the task participants were doing; 1-back for the NI study and a categorisation task for the GS study. It is possible that the difference in task could also influence the particular use of a similarity measure.

The relationship between feature selection and similarity measure are intricately linked (Medin, Goldstone, & Gentner, 1993). For naturalistic stimuli, constructing features can impose constraints on candidate similarity measures. Complementarily, if one supposes an agent is using a specific similarity measure then features that are consistent with the respective pairwise similarity matrices can be inferred. This interactive dependency between features and similarity measure poses a chicken or egg problem. What makes using neural activations as features for a stimulus valuable is that the complexity of feature selection is greatly reduced; the complexity is reduced in the sense that features can be spatially clustered and selected

through a cross-validated optimisation procedure (as performed for the study of neural representation of choice options), whereas this feat is considerably harder for natural images in pixel space (cf. convolutional neural networks, LeCun & Bengio, 1995). Another interesting property of feature spaces with respect to similarity measures, is that the number of features (i.e., the feature space dimensionality) is necessarily bounded. This is due to the curse of dimensionality which imposes a limit on dimensions given that similarity measures lose discriminative power in high dimensions; all points have comparable similarity between any other points in the space (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999; Friedman, J., Hastie, T., & Tibshirani, R., 2001). One can interpret the curse of dimensionality as imposing a limit on what is representable in the brain. Discovering whether this provides a useful upper bound on the dimensionality of cognitive representations is a recommended path for future studies.

The evaluation of similarity models (i.e., similarity measures) in chapter 5 provided privileged access to the study of representations; representations understood as static and morphological perspectives on information structures. Understanding representations in this way is useful but has its pitfalls; specifically, how should the temporal window be defined when studying choice option representations? For example, the Mahalanobis measure requires many internal computations, including the estimation of a covariance matrix. Thus, the representational view is unfolded into an algorithmic (i.e., processual) perspective once again, showing the inseparability of representation from algorithm. Furthermore, the temporal window under which choice option representations are studied is constrained by the measurement technology (i.e., fMRI). Perhaps other technologies such as electrophysiology, magnetoencephalography (MEG), or electroencephalography (EEG) provide more flexibility in this regard. Other questions arise regarding whether the representations that are used for the internal computations of the similarity model are indeed the correct ones (i.e, if beta weights for voxel activations are the correct representation or if there is a superior one, like *t* statistics, for example).

The claim was made that similarity models are a class of models that is best suited for the study of choice option representations. However, similarity models do not have to be conceptualised solely at the representational and algorithmic level; similarity models can be conceptualised at the computational level too (Lin, 1998). As mentioned in chapter 1, the distance axioms for similarity measures can be interpreted as a normative model of similarity representation. The case is made most clearly in the context of class separability where the

distance axioms provide the semantics under which class separability can be understood. Even more interesting is the interpretation that probabilities themselves can be understood as similarity measures in an event space (as opposed to a feature space as is the case in chapter 5), which presents an interesting link with Bayesian inference.

The analogy between optimal similarity measure and optimal class separation can be understood most clearly in the context of object recognition. Suppose an agent has a task which is to distinguish hammers from non-hammers (i.e., the classes are hammer and non-hammers). The optimal Bayes classifier would be given by computing the probability of the object being a hammer conditional on the observed features (Giraud, 2014), where unit probability would be equal to maximum similarity – or minimum distance – from the centre of mass of that conditional distribution. The problem with computing this optimal classifier is that the conditional distribution is not known so assumptions must be made about the nature of the distribution, such as parametric or smoothness assumptions of the distribution or classification function, respectively (Giraud, 2014).

Different assumptions about the probability event space will imply different similarity measures for the feature space. The intractability in computing the optimal Bayes classifier provides the justification for focusing on the optimal separating hyperplane instead of the conditional distribution. The optimal separating hyperplane separates the classes while maximizing the distance to the closest point from either class (Vapnik, 2013). Thus, the linear SVM constitutes a natural choice which sidesteps the intractability of the Bayes classifier by computing the optimal hyperplane for linearly separable classes. Although the methods in chapter 5 were based on a black-box decoding approach (i.e., picking the best classifier in terms of accuracy), these characteristics may help explain why linear SVM works well with this kind of data.

Nonetheless, linear separability is still a strong assumption (imposed by the linear SVM) which can be overcome by kernelised SVMs. Unfortunately, this approach brings the us back to the problem of making parametric assumptions, this time in the form of a suitable kernel. The linear separability between classes for fMRI data is interesting in itself, although there is confidence from previous studies that this indeed seems to be the case since linear SVMs work well for this kind of data (Misaki, Kim, Bandettini, & Kriegeskorte, 2010; Pereira, Mitchell, & Botvinick, 2009). It is still an open question if this linear separability represents a

true underlying brain function (i.e., due to the way downstream neurons read out information) or if it is just an artefact of the pre-processing steps for fMRI data (due to the application of general linear models prior to classification, for example).

6.4 Limitations of the studies

Below there is a section for each study in this work, denoting the limitations of each study in turn. A limitation that is common to all studies presented here is that all of them focus too much on the visual modality. The visual modality is preferred because it has rich structure and can be experimentally controlled with great detail. Of course, this limitation could be said of decision-making studies in general. The same criticism applies to relying heavily on verbal explanations of the experimental task. For the representation of choice options, both relying on verbal explanations and visual modality, is an adequate starting point but should not hinder progress in the study of other modalities. All studies as a group were meant to give different perspectives on the phenomenon of choice option representations through Marr's three levels of abstraction and the bottom-up/top-down approaches. However, if one would make the ideal study, such a study would seek to address all perspectives at once (if such a study is possible).

6.4.1 Delegation study

The limitations for the study presented in chapter 2 consisted mainly of the difficulty in accurately controlling for risk aversion which can contribute to the concavity of the utility function and enhance the size of the effect observed for the control premium. Although there was a gambling task that estimated risk aversion parameters, they were still estimated for a separate task and could be different for the delegation task. Providing full information during the delegation task (i.e., letting participants see the stimuli before choosing to delegate) could be described as undesirable given that many delegation choices in real world environments do not have full information up front. Furthermore, not providing more information on the characteristics of the advisor could be seen as a weakness given that a choice to delegate is highly specific to the individual you will delegate a choice to. Another weakness of the study is that perhaps participants might be more willing to delegate in circumstances where the effort involved in deliberating on the initial choice is too high (e.g., like thinking about tax returns or letting an accountant do them for you).

6.4.2 Preference integration study

A major weakness for the preference integration study in chapter 3 is the specification and fitting of the Bayesian model; the ratio between model parameters to observations is very high. Alarmingly so, this could mean that the model is overfitting the data which may well be true. However, the correlation of the KL divergence with the second confidence rating somewhat reduces this concern. The same can be said about the correlation of this quantity with dmPFC activity. The true objective of the model was not to generalise the behavioural data but to compute descriptive statistics that would reflect the underlying assumptions of Bayesian integration. Therefore, such statistics, like the KL divergence could be used as tools for probing certain brain ROIs in the hopes that correlation with activity in such areas would provide evidence for Bayesian integration of subjective value in the brain.

6.4.3 Heuristic studies

What was explained as a good feature of the design for the study in chapter 4 could actually be interpreted as a weakness. Instructed strategy use was meant as a methodological tool that would enable fine-tuned control of the decision strategy being implemented. However, researchers may be upset with how artificial this type of procedure seems, totally disconnected from any type of real world decision-making type scenario. Of course, instructed strategy use limits the possible comparisons made with other studies that rely on spontaneous use of decision strategies.

6.4.4 Neural representation of choice options study

The limitations for chapter 5 are mainly to do with its scope of permissible generalisations. The comparison of just two studies to claim that change in task or stimuli reveals a change in similarity measure by the brain should be taken with caution. More replications are needed and perhaps further fine-tuning of the comparison procedure between measures is warranted. For example, there is bias associated with similarity measures whenever there is noise present. There are ways of removing this bias in a cross-validated sense that may or may not have an impact on the performance of certain similarity measures (Walther et al., 2016). Another added complication is one that is common to many multivariate decoding approaches in fMRI; spatially correlated noise. This was seen as a strength in the output of the analysis where the findings showed that most ROIs use the same similarity measure. However,

at the same time this spatial correlation could make interpretation of this result more complicated. Although the analysis procedure used cross-validation to try to average out the effect of this noise, it is possible that there are still spatial correlations of the noise that are driving the results (Diedrichsen, Ridgway, Friston, & Wiestler, 2011). Finally, as mentioned at the end of chapter 5, although it is assumed that the differences between the GS and the NI study are due to effects of task and stimulus, other confounds such as differences in data quality (including issues concerning study design), cohort effects, or differences in fMRI equipment cannot be ruled out.

6.5 Recommendations for further research

A common theme throughout this work was trying to bridge between different levels of analysis and perspectives. Previously, others have recommended bridging between the computational and algorithmic levels (Griffiths et al., 2015) or between the algorithmic and implementational levels (Love, 2015, 2016). Complementarily, it has been argued here that the top-down and bottom-up approaches (focus on task goals or stimuli, respectively) are a different dimension that also needs to be addressed in the study of choice option representations.

Additionally, the argument was made that there is an alternating advantage to describing a cognitive phenomenon both in terms of process or in terms of representation. The *static* view provided by the representational perspective (i.e., the focus on the geometry of the information structures) should be studied at finer spatial and temporal resolutions. Hopefully, this path will lead to the construction of richer algorithmic perspectives that consider all the interesting transformations performed on data structures in the brain. For example, you could imagine using the 2x3 framework that was introduced here to study psychiatric conditions where individual differences in performance on a task could provide insights on how the cognitive representations are being adapted in these vulnerable populations. An example of this was mentioned as a future direction in the delegation study, where it's possible that depressive participants could show more optimal delegations than controls given that their control premium might be diminished as a by-product of their depressive symptoms; are their representations of the task different from controls or just the way they construct their cognitive representations and implement algorithms?

Marr's levels of abstraction also guide further research in the direction of the implementational issues, a topic that has been only scratched on the surface in this work. Although, the study on the neural representation of choice option similarity relations is a step in this direction. The analysis and findings of that study could benefit from more specific implementational details. For example, constructing a theory that specifies how a network of neurons (e.g., a leaky integrate and fire neuronal network) could even compute the Mahalanobis measure of similarity would bring this line of research closer to the implementation level. Other avenues of research could be deconstructing the task structure in more anterior brain areas (Schuck, Cai, Wilson, & Niv, 2016) or detailing the relationships between internal choice option representations and their external representation in a physical substrate (Kirsh, 2010).

New methodological tools from graph theory (Bassett & Sporns, 2017) and algebraic topology (Reimann et al., 2017) look promising in providing a framework that relates neural network structures to neural network function. From the point of view of evaluating similarity measures for choice option representations, the most adequate way of comparing neural network activations is not obvious. This is a general problem in graph theory since there is still no general consensus on the best way to compare networks (cf. Berlingerio, Koutra, Eliassi-Rad, & Faloutsos, 2013).

Other remaining open questions regarding similarity relations in the brain would be to investigate if different sensory modalities or different task demands would result in more varied similarity profiles for different brain regions. Relatedly, probing the different distance axioms in the brain seems like a natural extension of the psychological research (Tversky, 1977) that has already been accomplished in this spirit. Testing whether asymmetrical similarity measures are relevant for choice option representations is one suggestion (e.g., China is not as similar to North Korea than North Korea is to China).

Another natural progression for research on the role of choice option representations would be to test the influence of such representations on the computation of value and uncertainty. Would a stimulus set that requires Minkowski similarity representations affect value and confidence computations differently than a stimulus set that requires a measure such as Mahalanobis? How do neural representations of choice options influence value and confidence computations?

Also, the debate between which class of decision strategies best describes human decision-making (e.g., utility maximizing, Bayesian, compensatory like TAL, or non-compensatory like TTB) is still relevant to the proposal of the importance of choice option representations. The interaction between choice option representations and the broader context of a task will surely bring forth findings that delineate the boundary conditions for each class of decision strategies. An important challenge will be to match decision strategies to different environments according to multiple factors, such as stimulus format, task goals, time available, and cognitive resources.

6.6 On a more philosophical note

Trade-offs between task goals and stimuli can either be designed experimentally or fall out of natural relationships in the real-world. For example, humans created a goal for themselves which was to travel to the moon. The implications of such a task constrained the space of useful materials (i.e., stimuli) with which such a task could be achieved. Thus, finding the right materials and algorithms to achieve the task was constrained by the material implications (physical laws) of the task itself. Another way of seeing this is that to construct a task goal, it will necessarily refer to the physical and material space in which it will be realised. The work presented here assumed fixed tasks and stimuli but natural behaviours also allow for the freedom of choosing the task and the stimuli that the agent wishes to work with. On occasion, humans are free to construct a task based on the available materials, engaging in a bricolage of sorts (Lévi-Strauss, 1962).

This dissertation focused mainly on representations of choice options; representations framed in the context of a decision-making process. The issue of representations more generally has been treated by all the humanities and social sciences in one way or another. In philosophy, discussions go back to Aristotle (Hicks, 2015). Modern debates challenge the notion of representation itself, preferring the view that mental states are inherently intentional (i.e., only subserve behavioural goals) and are not semantically evaluable objects (Brentano, 1973; Dennett, 1989). Representations are also studied through social and cultural contexts (Leach, 1976; Turner, 1967) and as an intrinsic part of communication and linguistics (see semiology or semiotics in De Saussure, 2011; Peirce, 1974). These disciplines, in addition to acknowledging the effect that physical laws have on trade-offs between task and stimuli, emphasise the effect that representations have on each other through social processes.

In this work, I highlighted that transformations between representations should be understood as *process* at the algorithmic level. In cognitive science, we are accustomed to focus on singleton agents and the processes that occur confined therein. However, there is no reason not to extend our algorithmic theories in cognitive science to the multi-agent context. Hutchin's proposal on distributed cognition (Hutchins, 2010) is a step in this direction, which underlines the view that algorithmic processes are off-loaded into the physical and social environment (e.g., writing a book or engaging in dialogue, respectively). Focus on the social environment displays that the trade-off between task goals and stimuli is not only dictated by physical properties but also by socially relevant characteristics (e.g., suppose the task goal is seducing your future partner).

6.7 Concluding thoughts

Taken together, the findings of this thesis help provide a more complete understanding of the trade-off managed between stimuli and task goals by choice option representations. Specifically, this work aimed to show that task goals and stimuli are important interactive determinants of choice option representations that constrain these representations simultaneously. The findings indicate that personal beliefs and preferences are adapted to those of other people through the adjustment of beliefs regarding the self or of personal preferences (depending on the task), as well as confidence in that valuation. The adjustment of choice option representations is observed both in a task that establishes the self as a choice option and in a task that establishes a property of the self (personal preferences), with respect to retail products, as the choice option to be adjusted. Furthermore, the findings in this thesis show that decision strategies which operate over choice options, function differently depending on stimulus format and that stimulus format can directly influence choice option representations in the human brain. Overall, it is clear that choice option representations play a pivotal role in cognitive and decision-making processes, both from a bottom-up and a top-down perspective. Evidently, choice option representations need to be understood across all of Marr's three levels of abstraction (implementation, algorithm/representation, and computational goal). Constructing a complete theory of choice option representations will involve bridging across all three levels for dynamic decision-making processes starting from the initial presentation of task goals and stimulus set, up until the completion of said goals.

Appendices

A Supplementary materials of chapter 2

A.1 Summary of instructions given to participants

This experiment consists of 3 stages:

The first stage of the experiment is a training task. On each trial two shapes will be presented and your task will be to choose the shape that will deliver a better outcome. Your task will be to discover the rules, if any, that make some shapes better than others. There are two kinds of trials in this part of the experiment. In one kind of trial you will have the opportunity of winning 10 pounds or nothing. In the other type of trial, you will either lose 10 pounds or not lose any money at all. Instructions will be given onscreen, including which buttons you need to press during the experiment. I (the experimenter) will be outside should you have any questions during the task.

The second stage of the experiment is divided into two parts. On each trial, you will decide first whether to select between two shapes or to delegate the choice to an advisor. It is important to mention that these advisors are in fact artificial agents, whose advice was determined by a computer algorithm prior to the start of the experiment. On each trial, there will be a different advisor and the advice they offer is drawn from the responses that have already been computed previously. You will be presented with two pieces of information regarding the advisor: the advisor's success rate in the task and their fee. You will only pay this fee if the advisor selects the correct shape. If you choose the advisor, then the corresponding advisor will make the choice for you. If you decide to make the choice yourself, then you will just choose between the shapes as you did in the first stage of the experiment. It is also important to mention that the choices you make in this second part of the experiment will have an impact on your final compensation. Ten trials will be sampled at random and (depending on the choices you made) we will average the winnings and losses and add that to your final compensation. Other instructions, including which buttons you need to press during the experiment, will be given onscreen. I (the experimenter) will be outside should you have any questions during the task.

The third stage of the experiment is a gambling task. In this task, you must decide whether you want to participate in a 50-50 gamble or choose the “sure” option. The sure option will normally be an amount of money that is lower than what could be gained with the gamble option. Similarly, as in the second task, ten trials will be sampled at random and we will average the winnings and losses to add that to your final compensation. Other instructions, including which buttons you need to press during the experiment, will be given onscreen. I (the experimenter) will be outside should you have any questions during the task.

B Supplementary materials of chapter 4

B.1 List of cues (statistics for developing countries) with adjectives

"Competitiveness in medium enterprises"	(lower, higher)
"Price stability in cheap basic goods"	(less stable, more stable)
"Increased employment opportunities"	(lesser, greater)
"Public investment in infrastructure"	(smaller, larger)
"Decreased rates of infectious diseases"	(worse, better)
"Increased life expectancy for women"	(inferior, superior)
"Development of civic participation"	(weaker, stronger)

B.2 Stimuli sampling procedure

Classifying trials based on Take-the-Best difficulty results in Q1, Q2, Q3, Q4, Q5 trials (i.e. trials where looking at the first cue is sufficient, looking at the second is sufficient, etc.). Because only Q1, Q2 & Q3 trials were used, this reduces the space to 468 trials.

Classifying trials based on Tallying difficulty results in trials with Δ1, Δ2, Δ3, Δ4 & Δ5 difference between the amount of negatively-valued (red cross) and positively-valued (green checkmark) cues. Since only Δ1, Δ2 & Δ3 trials were used this reduces the space to 462 trials. Cross-tabulating Tallying & Take-the-Best trial spaces results in a space of 452 trials (Table B.1).

Table B.1
Cross tabulation between TAL & TTB trial types

		TAL		
		Δ3	*Δ2*	*Δ1*
TTB	*Q1*	40	100	180
	Q2	10	30	60
	Q3	2	8	20

In addition to balanced sampling across the two different difficulty classifications, the amount of non-discriminating cues can also affect difficulty in both experimental conditions, (whether participants use TAL or TTB). Trials range from having seven cues with discriminating power down to only three cues with discriminating power. Trials were sampled with the most diversity possible without repetition of stimuli (Table B.2).

Table B.2
Cross tabulation between TAL & TTB trial types sampled

		TAL			
		Δ3	*Δ2*	*Δ1*	*Total*
	Q1	6	3	3	12
TTB	*Q2*	5	5	2	12
	Q3	1	4	7	12
	Total	12	12	12	36

This results in a total of 36 trials that can be partitioned into 3 blocks of Q1, Q2, Q3 trials or equivalently as 3 blocks of Δ3, Δ2, or Δ1 trials respectively. Now adding the trials where the signs are reversed -for right-left presentation instead of left-right presentation of the stimuli- gives a total of 72 trials.

Globally, all 72 trials can also be partitioned into trials where all cues can discriminate between options (n = 8), trials where six cues can discriminate between options (n = 8), trials where five cues can discriminate between options (n = 28), trials where four cues can discriminate between options (n = 16) and trials where three cues can discriminate between options (n = 12). This can be seen in Table B.3.

Table B.3
Cross tabulation of trial difficulty with amount of non-discriminating cues for TAL & TTB

Tallying					**Take-the-Best**				
# of Equals Signs	**Δ3**	**Δ2**	**Δ1**	**Total**	**# of Equals Signs**	**Q1**	**Q2**	**Q3**	**Total**
0	6	0	2	6	0	8	0	0	6
1	0	8	0	8	1	4	4	0	8
2	18	0	10	28	2	8	12	8	28
3	0	16	0	16	3	2	6	8	16
4	0	0	12	12	4	2	2	8	12
Total	24	24	24	72	**Total**	24	24	24	72

B.3 Cue subset models for TAL & TTB for Heuristic Study 1

B.3.1 Subset models for the time pressure phase

In addition to comparing participants' vector of responses to what the respective heuristic predicted for each condition, these responses were also compared to what these heuristics would predict for subsets of the cues. These models were applied exclusively to the time pressure phase as shown in Figure B.1. The models were classified in a two by two factorial design (heuristic model x model class), giving four model classes in total for each experimental condition. The first factor consisted of modeling participant responses for TAL subsets (n = 128) and for TTB subsets (n = 13700) for both conditions to see whether TAL subsets could better account for the data in the TTB condition or vice versa. The second factor (model class) was constructed based on the way cues were defined, either featural or positional. The featural models represent tracking cues based on their names (e.g. "Public investment in infrastructure") and the positional models represent tracking just the positions of the cues (e.g. just looking at the first three cues). A 2x2x2 ANOVA (experimental condition x heuristic model x model class) for five-fold cross-validated predictive accuracy of the models showed a main effect of model class $F_{1,177} = 172.86$, $p < 0.001$, $\eta^2 = 0.49$, a main effect of heuristic $F_{1,177} = 305.49$, $p < 0.001$, $\eta^2 = 0.63$, and of their interaction $F_{1,177} = 326.32$, $p < 0.001$, $\eta^2 = 0.65$. There was also an effect of experimental condition $F_{1,177} = 774.59$, $p < 0.001$, $\eta^2 = 0.81$, its interaction with heuristic model $F_{1,177} = 233.36$, $p < 0.001$, $\eta^2 = 0.57$, and the three-way interaction of model class with heuristic model and experimental condition $F_{1,177} = 79.15$, $p < 0.001$, $\eta^2 = 0.31$.

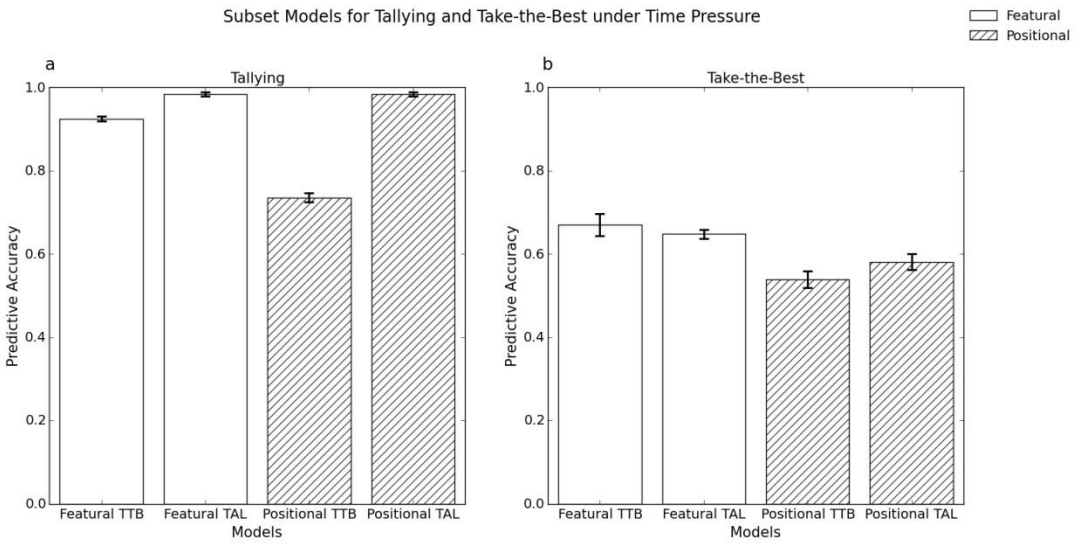

Figure B.1 Subset models under time pressure.

On the left is (a) mean predictive accuracy from five-fold cross validation for the four classes of subset models for TAL in the time pressure phase and on the right is (b) the same for TTB. Error bars are within-participant confidence intervals.

It is important to note that distinguishing between Positional TAL models and Featural TAL models was not possible due to equivalent predictions of both classes of models when participants were using the full set of cues. This section is concluded with an exposition of the actual models considered for this analysis.

Number of TAL Subset Models:

$$|TAL| = 2^n = 128 \tag{B.1}$$

where n is the total number of cues ($n = 7$).

Number of TTB Subset Models:

$$|TTB| = \sum_{k=1}^{n=7} \binom{n}{k} k! = 13700 \tag{B.2}$$

Doubling the number of models to distinguish between positional and featural models, and accounting for duplicates, gives a total of 27639 subset models per participant.

B.4 Practice phase analysis for Heuristic Studies 1 and 2

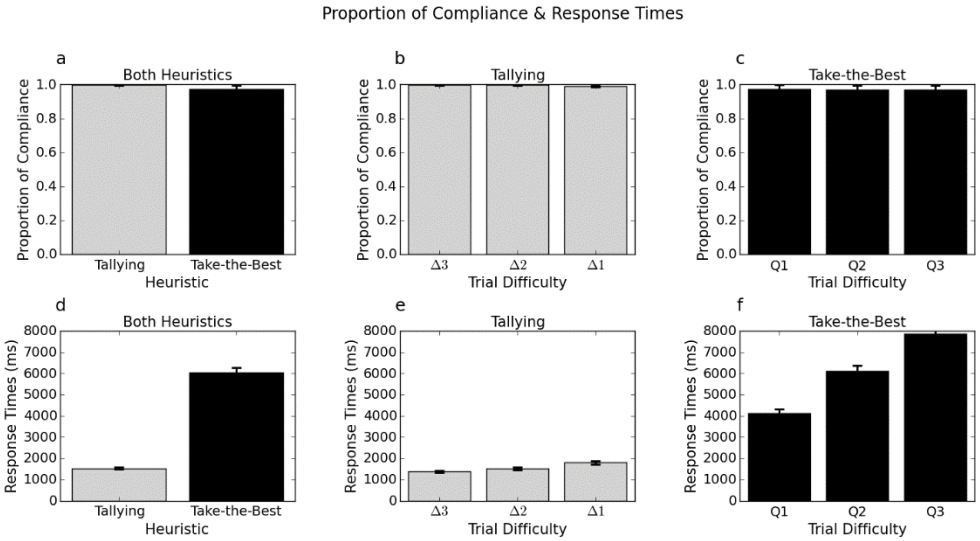

Figure B.2 Results for practice phase of Heuristic Study 1 after applying exclusion criteria (*n* = 179).

The figure shows the proportion of compliance with (a) both heuristics with TAL in gray and TTB in black, (b) the proportion of compliance in the TAL condition displayed by degrees of difficulty (Deltas), and (c) the proportion of compliance in the TTB condition displayed by degrees of difficulty (Qs). Shows the response times for (d) both heuristics, (e) the response times in the TAL condition displayed by degrees of difficulty (Deltas), and (f) the response times in the TTB condition displayed by degrees of difficulty (Qs). For all panels, error bars are 95% within-participants confidence intervals.

B.4.1 Heuristic Study 1

Proportion of compliance was analysed using a 2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB) and a within-participants factor of trial difficulty (3 levels of trial difficulty). Main effects were observed for heuristic, $F_{1, 177} = 74.29$, $p < 0.001$, $\eta^2 = 0.296$. However, effects were not observed for trial difficulty $F_{2, 354} = 1.40$, $p = 0.248$, $\eta^2 = 0.008$, nor of the interaction between trial difficulty and heuristic, $F_{2, 354} = 0.767$, $p = 0.465$, $\eta^2 = 0.004$. These results suggest that participants had a slightly harder time learning TTB although both heuristics were close to ceiling (see panel a of Figure B.2). Interestingly, trial difficulty was not a relevant factor (with respect to compliance) during the practice phase (see panels b and c of Figure B.2).

Response times were analysed using a 2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB) and with a within-participants factor of trial difficulty (3 levels of trial difficulty). Main effects were observed for heuristic, $F_{1, 177} = 456.69$, $p < 0.001$, $\eta^2 = 0.721$ and trial difficulty, $F_{2, 354} = 398.59$, $p < 0.001$, $\eta^2 = 0.692$, as well as for the interaction of heuristic and trial difficulty, $F_{2, 354} = 253.75$, $p < 0.001$, $\eta^2 = 0.589$. These results confirm the fact that TTB was harder to implement in this experiment given the slower response times (see panel d of Figure B.2). Furthermore, trial difficulty did not affect response times for TAL (panel e of Figure B.2) but did affect response times for TTB (panel f of figure B.2).

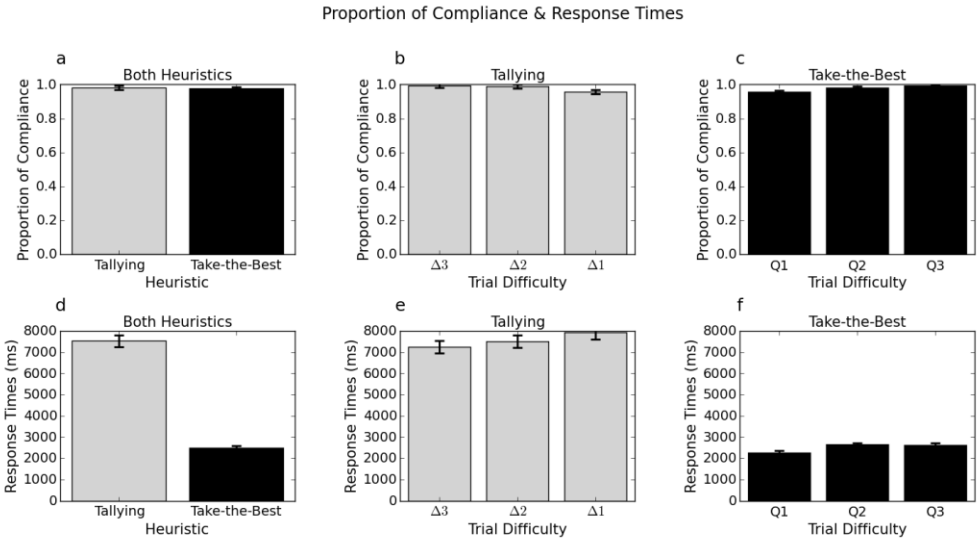

Figure B.3 Results for practice phase of Heuristic Study 2 after applying exclusion criteria (n = 173).

The figure shows the proportion of compliance with (a) both heuristics with TAL in gray and TTB in black, (b) the proportion of compliance in the TAL condition displayed by degrees of difficulty (Deltas), and (c) the proportion of compliance in the TTB condition displayed by degrees of difficulty (Qs). Shows the response times for (d) both heuristics, (e) the response times in the TAL condition displayed by degrees of difficulty (Deltas), and (f) the response times in the TTB condition displayed by degrees of difficulty (Qs). For all panels, error bars are 95% within-participants confidence intervals.

B.4.2 Heuristic Study 2

Proportion of compliance was analysed using a 2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB) and a within-participants factor of trial difficulty (3 levels of trial difficulty). Main effects were not observed for heuristic, $F_{1, 171} = 0.295$, $p = 0.588$, $\eta^2 = 0.002$. However, effects were not observed for trial difficulty $F_{2, 342} = 5.87$, $p = 0.003$, $\eta^2 = 0.033$, and of the interaction between trial difficulty and heuristic, $F_{2, 342}$

= 60.59, $p < 0.001$, $\eta^2 = 0.26$. These results resemble the fact that participants adequately learned the use of both heuristics and that trial difficulty affected their learning (see panels a, b and c of Figure B.3). The fact that TTB shows better compliance with harder trials is unexpected but is also in line with the results seen during the test phase.

Response times were analysed using a 2x3 mixed-design ANOVA with a between-participant factor of heuristic (TAL, TTB) and with a within-participants factor of trial difficulty (3 levels of trial difficulty). Main effects were observed for heuristic, $F_{1, 171} = 309.58$, $p < 0.001$, $\eta^2 = 0.644$ and trial difficulty, $F_{2, 342} = 21.33$, $p < 0.001$, $\eta^2 = 0.111$, as well as for the interaction of heuristic and trial difficulty, $F_{2, 342} = 3.94$, $p = 0.02$, $\eta^2 = 0.022$. These results confirm the fact that TAL was harder to implement in this experiment given the slower response times (see panel d of Figure B.3).

C Supplementary materials of chapter 5

C.1 Regions of interest from the Harvard-Oxford atlas

All regions from the Harvard-Oxford atlas were included in the analyses except for cerebral white matter, the lateral ventricles, left and right cerebral cortex, and the brain stem. This results in 48 cortical regions and 7 subcortical regions; doubling for lateralisation results in the 110 regions of interest.

C.1.1 Cortical regions of interest

Frontal Pole

Insular Cortex

Superior Frontal Gyrus

Middle Frontal Gyrus

Inferior Frontal Gyrus, pars triangularis

Inferior Frontal Gyrus, pars opercularis

Precentral Gyrus

Temporal Pole

Superior Temporal Gyrus, anterior division

Superior Temporal Gyrus, posterior division

Middle Temporal Gyrus, anterior division

Middle Temporal Gyrus, posterior division

Middle Temporal Gyrus, temporooccipital part

Inferior Temporal Gyrus, anterior division

Inferior Temporal Gyrus, posterior division

Inferior Temporal Gyrus, temporooccipital part

Postcentral Gyrus

Superior Parietal Lobule

Supramarginal Gyrus, anterior division

Supramarginal Gyrus, posterior division

Angular Gyrus

Lateral Occipital Cortex, superior division

Lateral Occipital Cortex, inferior division

Intracalcarine Cortex

Frontal Medial Cortex

Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)

Subcallosal Cortex

Paracingulate Gyrus

Cingulate Gyrus, anterior division

Cingulate Gyrus, posterior division

Precuneous Cortex

Cuneal Cortex

Frontal Orbital Cortex

Parahippocampal Gyrus, anterior division

Parahippocampal Gyrus, posterior division

Lingual Gyrus

Temporal Fusiform Cortex, anterior division

Temporal Fusiform Cortex, posterior division

Temporal Occipital Fusiform Cortex

Occipital Fusiform Gyrus

Frontal Operculum Cortex

Central Opercular Cortex

Parietal Operculum Cortex

Planum Polare

Heschl's Gyrus (includes H1 and H2)

Planum Temporale

Supracalcarine Cortex

Occipital Pole

C.1.2 Subcortical regions of interest

Thalamus, Caudate, Putamen, Pallidum, Hippocampus, Amygdala, Accumbens

C.2 Task descriptions and acquisition parameters

C.2.1 Geometric shapes (GS) study

The GS study presented sixteen objects in total, which varied on four different binary features: (colour: red or green, shape: circle or triangle, size: large or small, and position: right or left). Participants in this study were trained to do a categorisation task. They were first trained on five objects of one category and four of the other (nine objects total during training) with twenty repetitions of each object. During the anatomical scan, participants saw four more repetitions of the training items as a refresher. Then during the functional scanning phase, participants were asked to categorise the nine familiar objects they saw during the training phase and seven novel objects they had not seen before. Each trial during the functional scanning phase lasted 10 seconds; 3.5 seconds where one of the sixteen objects (nine training stimuli and seven novel transfer stimuli) was presented after which a fixation cross was presented for 6.5 seconds. No feedback was provided during this phase. Each stimulus was presented three times within a run across six runs resulting in each stimulus being presented a total of eighteen times during the functional scanning phase – except for one participant who only participated in five runs of the scanning phase.

Whole-brain imaging data were acquired on a 3.0T GE Signa MRI system (GE Medical Systems). Structural images were acquired using a T2-weighted flow-compensated spin-echo pulse sequence (TR=3s; TE=68ms, 256x256 matrix, 1x1mm in-plane resolution) with thirty-three 3-mm thick oblique axial slices (0.6mm gap), approximately 20° off the AC-PC line. Functional images were acquired with an echo planar imaging sequence using the same slice prescription as the structural images (TR=2s, TE=30.5ms, flip angle=73°, 64x64 matrix, 3.75x3.75 in-plane resolution, bottom-up interleaved acquisition, 0.6mm gap). An additional high-resolution T1-weighted 3D SPGR structural volume (256x256x172 matrix, 1x1x1.3mm voxels) was acquired for registration and cortex parcellation.

C.2.2 Natural images (NI) study

The NI study presented fifty-four objects in total, which varied in two ways. The 54 stimulus items were conceived to either be organised by category (6 categories: minerals, animals, fruits/vegetables, music, sports, or tools) or by their silhouette (9 silhouettes) which cut orthogonally across the category distinction. Participants in this study were asked to perform a 1-back real-world size judgment task (i.e., to respond according to whether the object on the previous trial was larger or smaller than the current image on screen). Participants were scanned on two separate sessions (different days). Each session consisted of eight functional scanning runs resulting in sixteen runs total – except for one participant for which four of the runs of the first session were lost due to scanning issues. Each one of the fifty-four objects were presented twice within each run in a randomised sequence. This resulted in each object being presented a total of thirty-two times (or twenty-four times for the participant that only had twelve runs). On each trial, each object was presented for 1.5 seconds after which a fixation cross was presented for 1.5 seconds. Each run started with a fixation cross for fourteen seconds and ended with a fixation cross for fourteen seconds. Thirty-six fixation trials lasting three seconds each were also randomly presented within each run.

Data collection was performed on a 3T Philips scanner with a 32-channel coil at the Department of Radiology of the University Hospitals Leuven. MRI volumes were collected using echo planar (EPI) T2*-weighted scans. Acquisition parameters were as follows: repetition time (TR) of 2 s, echo time (TE) of 30 ms, flip angle (FA) of 90°, field of view (FoV) of 216 mm, and matrix size of 72x72. Each volume comprised 37 axial slices (covering the whole brain) with 3 mm thickness and no gap. The T1-weighted anatomical images were acquired with an MP-RAGE sequence, with 1x1x1 mm resolution.

C.3 Similarity measures

Only similarity measures that presented a mean Spearman correlation within three median absolute deviations away from the group average (group refers to distances here) were presented in the Results section. Measures that did not meet these criteria were considered outliers. The median Spearman correlation across the 18 similarity measures evaluated was 0.203 for the GS study 0.125 and for the NI study and their median absolute deviation was 0.0482 for the GS study and 0.0234 for the NI study. The mean Spearman correlations (across participants) and the standard deviations for the measures that were more than three median

absolute deviations away from the group average were: Bhattacharya without covariance matrix regularisation (mean = 0.001 and s.d. = 0.004 for the GS study, mean = 0.0002 and s.d. = 0.0006 for the NI study), Bhattacharya (d) (with diagonal regularisation) (mean = -0.0005 and s.d. = 0.003 for the GS study, mean = -0.0001 and s.d. = 0.0007 for the NI study), variance of information (mean = -0.04 and s.d. = 0.037 for the GS study, mean = -0.012 and s.d. = 0.004 for the NI study), and distance correlation (mean = -0.037 and s.d. = 0.026 for the GS study, mean = -0.0009 and s.d. = 0.0038 for the NI study). These statistics were computed across the 110 original ROIs. Below is a list of the equations for each measure considered.

For two classes represented as vectors

$$X = (x_1, x_2, ..., x_n) \text{ and } Y = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$$

where each component is computed as the arithmetic mean across *m* observations (trial-by-trial β coefficients) per class, per run, and *n* is the number of voxels. This notation is valid except for where these vectors show subscripts denoting individual observations as opposed to mean vectors (this is only the case when discussing distance correlation).

C.3.1 Dot product

$$XY^T \tag{C.1}$$

C.3.2 Cosine measure

$$\frac{XY^T}{\|X\|_2\|Y\|_2} \tag{C.2}$$

where $\|\cdot\|_2$ denotes the L2 (Euclidean) norm.

C.3.3 Minkowski measure

The (negative) Minkowski measure is:

$$-\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p} \tag{C.3}$$

For the city-block distance p = 1, for the Euclidean distance p = 2, and for the Chebyshev distance $p = \infty$.

C.3.4 Pearson correlation

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{C.4}$$

where $\bar{x}$ and $\bar{y}$ are the component-wise arithmetic means of vectors X and Y, respectively.

C.3.5 Spearman correlation

$$1-\frac{6\sum_{i=1}^{n}(rg(x_i)-rg(y_i))^2}{n(n^2-1)} \tag{C.5}$$

where $rg(x_i)$ and $rg(y_i)$ are the ranks of the values x_i and y_i, respectively.

C.3.6 Mahalanobis measure

The (negative) Mahalanobis measure between two random vectors coming from the same multivariate normal distribution is:

$$-\sqrt{(X-Y)^T\Sigma^{-1}(X-Y)} \tag{C.6}$$

where Σ is the n x n covariance matrix between voxels.

C.3.7 Bhattacharyya measure

The (negative) Bhattacharyya measure between two multivariate normal distributions $\mathcal{N}(X,\Sigma_X)$ and $\mathcal{N}(X,\Sigma_Y)$, where each voxel covariance matrix Σ_X and Σ_Y is estimated separately for each class X and Y, respectively, is:

$$-\left(\frac{1}{8}(X-Y)^T\bar{\Sigma}^{-1}(X-Y)+\frac{1}{2}ln\left(\frac{det\bar{\Sigma}}{\sqrt{det\Sigma_X det\Sigma_Y}}\right)\right) \tag{C.7.1}$$

where

$$\bar{\Sigma}=\frac{\Sigma_X+\Sigma_Y}{2} \tag{C.7.2}$$

C.3.8 Distance correlation

$$\frac{dCov(X,Y)}{dVar(X)dVar(Y)} \tag{C.8.1}$$

where $dCov^2(X,Y)$ is

$$\frac{1}{m^2}\sum_{j=1}^{m}\sum_{k=1}^{m}A_{j,k}Bj,k \tag{C.8.2}$$

where $dVar^2(X)$ is

$$\frac{1}{m^2}\sum_{j=1}^{m}\sum_{k=1}^{m}A_{j,k}^2 \tag{C.8.3}$$

where $A_{j,k}$ is the matrix computed from doubly-centreing the matrix $a_{j,k}$ (subtracting row and column means while adding the grand mean), where

$$a_{j,k}=||X_j-X_k||_2 \tag{C.8.4}$$

Thus, $B_{j,k}$ is computed from $b_{j,k}$, where

$$b_{j,k}=||Y_j-Y_k||_2 \tag{C.8.5}$$

These pairwise distance matrices are computed from distances between observations.

C.3.9 Variation of information

For two classes X and Y represented as two multivariate Gaussian distributions, the (negative) Variation of information is

$$-VI(X;\ Y) = -(H(X) + H(Y) - 2I(X;Y)) \tag{C.9.1}$$

where *H(X)* is the entropy of *X* and *I(X;Y)* is the mutual information between *X* and *Y*.

For a multivariate Gaussian *X*, *H(X)* is

$$\frac{1}{2}ln((2\pi e)^n det\Sigma_X) \tag{C.9.2}$$

and the mutual information between *X* and *Y* is

$$\frac{1}{2}ln(\frac{det\Sigma_X det\Sigma_Y}{det\Sigma^*}) \tag{C.9.3}$$

where Σ^*

$$= \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \tag{C.9.4}$$

and Σ_{XY} is the between-class voxel covariance matrix. Σ_{YX} is the transpose of Σ_{XY}.

C.4 Covariance matrix regularisation

C.4.1 Diagonal regularisation

Diagonal regularisation for a covariance matrix $\sum$ was computed as $\sum \circ I$, where $\circ$ is the Hadamard product (element-wise multiplication) and I is the identity matrix.

C.4.2 Ledoit-Wolf regularisation

Ledoit-Wolf regularisation for a covariance matrix $\sum$ was computed as (1 - shrinkage) $\sum$ + (shrinkage)(μ)I, where $\mu = trace(\sum)/n$ and the optimal shrinkage parameter is a value between 0 and 1 estimated according to the derivation in (Ledoit & Wolf, 2004).

References

Allefeld, C., & Haynes, J. D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage, 89*, 345–357. http://doi.org/10.1016/j.neuroimage.2013.11.043

Alloy, L. B., & Clements, C. M. (1992). Illusion of control: invulnerability to negative affect and depressive symptoms after laboratory and natural stressors. *Journal of Abnormal Psychology*, *101*(2), 234.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews. Neuroscience*, *7*(4), 268.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Angelaki, D. E., Gu, Y., & DeAngelis, G. C. (2009). Multisensory integration: psychophysics, neurophysiology, and computation. *Current Opinion in Neurobiology*, *19*(4), 452–458.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*(1), 124.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews. Neuroscience*, *7*(5), 358.

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*(1), 13–20.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge: Cambridge University Press.

Bargh, J. A. (1994). The four horsemen of automaticity: Intention, awareness, efficiency, and control as separate issues. *Handbook of social cognition, 1*, 1–40.

Barron, H. C., Garvert, M. M., & Behrens, T. E. J. (2015). Reassessing VMPFC: Full of confidence? *Nature Neuroscience*, *18*(8), 1064–1066. http://doi.org/10.1038/nn.4076

Bartlett, M. S. (1951). The effect of standardization on a χ 2 approximation in factor analysis. *Biometrika*, *38*(3/4), 337–344.

Bartling, B., Fehr, E., & Herz, H. (2014). The intrinsic value of decision rights. *Econometrica, 82*(6), 2005-2039.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823*.

Baumeister, R. F. (2002). Ego depletion and self-control failure: An energy model of the self's executive function. *Self and Identity, 1*(2), 129–136.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 58–80.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47*(1), 129–141.

Beattie, J., Baron, J., Hershey, J. C., & Spranca, M. D. (1994). Psychological determinants of decision attitude. *Journal of Behavioral Decision Making*, *7*(2), 129–144. http://doi.org/10.1002/bdm.3960070206

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science, 324*(5931), 1160–1164.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. http://doi.org/10.1038/nature07538

Bell, D. E., & Raiffa, H. (1988). *Decision making: Descriptive, normative, and prescriptive interactions*. Cambridge: Cambridge University Press.

Benassi, V. A., Sweeney, P. D., & Dufour, C. L. (1988). Is there a relation between locus of control orientation and depression? *Journal of Abnormal Psychology, 97*, 357-367.

Ben Zur, H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica, 47*(2), 89–104. http://doi.org/10.1016/0001-6918(81)90001-9

Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(1), 107–129. http://doi.org/10.1037/0278-7393.33.1.107

Berlingerio, M., Koutra, D., Eliassi-Rad, T., & Faloutsos, C. (2013). Network similarity via multiple social theories. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on* (pp. 1439–1440).

Berns, G. S., Capra, C. M., Moore, S., & Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *NeuroImage, 49*(3), 2687–2696. http://doi.org/10.1016/j.neuroimage.2009.10.070

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International conference on database theory* (pp. 217–235).

Bishop, C. (2007). *Pattern Recognition and Machine Learning* (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. New York, NY: Springer.

Bown, N. J., Read, D., & Summers, B. (2003). The lure of choice. *Journal of Behavioral Decision Making*, *16*(4), 297.

Bracci, S., & de Beeck, H. O. (2016). Dissociations and associations between shape and

category representations in the two visual pathways. *Journal of Neuroscience*, *36*(2), 432–444.

Brehm, S. S., & Brehm, J. W. (2013). *Psychological reactance: A theory of freedom and control*. Cambridge, MA: Academic Press.

Brentano, F. (2014). *Psychology from an empirical standpoint*. London: Routledge.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*(1), 2–16.

Bröder, A. (2003). Decision making with the "adaptive toolbox": influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(4), 611–625. http://doi.org/10.1037/0278-7393.29.4.611

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review*, *14*(5), 895–900. http://doi.org/10.3758/BF03194118

Bröder, A., & Schiffer, S. (2003). Take the best versus simultaneous feature matching: probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology. General*, *132*(2), 277–293. http://doi.org/10.1037/0096-3445.132.2.277

Bröder, A., & Schiffer, S. (2006). Stimulus format and working memory in fast and frugal strategy selection. *Journal of Behavioral Decision Making*, *19*(4), 361–380. http://doi.org/10.1002/bdm.533

Bromberg-Martin, E. S., & Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. *Nature Neuroscience*, *14*(9), 1209–1216.

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860–1862.

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. *Blackwell Handbook of Judgment and Decision Making*, 133–154.

Calhoun, V. D., Liu, J., & Adal\i, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, *45*(1), S163--S172.

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology*, *20*(13), 1165–1170. http://doi.org/10.1016/j.cub.2010.04.055

Campbell-Meiklejohn, D. K., Kanai, R., Bahrami, B., Bach, D. R., Dolan, R. J., Roepstorff, A., & Frith, C. D. (2012). Structure of orbitofrontal cortex predicts social influence. *Current Biology*, *22*(4), R123–R124. http://doi.org/10.1016/j.cub.2012.01.012

Canguilhem, G. (2012). *On the Normal and the Pathological* (Vol. 3). New York, NY: Springer.

Catania, A. C. (1981). Freedom of choice: A behavioral analysis. *The Psychology of Learning and Motivation*, *14*, 97–145.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010). Separate channels for processing form, texture, and color: evidence from fMRI adaptation and visual object agnosia. *Cerebral Cortex*, *20*(10), 2319–2332.

Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of Affective Decision Making How Do Feelings Predict Choice? *Psychological Science*, *27*(6), 763–775.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291. http://doi.org/10.1016/j.tics.2006.05.007

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A., & Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, *10*(Mar), 747–776.

Churchland, P. M. (1993). State-space semantics and meaning holism. *Philosophy and Phenomenological Research*, *53*(3), 667–672.

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. http://doi.org/10.1146/annurev.psych.55.090902.142015

Clark, T. S., & Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods*, *3*(2), 399–408. http://doi.org/10.1017/psrm.2014.32

Clithero, J. A., & Rangel, A. (2013). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. http://doi.org/10.1093/scan/nst106

Cockburn, J., Collins, A. G. E., & Frank, M. J. (2014). A Reinforcement Learning Mechanism Responsible for the Valuation of Free Choice. *Neuron*, *83*(3), 551–557. http://doi.org/10.1016/j.neuron.2014.06.035

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *106*(23), 9163–9168. http://doi.org/10.1073/pnas.0807721106

Corter, J. E. (1987). Similarity, confusability, and the density hypothesis. *Journal of Experimental Psychology: General*, *116*(3), 238.

Costello, F., & Watts, P. (2014). Surprisingly rational: probability theory plus noise explains

biases in judgment. *Psychological Review*, *121*(3), 463.

Craig, A. D., & Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1).

Crump, M. J. C., McDonnell, J. V, & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In *Simple heuristics that make us smart* (pp. 97–118). New York, NY: Oxford University Press.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. http://doi.org/10.1038/nature04766

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience* (Vol. 806). Cambridge, MA: MIT Press.

De Finetti, B. (2017). *Theory of probability: A critical introductory treatment* (Vol. 6). London: John Wiley & Sons.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.

De Saussure, F. (2011). *Course in general linguistics*. New York, NY: Columbia University Press.

Deichmann, R., Gottfried, J. A., Hutton, C., & Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage*, *19*(2), 430–441.

Dennett, D. C. (1989). *The intentional stance*. Cambridge, MA: MIT press.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A Meta-analysis of Functional Neuroimaging Studies of Self- and Other Judgments Reveals a Spatial Gradient for Mentalizing in Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *24*(8), 1742–1752. http://doi.org/10.1162/jocn_a_00233

Diedrichsen, J., & Kriegeskorte, N. (2016). *Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis*. *bioRxiv*. http://doi.org/10.1101/071472

Diedrichsen, J., Ridgway, G. R., Friston, K. J., & Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: A pattern-component model. *NeuroImage*, *55*(4), 1665–1678. http://doi.org/10.1016/j.neuroimage.2011.01.044

Domenech, P., & Koechlin, E. (2015). Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, *1*, 101–106.

http://doi.org/10.1016/j.cobeha.2014.10.007

Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486. http://doi.org/10.1126/science.1252254

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review*, *115*(1), 199–213. http://doi.org/10.1037/0033-295X.115.1.199

Durstewitz, D., Vittoz, N. M., Floresco, S. B., & Seamans, J. K. (2010). Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, *66*(3), 438–448. http://doi.org/10.1016/j.neuron.2010.03.029

Egan, L. C., Bloom, P., & Santos, L. R. (2010). Choice-induced preferences in the absence of choice: Evidence from a blind two choice paradigm with young children and capuchin monkeys. *Journal of Experimental Social Psychology*, *46*(1), 204–207.

Egan, L. C., Santos, L. R., & Bloom, P. (2007). The origins of cognitive dissonance evidence from children and monkeys. *Psychological Science*, *18*(11), 978–983.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 201602413.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. http://doi.org/10.1038/415429a

Fechner, G. (1966). Elements of psychophysics. Vol. I. Transl. HE Adler, ed. DH Howes, EG Boring. New York: Rinehart & Winston. From German

Fehr, E., Herz, H., & Wilkening, T. (2012). The Lure of Authority : Motivation and Incentive Effects of Power The Lure of Authority : Motivation and Incentive Effects of Power, *103*(November), 1325–1359. http://doi.org/10.1257/aer.103.4.1325

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. http://doi.org/10.1016/j.tics.2004.05.002

Fishburn, P. C. (1967). Methods of estimating additive utilities. *Management Science*, *13*(7), 435–453.

Flandin, G., & Friston, K. J. (2016). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *arXiv Preprint arXiv:1606.08199*.

Fodor, J. A., & Lepore, E. (1992). Holism: A shopper's guide. Oxford: Blackwell.

Fodor, J., & Lepore, E. (1999). All at sea in semantic space: Churchland on meaning similarity. *The Journal of Philosophy*, *96*(8), 381–403.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1).

Springer series in statistics. Berlin: Springer.

Friedman, M. (1953). *Essays in positive economics*. Chicago, IL: University of Chicago Press.

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

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology, 15*(17), R644--R645.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind." *Trends in Cognitive Sciences, 7*(2), 77–83. http://doi.org/10.1016/S1364-6613(02)00025-6

Garzon, F. C. (2000). State space semantics and conceptual similarity: reply to Churchland. *Philosophical Psychology*, *13*(1), 77–95.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevelhierarchical models* (Vol. 1). New York, NY: Cambridge University Press.

Gerber, A. J., Posner, J., Gorman, D., Colibazzi, T., Yu, S., Wang, Z., … Peterson, B. S. (2008). An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces. *Neuropsychologia*, *46*(8), 2129–2139.

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. http://doi.org/10.1126/science.aac6076

Gescheider, G. A. (2013). *Psychophysics: the fundamentals*. Hove, England: Psychology Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482. http://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*(4), 684.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart* (pp. 3–34). Oxford: Oxford University Press.

Gigerenzer, G., Todd, P. M., & the ABC Group. (1999). *Simple heuristics that make us smart*. (G. Gigerenzer, P. M. Todd, & the ABC Group, Eds.). Oxford: Oxford University Press.

Gläscher, J. (2009). Visualization of group inference data in functional neuroimaging. *Neuroinformatics, 7*(1), 73–82.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Glimcher, P. W. (2008). Neuroeconomics. *Scholarpedia*, *3*(10), 1759.

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision making and the brain*.

Cambridge, MA: Academic Press.

Glockner, A., & Betsch, T. (2008). Multiple-Reason Decision Making Based on Automatic Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1055. http://doi.org/10.1037/0278-7393.34.5.1055

Glöckner, A., & Betsch, T. (2012). Decisions beyond boundaries: When more information is processed faster than less. *Acta Psychologica*, *139*(3), 532–542. http://doi.org/10.1016/j.actpsy.2012.01.009

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*(6666), 481.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, *130*(5), 769.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Guest, O., & Love, B. C. (2017). What the success of brain imaging implies about the neural code. *Elife*, *6*, e21397.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, *105*(18), 6741–6746. http://doi.org/10.1073/pnas.0711099105

Hancock, E., & Pelillo, M. (2011). *Similarity-Based Pattern Recognition*. New York, NY: Springer.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V, & Pollmann, S. (2009). PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, *324*(5927), 646–648.

Harter, S. (1999). *The construction of the self: A developmental perspective.* New York, NY: Guilford Press.

Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013). Topographic representation of numerosity in the human parietal cortex. *Science*, *341*(6150), 1123–1126.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258.

Haxby, J. V, Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, (June), 435–456. http://doi.org/10.1146/annurev-neuro-062012-170325

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939. http://doi.org/10.1038/nn.2856

Henson, R. N. A., Rugg, M. D., Shallice, T., Josephs, O., & Dolan, R. J. (1999). Recollection and familiarity in recognition memory: an event-related functional magnetic resonance imaging study. *Journal of Neuroscience*, *19*(10), 3962–3972.

Hicks, R. D. (2015). *Aristotle De Anima*. Cambridge, MA: Cambridge University Press.

Higgins, E. T. (2005). Value from regulatory fit. *Current Directions in Psychological Science*, *14*(4), 209–213.

Holden, M., Hill, D. L. G., Denton, E. R. E., Jarosz, J. M., Cox, T. C. S., Rohlfing, T., … Hawkes, D. J. (2000). Voxel similarity measures for 3-D serial MR brain image registration. *IEEE Transactions on Medical Imaging*, *19*(2), 94–102.

Howe, C. Q., & Purves, D. (2005). The M{ü}ller-Lyer illusion explained by the statistics of image--source relationships. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1234–1239.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging* (Vol. 1). Sunderland, MA: Sinauer Associates Sunderland.

Hume, D. (1793). *An inquiry concerning human understanding* (Vol. 3). Oxford: Oxford University Press.

Hutchins, E. (2010). Cognitive Ecology. *Topics in Cognitive Science*, *2*(4), 705–715. http://doi.org/10.1111/j.1756-8765.2010.01089.x

Ito, M. (1989). Long-term depression. *Annual Review of Neuroscience*, *12*(1), 85–102.

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, *78*(3), 563–573. http://doi.org/10.1016/j.neuron.2013.03.023

Izuma, K., Akula, S., Murayama, K., Wu, D.-A., Iacoboni, M., & Adolphs, R. (2015). A causal role for posterior medial frontal cortex in choice-induced preference change. *Journal of Neuroscience*, *35*(8), 3598–3606.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Jennrich, R. I. (1970). An asymptotic $χ$2 test for the equality of two correlation matrices. *Journal of the American Statistical Association*, *65*(330), 904–912.

Jones, E., Oliphant, T., Peterson, P., & others. (n.d.). {SciPy}: Open source scientific tools for {Python}. Retrieved from http://www.scipy.org/

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.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607. http://doi.org/10.1016/S0364-0213(02)00083-6

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard--easy effect. *Psychological Review*, *107*(2), 384.

K\Hoszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, *121*(4), 1133–1165.

Kahneman, D., Slovic, P., & Tversky, A. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, *246*, 160-173.

Khader, P. H., Pachur, T., & Jost, K. (2013). Automatic activation of attribute knowledge in heuristic inference from memory. *Psychonomic Bulletin & Review*, *20*(2), 372–7. http://doi.org/10.3758/s13423-012-0334-7

Khader, P. H., Pachur, T., Meier, S., Bien, S., Jost, K., & Rösler, F. (2011). Memory-based Decision-making with Heuristics: Evidence for a Controlled Activation of Memory Representations. *Journal of Cognitive Neuroscience*, *23*(11), 3540–3554. http://doi.org/10.1162/jocn_a_00059

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309. http://doi.org/10.1152/jn.00024.2007

Kirsh, D. (2010). Thinking with external representations. *Ai & Society*, *25*(4), 441–454.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement Learning Signal Predicts Social Conformity. *Neuron*, *61*(1), 140–151. http://doi.org/10.1016/j.neuron.2008.11.027

Klucharev, V., Munneke, M. A. M., Smidts, A., & Fernández, G. (2011). Downregulation of

the posterior medial frontal cortex prevents social conformity. *Journal of Neuroscience*, *31*(33), 11934–11940.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.

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. http://doi.org/10.1016/j.conb.2015.12.007

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. http://doi.org/10.1038/nature02169

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

Kray, L. (2000). Contingent Weighting in Self-Other Decision Making. *Organizational Behavior and Human Decision Processes*, *83*(1), 82–106. http://doi.org/10.1006/obhd.2000.2903

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4. http://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*(2), 180.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*, 445-463.

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *arXiv Preprint arXiv:1604.00289*.

Lane, R. D., Chua, P. M. L., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia*, *37*(9), 989–997.

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76–94.

Leach, E. (1976). *Culture and Communication: the logic by which symbols are connected. An introduction to the use of structuralist analysis in social anthropology*. Cambridge, MA: Cambridge University Press.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167. http://doi.org/10.1038/nn.4064

LeCun, Y., Bengio, Y., & others. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10), 1995.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review*, *11*(2), 343–352. http://doi.org/10.3758/BF03196581

Leonards, U., Sunaert, S., Van Hecke, P., & Orban, G. A. (2006). Attention mechanisms in visual search—an fMRI study. *Journal of Cognitive Neuroscience*, *12*, 61-75.

Leotti, L. A., & Delgado, M. R. (2014). The Value of Exercising Control Over Monetary Gains and Losses. *Psychological Science*, *25*(2), 596–604. http://doi.org/10.1177/0956797613514589

Leotti, L. A., Iyengar, S. S., & Ochsner, K. N. (2010). Born to choose: The origins and value of the need for control. *Trends in Cognitive Sciences*, *14*(10), 457–463.

Leotti, L. a, & Delgado, M. R. (2011). The inherent reward of choice. *Psychological Science*, *22*(10), 1310–8. http://doi.org/10.1177/0956797611417005

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. http://doi.org/10.1016/j.conb.2012.06.001

Lieberman, M. D., Ochsner, K. N., Gilbert, D. T., & Schacter, D. L. (2001). Do amnesics exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science*, *12*(2), 135–140.

Lin, D., & others. (1998). An information-theoretic definition of similarity. In *Icml* (Vol. 98, pp. 296–304).

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, *227*(4693), 1435–1441.

Lisman, J. E., Fellous, J.-M., & Wang, X.-J. (1998). A role for NMDA-receptor channels in working memory. *Nature Neuroscience*, *1*(4), 273–275.

Locke, J. (1975). An essay concerning human understanding, ed. PH Nidditch. Oxford: Clarendon Press.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science*, *14*(4), 195–199. http://doi.org/10.1111/j.0963-7214.2005.00363.x

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, *7*(2), 230–242.

Love, B. C. (2016). Cognitive models as bridge between brain and behavior. *Trends in Cognitive Sciences*, *20*(4), 247–248.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309–332. http://doi.org/10.1037/0033-295X.111.2.309

Lynch, M. (1990). The similarity index and DNA fingerprinting. *Molecular Biology and Evolution*, *7*(5), 478–484.

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*(20), 2023–2027. http://doi.org/10.1016/j.cub.2013.08.035

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General*, *111*(1), 1.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375.

McFadden, D. (2001). Economic choices. *The American Economic Review*, *91*(3), 351–378.

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron*, *84*(4), 870–881. http://doi.org/10.1016/j.neuron.2014.10.013

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254.

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Computational Biology*, *11*(6), 1–25. http://doi.org/10.1371/journal.pcbi.1004305

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. http://doi.org/10.1016/j.neuron.2015.09.039

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *hlt-Naacl* (Vol. 13, pp. 746–751).

Monti, M. M. (2011). Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Frontiers in Human Neuroscience*, *5*.

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*(4), 406–419. http://doi.org/10.1037/a0024377

Mozer, M., & Baldwin, D. (2008). Experience-Guided Search: A Theory of Attentional Control. *Advances in Neural Information Processing Systems 20*, 1033–1040. http://doi.org/10.1037/e527342012-076

Mumford, J. A., Poline, J.-B., & Poldrack, R. A. (2015). Orthogonalization of regressors in fMRI models. *PLoS One*, *10*(4), e0126255.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636–2643.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT press.

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences*, *9*(1), 11–15. http://doi.org/10.1016/j.tics.2004.11.005

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(1), 53–65. http://doi.org/10.1037/0278-7393.29.1.53

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best." *Organizational Behavior and Human Decision Processes*, *91*(1), 82–96. http://doi.org/10.1016/S0749-5978(02)00525-3

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. J. (2012). An Agent Independent Axis for Executed and Modeled Choice in Medial Prefrontal Cortex. *Neuron*, *75*(6), 1114–1121. http://doi.org/10.1016/j.neuron.2012.07.023

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4). http://doi.org/10.1371/journal.pcbi.1003553

Norman, K. A., & O'reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-

voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005

Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517–524.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(1), 25–53.

O'Reilly, J. X., Jbabdi, S., & Behrens, T. E. J. (2012). How can a Bayesian approach inform neuroscience? *European Journal of Neuroscience*, *35*(7), 1169–1179. http://doi.org/10.1111/j.1460-9568.2012.08010.x

O'Reilly, J. X., Schuffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38), E3660–E3669. http://doi.org/10.1073/pnas.1305373110

Oaksford, M., & Chater, N. (1998). *Rational models of cognition*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357. http://doi.org/10.1016/S1364-6613(00)01699-5

Over, D. E. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 3–18). Malden, MA: Blackwell Publishing

Owens, D., Grossman, Z., & Fackler, R. (2014). The control premium: A preference for payoff autonomy. *American Economic Journal: Microeconomics*, *6*(4), 138–161. http://doi.org/10.1257/mic.6.4.138

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 376–404. http://doi.org/10.1167/5.5.1

Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology, 74*, 59-6.

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: a meta-analysis of research findings. *Psychological Bulletin*, *134*(2), 270.

Pavlov, I. P., & Anrep, G. V. (2003). *Conditioned reflexes*. North Chelmsford, MA: Courier Corporation.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The adaptive decision maker. *The Adaptive Decision Maker*, *45*(7), 352. http://doi.org/10.1057/jors.1994.133

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2013). The neural representation of unexpected uncertainty during value-based decision making. *Neuron*,

79(1), 191–201.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Peirce, C. S. (1974). *Collected papers of charles sanders peirce* (Vol. 5). Cambridge, MA: Harvard University Press.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classi!ers and fMRI: A tutorial overview. *NeuroImage*, *45*, S199–S209. http://doi.org/10.1016/j.neuroimage.2008.11.007.Machine

Peters, E., Dieckmann, N. F., Västfjäll, D., Mertz, C. K., Slovic, P., & Hibbard, J. H. (2009). Bringing meaning to numbers: the impact of evaluative categories on decisions. *Journal of Experimental Psychology. Applied*, *15*(3), 213–227. http://doi.org/10.1037/a0016978

Platzer, C., & Bröder, A. (2012). Most people do not ignore salient invalid cues in memory-based decisions. *Psychonomic Bulletin & Review*, *19*(4), 654–661. http://doi.org/10.3758/s13423-012-0248-4

Platzer, C., Bröder, A., & Heck, D. W. (2014). Deciding with the eye: How the visually manipulated accessibility of information in memory influences decision behavior. *Memory & Cognition*, *42*(4), 595–608.

Polman, E. (2010). Information distortion in self-other decision making. *Journal of Experimental Social Psychology*, *46*(2), 432–435. http://doi.org/10.1016/j.jesp.2009.11.003

Posner, M. I., & Presti, D. E. (1987). Selective attention and cognitive control. *Trends in Neurosciences*, *10*(1), 13–17. http://doi.org/10.1016/0166-2236(87)90116-0

Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J. R., & Hampton, J. A. (2015). Progress and current challenges with the quantum similarity model. *Frontiers in Psychology*, *6*.

Pothos, E. M., & Busemeyer, J. R. (2014). In search for a standard of rationality. *Frontiers in Psychology*, *5*.

Pothos, E. M., Busemeyer, J. R., Shiffrin, R. M., & Yearsley, J. M. (2017). The rational status of quantum cognition. *Journal of Experimental Psychology: General*, *146*(7), 968.

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, *120*(3), 679.

Pothos, E. M., & Trueblood, J. S. (2015). Structured representations in a quantum probability model of similarity. *Journal of Mathematical Psychology*, *64*, 35–43.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.

Pratte, M. S., & Tong, F. (2017). Integrating theoretical models with functional neuroimaging. *Journal of Mathematical Psychology*, *76*, 80–93.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 497–527.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, *68*(5), 1281–1292.

Ramsey, F. P. (1931). Truth and probability (1926). *The Foundations of Mathematics and Other Logical Essays*, 156–198.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews. Neuroscience*, *9*(7), 545–556.

Rao, H., Korczykowski, M., Pluta, J., Hoang, A., & Detre, J. A. (2008). Neural correlates of voluntary and involuntary risk taking in the human brain: An fMRI Study of the Balloon Analog Risk Task (BART). *NeuroImage*, *42*(2), 902–910. http://doi.org/10.1016/j.neuroimage.2008.05.046

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Reimann, M. W., Nolte, M., Scolamiero, M., Turner, K., Perin, R., Chindemi, G., … Markram, H. (2017). Cliques of Neurons Bound into Cavities Provide a Missing Link between Structure and Function. *Frontiers in Computational Neuroscience*, *11*, 48.

Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. *Medical Decision Making*, *28*(6), 850-865.

Reynolds, J. H., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron*, *37*(5), 853–863.

Riefer, P. S., Prior, R., Blair, N., Pavey, G., & Love, B. C. (2017). Coherency-maximizing exploration in the supermarket. *Nature Human Behaviour*, *1*, 17.

Rieskamp, J., & Otto, P. E. (2006). SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Rolls, E. T., Grabenhorst, F., & Deco, G. (2010). Choice, difficulty, and confidence in the brain. *NeuroImage*, *53*(2), 694–706. http://doi.org/10.1016/j.neuroimage.2010.06.073

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1987). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT press.

Russo, J. E., & Dosher, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 676–696. http://doi.org/10.1037/0278-7393.9.4.676

Rutledge, R. B., De Berker, A. O., Espenhahn, S., Dayan, P., & Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*.

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience*, *30*(40), 13525–13536.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2015). Dopaminergic modulation of decision making and subjective well-being. *Journal of Neuroscience*, *35*(27), 9811–9822.

Rutledge, R. B., Smittenaar, P., Zeidman, P., Brown, H. R., Adams, R. A., Lindenberger, U., … Dolan, R. J. (2016). Risk taking for potential reward decreases across the lifespan. *Current Biology*, *26*(12), 1634–1639.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science*, *310*(5752), 1337–1340. http://doi.org/10.1126/science.1115270

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. http://doi.org/10.1007/BF00055564

Griffiths, T. L., Navarro, D. J., & Sanborn, A. N. (2006, January). A More Rational Model of Categorization. In *Proceedings of the Cognitive Science Society* (Vol. 28, No. 28).

Chicago Scellier, B., & Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, *11*.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *12*(5), 1207–1245.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*(6), 1402–1412.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(June 1994), 1593–1599. http://doi.org/10.1126/science.275.5306.1593

Shapiro Jr, D. H., Schwartz, C. E., & Astin, J. A. (1996). Controlling ourselves, controlling our world: Psychology's role in understanding positive and negative consequences of seeking and gaining control. *American Psychologist*, *51*(12), 1213.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941--R945.

Sharot, T., De Martino, B., & Dolan, R. J. (2009). How choice reveals and shapes expected hedonic outcome. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(12), 3760–3765. http://doi.org/10.1523/JNEUROSCI.4972-08.2009

Sharot, T., Riccardi, A. M., Raio, C. M., & Phelps, E. a. (2007). Neural mechanisms mediating optimism bias. *Nature*, *450*(7166), 102–105. http://doi.org/10.1038/nature06280

Sharot, T., Shiner, T., & Dolan, R. J. (2010). Experience and choice shape expected aversive outcomes. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(27), 9209–9215. http://doi.org/10.1523/JNEUROSCI.4770-09.2010

Sharot, T., Velasquez, C. M., & Dolan, R. J. (2010). Do Decisions Shape Preference?: Evidence From Blind Choice. *Psychological Science*, *21*(9), 1231–1235. http://doi.org/10.1177/0956797610379235

Shepard, R. N., & others. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237–250.

Simon, H. A. (1978). Rationality as Process and as Product of a Thought. *American Economic Review*, *68*(2), 1. http://doi.org/10.2307/1816653

Simon, H. A. (1982). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). Cambridge, MA: MIT press.

Sjöström, J., & Gerstner, W. (2010). Spike-timing dependent plasticity. *Spike-Timing Dependent Plasticity*, *35*.

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: why we never think alone*. New York, NY: Penguin.

Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L., & Beckmann, C. F. (2014). Group-PCA for very large fMRI datasets. *NeuroImage*, *101*, 738–749.

Soucy, E. R., Albeanu, D. F., Fantana, A. L., Murthy, V. N., & Meister, M. (2009). Precision and diversity in an odor map on the olfactory bulb. *Nature Neuroscience*, *12*(2), 210–220.

Spence, K. W. (1952). The nature of the response in discrimination learning. *Psychological Review*, *59*(1), 89.

Stone, E. R., & Allgaier, L. (2008). A Social Values Analysis of Self–Other Differences in Decision Making Involving Risk. *Basic and Applied Social Psychology*, *30*(2), 114–129. http://doi.org/10.1080/01973530802208832

Strauss, C. L. (1962). *Savage mind*. Chicago, IL: University of Chicago Press.

Studer, B., Apergis-Schoute, A. M., Robbins, T. W., & Clark, L. (2012). What are the odds? The neural correlates of active choice during gambling. *Frontiers in Neuroscience*, *6*(APR), 1–16. http://doi.org/10.3389/fnins.2012.00046

Sturm, T. (2012). The "rationality wars" in psychology: Where they are and where they could go. *Inquiry*, *55*(1), 66–81.

Summerfield, C., & Koechlin, E. (2008). A Neural Representation of Prior Information during Perceptual Inference. *Neuron*, *59*(2), 336–347. http://doi.org/10.1016/j.neuron.2008.05.021

Suzuki, S. (1997). Effects of number of alternatives on choice in humans. *Behavioural Processes*, *39*(2), 205–214. http://doi.org/10.1016/S0376-6357(96)00049-6

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., … Nakahara, H. (2012). Learning to Simulate Others' Decisions. *Neuron*, *74*(6), 1125–1137. http://doi.org/10.1016/j.neuron.2012.04.030

Syed, E. C. J., Grima, L. L., Magill, P. J., Bogacz, R., Brown, P., & Walton, M. E. (2015). Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nature Neuroscience*, *19*(December), 1–6. http://doi.org/10.1038/nn.4187

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*(1), 235.

Tervo, D. G. R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & Karpova, A. Y. (2014). Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*, *159*(1), 21–32. http://doi.org/10.1016/j.cell.2014.08.037

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences*, *94*(26), 14792–14797.

Thompson, K. G., & Bichot, N. P. (2005). A visual salience map in the primate frontal eye field. *Progress in Brain Research*, *147*, 249–262.

Thompson, S. C. (1999). Illusions of control: How we overestimate our personal influence. *Current Directions in Psychological Science*, *8*(6), 187–190. http://doi.org/10.1111/1467-8721.00044

Todd, P. M., & Gigerenzer, G. (2000). Précis of Simple heuristics that make us smart. *The Behavioral and Brain Sciences*, *23*(5), 727-741-780. http://doi.org/10.1017/S0140525X00003447

Trueblood, J. S., Pothos, E. M., & Busemeyer, J. R. (2014). Quantum probability theory as a common framework for reasoning and similarity. *Frontiers in Psychology*, *5*.

Turner, V. W. (1967). *The forest of symbols: Aspects of Ndembu ritual* (Vol. 101). Ithaca, NY: Cornell University Press.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*(2), 123–154. http://doi.org/10.1037/0033-295X.89.2.123

Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making* (pp. 141–162). New York, NY: Springer.

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.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *The American Economic Review*, 204–217.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142.

van Rossum, M. C. W. (2001). A novel spike distance. *Neural Computation*, *13*(4), 751–763.

Vapnik, V. (2013). *The nature of statistical learning theory*. New York, NY: Springer.

Vilares, I., & Kording, K. (2011). Bayesian models: The structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*(1), 22–39. http://doi.org/10.1111/j.1749-6632.2011.05965.x

Von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton, NJ: Princeton university press.

Wagner, A. D., Maril, A., Bjork, R. A., & Schacter, D. L. (2001). Prefrontal contributions to executive control: fMRI evidence for functional distinctions within lateral prefrontal cortex. *Neuroimage*, *14*(6), 1337–1347.

Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General*, *144*(1), 7.

Waldfogel, J. (1993). The deadweight loss of Christmas. *The American Economic Review*,

83(5), 1328–1336.

Weber, M., & Osherson, D. (2010). Similarity and induction. *Review of Philosophy and Psychology*, *1*(2), 245–264.

Weiskopf, N., Hutton, C., Josephs, O., & Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. *Neuroimage*, *33*(2), 493–504.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *15*(3), 419–433. http://doi.org/2527952

Xing, E. P., Jordan, M. I., Russell, S. J., & Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems* (pp. 521–528).

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B*, *367*(1594), 1310–1321.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience*, *25*(11), 3002–3008.

Yoshida, W., & Ishii, S. (2006). Resolution of Uncertainty in Prefrontal Cortex. *Neuron*, *50*(5), 781–789. http://doi.org/10.1016/j.neuron.2006.05.006

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. http://doi.org/10.1016/j.neuron.2005.04.026

Zucker, R. S. (1989). Short-term synaptic plasticity. *Annual Review of Neuroscience*, *12*(1), 13–31.