Constructing Value in the Medial Prefrontal Cortex

Keno Jüchems

A dissertation submitted in partial fulfilment of the requirements for the degree of *Doctor of Philosophy* in the University of Oxford

Wolfson College

Trinity Term 2019

Constructing Value in the Medial Prefrontal Cortex

Keno Jüchems

A dissertation submitted in partial fulfilment of the requirements for the degree of *Doctor of Philosophy* in the University of Oxford.

Wolfson College, Oxford Trinity Term 2019

Approximate word count: 48,000

Humans and other animals are remarkably adept at integrating a vast array of information into their choices. For instance, they seamlessly integrate their own internal states and goals (e.g. being thirsty or aspiring to obtain a PhD) into their actions without extensive training. How does the brain enable them to do this efficiently? In this thesis, I will investigate three key components of this behaviour: i) how humans track accumulation of resources as information about the current state of their goals, ii) how they arbitrate between two competing goals, and iii) how remembered value information is efficiently re-weighted to support pursuit of novel goals. All reported experiments employed functional magnetic resonance imaging to localize neural responses.

In the first line of research, covering points i) and ii), I will review two experiments, which investigate the neural correlates in the prefrontal cortex (PFC) during goal-directed decision-making. In Chapter 3, I report results from a study in which participants accumulated rewards from risky gambles over time. A key value-related area in the prefrontal cortex (PFC), the ventromedial PFC, encoded both the momentary receipt of reward and the participants' current cumulative rewards. Next, in Chapter 4, I ask how humans arbitrate between two competing goals based on such latent accumulation signals. Another PFC region, the anterior cingulate cortex (ACC), reflected both the pressure to act on one goal over another as well as participants' likelihood to act according to this pressure. Importantly, the vmPFC encoded how much participants' choices redressed the imbalance between goals, indicating that its value correlates may represent an integration of multiple goals and momentary receipt. If goals exert pressure to re-direct choices towards certain courses of action, then the areas supporting such decision-making must be able to flexibly re-weight information according to these new goals. Thus, in Chapter 5, I present a study showing that the frontopolar cortex as well as the ACC are key areas supporting this dynamic re-weighting and the representation of opportunity costs. Lastly, I will review a framework of primary rewards, homeostatic reinforcement learning, and how it may be adapted to encompass cognitive goals.

This work was funded by a 3-year Advanced Quantitative Methods studentship from the Economic and Social Research Council.

Statement of previous acceptance of thesis for degree

The following part of the thesis has been accepted for the degree of M.Sc. in Cognitive Science at the University of Osnabrück, Germany, on 16/07/2015:

Chapter Three: Ventromedial prefrontal cortex encodes a latent estimate of cumulative reward

This chapter represents an extensive redevelopment of my M.Sc. thesis. While the underlying neural imaging data are identical, the analyses have been re-worked almost completely and the behavioural data has been supplemented with more sophisticated methods. Both the background and discussion have improved from the M.Sc. thesis and the text has been re-written in its entirety. Thus, although much of the underlying data is shared, I have invested a considerable amount of work over the course of this DPhil to improve the project and the following chapters depend in part on this work. Thus, in summary, it appears justified to include this chapter for completeness.

Contribution to thesis

Any scientific work is inevitably the product of the contributions of many people. The following lists how others contributed to the work laid out in this thesis in Chapters 3-6.

My supervisor, Prof Christopher Summerfield, provided support for the work in all data chapters, from conceptualization, design, data analysis, to the writing of these chapters.

Chapter 3 is the result of a collaboration of myself and Jan Balaguer, who is joint first author on the published paper. We conceptualized and designed the experiment and collected the data together. He programmed the experiment, and provided software for data analysis, prepared the figures based on the data I provided, and provided comments on a draft paper, whereas I analysed the data and wrote the paper together with my supervisor. Prof María Ruz assisted with data collection in Granada. The results of this study were published in the journal *Neuron* and an early version of the results formed my MSc thesis.

In Chapter 4, my two supervisors, Prof Summerfield and Prof Jill O'Reilly, and I conceptualized and designed the experiment. I programmed the experiment, collected and analysed the data, and wrote the paper together with Prof Summerfield. Jan Balaguer provided software for data analysis and helped with data collection together with Santiago Herce-Castanón and Prof María Ruz. The results of this study were published in the journal *Neuron*.

In Chapter 5, Prof Summerfield, Dr Andreas Jarvstad, and I conceptualized and designed the experiment, while I programmed the experiment, collected and analysed the data. Hannah Tickle and Prof Ruz helped with data collection and Jan Balaguer provided software for data analysis.

Chapter 6 was written jointly with Prof Summerfield as an invited submission for the journal *Trends in Cognitive Sciences* as an Opinion piece.

Acknowledgements

Completing a PhD takes up a large chunk of time, but I have been very fortunate during this part of my life that it was a great experience and I would do most of it again. The reason that the science went relatively smoothly is to be found in a great lab and supervisor.

First and foremost, I would like to thank Chris for being a great mentor throughout the last five years, for teaching me to be bolder and more ambitious, for ever great new analysis ideas, for fun experimental tasks that didn't involve Gabor patches, for continuing my stipend for a fourth year, for funding many great conference trips with the lab to Paris and elsewhere, for two fun lab retreats and many great summer and Christmas parties (with varying quality of Secret Santa gifts, but always good quizzes). Finally, I want to thank you for saying my writing was great and, despite this assertion, for continually changing the bulk of it to make it better.

Hannah Tickle for being the social heart of the lab, Fabrice Luyckx for many pizza lunches at the Rickety Press and great friendship, Andreas Jarvstad for many great discussions about economic decision-making, Timo Flesch for being a great desk buddy working on a similar schedule as me. Jan Balaguer for taking me under your wings when I first arrived and teaching me JavaScript, fMRI analysis, coding tricks, for being the designated Spanish-speaker for data collection, and for creating outstanding figures which have found its way into one of the chapters of this thesis, and our shared-first author paper. And everyone else in the lab for making it such a fun and inspiring place to work in.

The 1970s for using Asbestos as a building material so the department had to relocate mid-PhD to a much nicer building in a better location.

María Ruz for giving us the opportunity to run our experiments in Granada, which was always a welcome excuse to spend some time in Spain, despite having to spend a good 13 hours each day in a windowless basement.

Nathaniel Daw and his lab for making a short visit to the Princeton Neuroscience Institute a great experience I would not want to miss. And New Jersey Transit for running (mostly) on schedule.

I also wish to thank the Economic and Social Research Council for their generous funding and for providing financial support to spend some time at Princeton. And Wolfson College for having been a great academic home, egalitarian and unburdened by the many awkward traditions of Oxford colleges.

Lastly, and most importantly, James for always being there, for being the best friend and husband I could ask for. Without you I'm sure I would have suffered a nervous breakdown or two every time I had to finish a project.

Table of Contents

General Introduction

What is the life best lived? What is the normative solution to life? This question pervades philosophy since antiquity (Aristotle and Crisp, 2000) and occupies a sizeable portion of shelving space at high-street book vendors. Similarly, the internet is host to many lively discussions about which historical figure or culture lived life in the best way possible. These popular attempts at general answers can be divided broadly into two categories: Those that emphasize finding a good *goal* in life, and those that emphasize identifying likely *paths* towards success once a goal has been set. However, despite many centuries of investigation, it seems no satisfying solution has been found to this problem. Thus, to disappoint the reader early on, this thesis will hold no answers to what may constitute a good decision or indeed the good *life*.

What, then, can we investigate if no normative solution to life exists? What statements of sufficient generality can we make about behaviour? One promising avenue is to study the *mechanism* by which animals enact behaviour in the world (Krakauer et al., 2017; Tinbergen, 1963). After all, much of the biological machinery, from genes to sensory systems and brain networks, is shared across individuals within a species and to a large extent along the phylogenetic tree (Krubitzer, 1995; Neubert et al., 2015; Sallet et al., 2013; Wallis, 2011). Thus, at least the fundamentals of decision-making may be studied using paradigms with high generality, such as preferential choice (Lebreton et al., 2009), foraging (Hayden et al., 2011; Kolling et al., 2012), navigation (Balaguer et al., 2016; Doeller et al., 2010),

or associative and reinforcement learning (Schultz et al., 1993, 1997; Sharpe et al., 2019a), to name only those relevant to this thesis.

Thus, the focus of this thesis will be on the *how* and *why* of decision-making: First, the animal selects a goal, a *why*, of behaviour, before pursuing actions (the *how*) that further its progress to this goal. I will identify these two key mechanisms with the terms goal selection (*why*) and goal pursuit (*how*), following (O'Reilly et al., 2014). Hence, this thesis is chiefly concerned with decisions for which we may reasonably assume that they are made in a goal-directed fashion, as opposed to those that are simply made out of habit or which form part of innate reflexes. Within the division usually drawn between model-based (goal-directed) and model-free (habitual) behaviour (Daw et al., 2005; Dayan, 2009; Dolan and Dayan, 2013), this thesis will be biased towards the model-based side. The experiments reported will thus rely on behaviour that requires some form of planning and on-the-fly computation of values, as opposed to the execution of stereotyped behaviours based on *cached* values, i.e. based on remembered averages. Much of the thesis will draw on one key motivating question: how do we choose to prioritize one goal over another? I will argue that we know relatively little about the processes underlying this goal selection, both in terms of observed behaviour as well as in its neural underpinnings (Mansouri et al., 2017; O'Reilly et al., 2014).

I will begin by splitting this general motivating question into three more specific sub-questions. These questions broadly map onto the studies reported in Chapters Three-Five, and the summarizing Chapter Six, and shall thus set the scene for a broader review of the literature relevant to my topic:

I) What factors contribute to the cost function of behaviour?

Cost function is a central concept in any theory of decision-making. In a broad sense, it describes what counts as success and failure and assigns value to objects and situations (called states) that may be encountered in the world. Other terms for this function include objective function, reward function, or loss function. The decision-maker's goal is then to maximize the long-run output (or minimize the loss) of this function, e.g. choosing options that lead to greatest success over an arbitrary time horizon. A popular variant of this is the expected utility model (Von Neumann and Morgenstern, 1944), which assigns *utility* to options. This utility may differ in magnitude and ordering from any objective standard - e.g. it permits that one assigns higher utility to £20 than to £50 in some situations as long as one is consistent in this choice. One key point of contention within the larger framework of utility models has been the question in how far utility depends on a point of reference, or a "context". For example, utility is known to vary as a function of external factors, such as the framing of choice (De Martino et al., 2006; Tversky and Kahneman, 1981), the available choice set including irrelevant alternatives (Tsetsos et al., 2016, 2010), or training curricula (Bavard et al., 2018; Klein et al., 2017). Similarly, utility may also depend on more subjective factors, such as the wealth and expectation incurred from previous choices (Friedman and Savage, 1948; Kőszegi and Rabin, 2006; Markowitz, 1952), as well as factors relating to the underlying neural substrates such as the level of noise or uncertainty in inferring the best course of action (Li et al., 2018; Polanía et al., 2018; Woodford, 2012). The focus of this thesis will be mostly on these "subjective" factors.

I will discuss some of the serious problems raised by context-dependence for utility maximization models which often assume that preferences are invariant to (irrelevant) context and do not change over time (Bruckner, 2009; Mas-Colell et al., 1995). In Chapter Two I will review these problems from a general perspective, as well as in the context of an experiment in Chapter Three.

II) How do we monitor two concurrent goals?

Humans and other animals must satisfy several competing goals simultaneously, e.g. feed, hydrate, groom, maintain social ties, etc. Many primary goals exist only within a small range of permissible values, where the optimal "set-point" is regulated via a mechanism called "homeostasis" (Berridge, 2004; Cannon, 1929; Carpenter, 2004; Keramati and Gutkin, 2014). These primary mechanisms are well-preserved across evolution with many converging on the hypothalamus which in turn regulates behaviour often by way of specialized pathways, e.g. the leptin/ghrelin pathway for energy regulation (Nogueiras et al., 2008). Many higher-order goals associated with sophisticated, flexible behaviour rely on neocortical brain areas, such as the prefrontal cortex – for instance, foraging behaviour, career advancement, or social decision-making. How then does the organism know whether to eat or to continue writing a paper (or thesis), whether to clean the house or read a book? Naturally, it seems one should weigh the costs and benefits of each course of action and choose the one with greatest net benefit. However, how does an increasingly untidy house increase the value of cleaning and how does one compare this to an encroaching deadline? In other words, how does one establish the value of a goal – independently of any *specific* course of action – to engage in pursuit of that goal whilst ignoring others? I will argue in the tradition of drive

reduction theories (Hull, 1943) and homeostatic reinforcement learning (Keramati and Gutkin, 2014) that distance to a desired setpoint, or goal, is a widely available signal in the brain and show in an fMRI study in Chapter Four how this distance to goal may be employed to contrast competing goals.

III) How do we maintain choice flexibility across contexts?

Humans and animals adapt remarkably quickly to new contexts and have the ability to flexibly adjust behaviour to fit changing requirements (Mante et al., 2013). This poses a significant challenge to the brain: How should it store this information to allow flexible readout according to new goals whilst ignoring irrelevant features? If the brain only stores the average value it received in previous contexts it will continually have to re-learn these when circumstances change. Conversely, remembering all possible features quickly overburdens the memory system. Consider, for instance, a merchant needing to select goods to bring to market. Some days, the buyers tend to be younger and older on others. If the merchant remembers the average earnings for each good, they will perform decently on average – however, if they were to select those goods that are maximally valuable to younger buyers on market days with a lot of young people and vice versa for older people, they can maximize their profits on average.

Much research on the "common currency" (i.e. utility) of neuronal value encoding ignores this problem by focussing on final value on which choices are based. However, we know relatively little about the flow of information that inhibits (or enhances) features of remembered items when these become irrelevant (or relevant), such as when the merchant needs to select goods for a "young" market

day. Previous research suggests that the prefrontal cortex is critical to such computations (Mante et al., 2013; Pelletier and Fellows, 2018) and works in concert with the hippocampus (Shadlen and Shohamy, 2016), often via direct, monosynaptic connections (Öngür et al., 2003). However, it is unclear what value-related information is encoded in the prefrontal cortex. Does the PFC only represent the relevant information, or does it also represent the information pertaining to other, currently inactive goals? In other words, at what processing stage does the brain distinguish between relevant and irrelevant?

Investigating these questions is particularly important, because they may shed light on another important concept from economics: opportunity costs. If the merchant sells goods on one market day, they will not be available on the next where they may fetch a higher price. Thus, while information may be currently irrelevant to a goal at hand, it may nevertheless carry some weight in decisions as the representation of "missed opportunities". I will explore the balance between a maximally flexible, full encoding of all available information and a more restrictive but memory efficient encoding of the relevant information in an fMRI study in Chapter Five.

Equipped with these three questions, we may ask: How does one select the best action? Inherent in any such consideration is the concept of *value*. As we have defined earlier, animals are often assumed to behave in a way that maximizes the value they receive through interactions with the world. The final chapter of this thesis will argue that in many ways this is looking at the problem from the wrong end. Once a decision has been made, it is easy to judge its utility based on the outcome. However, how can one identify an appropriate objective, weight factors

in light of this objective and maintain a reasonable degree of flexibility to adjust if things go wrong?

The remaining introductory chapters will describe and critically appraise some common answers to this question. Chapter One is mostly concerned with the *objective* function and the concept of *value*. I will broadly review research and theory from economics, before outlining behavioural findings from psychology as well as their theoretical descriptions by computational models.

Where Chapter One was mainly concerned with the objective function and the way decisions over values can be made, Chapter Two will discuss how value may be learned and inferred from experience. I will introduce some common algorithms to accomplish this as well as some ways in which these can be augmented to build an accurate description of the environment. I will then review how the constructs introduced in Chapter One relate to biological substrates, such as neuronal spiking activity and its correlated, observable measures such as oxygenated blood flow. I will outline how these findings have broadened our understanding of the underlying *mechanism* employed by the brain to solve decision-making. Finally, I will outline the most immediately relevant findings for this thesis and how they employ the notion of a "context" in which decisions are made.

Chapter One: Value and utility

The purpose of this chapter is two-fold: First, I will review aspects of economic theory and thought that are relevant to the experimental chapters that follow. These include the notion of utility (or value), rationality and optimality, and the *target* of choices which is to maximize utility. The second aim is to critique the concepts just introduced in the light of their possible *implementation* in the nervous system and their ecological validity. Much of this criticism will rest on the distinction between descriptive and explanatory accounts of behaviour and different notions of the objective that biological agents maximize (or minimize).

A brief description of the term value in economics

Value is a strangely elusive concept in economics. As the field of economics changed over time, so did its most central concept. At first, economics was concerned with the *objective* value of an object derived from the amount of labour (its labour *cost*) needed to produce an object (Smith, 1786). However, in the second half of the 19th century, an object's value was understood to be reflect the pleasure of consuming (or possessing) said object: "A great theory of economy can only be attained by going back to the great springs of human action – the feelings of pleasure and pain" (Jevons, 1866). Central to this second view is that every market participant will derive pleasure from different actions and purchases – value is inherently *subjective*. Thus, we may say that any object will have different *subjective utility* to different people, which in the aggregate results in the demand of a certain good. This interpretation of value dominates much of modern, neoclassical economics as well as most studies in value-based decision-making.

If subjective utility forms the basis of decision-making, one is inevitably led to this question: How does one know whether a good has high or low utility to a person? Jevons (1866) answers this as follows: "Our choice of one course out of two or more proves that [...] this course promises the greatest balance of pleasure". Hence, by observing a person's preferences, one may infer the utility of the options amongst which the person chooses - this concept is known as *revealed preferences*. Of course, this requires the central assumption that one behaves *as if* choices maximized long-run subjective utility. If one only observes purchases made on behalf of someone else one will learn nothing about the buyer. In general, however, this seems to be a reasonable assumption. For instance, withdrawing one's hand from a hot stove implies preferring a healthy hand. In other words, choices reveal the preferences which drove the choices in the first place. The reader would be forgiven for thinking that this sounds eerily circular. Indeed, much criticism has been levelled at such a view of behaviour (Etzioni, 1986) and related concepts in economics such as the law of demand (Albert et al., 2012). Nevertheless, the utility of this concept is easily demonstrated by the fact that it enables economists to describe behaviour (which revealed preferences) in a mathematically sound way. Most importantly, it allows for predictions to be made between uncertain prospects without the need to observe all possible outcomes. Indeed, it is this mathematical foundation of decisions involving risk that has given economics a considerable boost since the mid-20th century, due in large part to Von Neumann & Morgenstern (1944) and Samuelson (1947). In turn this has lead economic thought to dominate subdisciplines of other fields (Hirshleifer, 1985).

Objective and subjective utility

What defines a subjectively best option? Much of economics is deliberately agnostic to this question. The amount of utility derived from consumption of any good varies across individuals and may depend on their age, status, wealth, culture, and many other factors. For instance, it was recognized early on that even money will be of different utility to different people: a gain of £100 is seen as more valuable to a person who possesses nothing than to a billionaire (Bernoulli, 1954). This concept is known as *diminishing marginal utility*. Despite this variability across individuals, preferences are mostly assumed to be fixed within any given individual.

In a seminal book on the topic of utility, Von Neumann & Morgenstern (1944) formalized Bernoulli's view into a mathematical account. Their work is mainly concerned with decisions under risk, so their terminology refers to lotteries. However, their theory is of sufficient generality to cover utility functions defined over risk-less decisions. They invoke four axioms which together define any complete function that maps a lottery onto its utility. As each of these axioms will appear in subsequent sections, I will introduce them here:

1. Completeness. For any pair of lotteries $\{A,B\} \in X$, the agent can decide that either $u(A) < u(B)$, $u(A) = u(B)$, or $u(A) > u(B)$. Thus, the agent has complete preferences over the set.
2. Transitivity. If $u(A) < u(B)$, and $u(B) < u(C)$, then $u(A) < u(C)$. In other words, the agent is consistent across three or more lotteries.
3. Continuity. If $u(A) < u(B) < u(C)$, then we can find a probability p, such that $p*u(A) + (1-p) * u(C) = u(B)$. In other words, one can mix the inferior and superior options to equal the utility of the mediocre option.

4. Independence of irrelevant alternatives. If u(A) < u(B), we can find a probability and another lottery C, such that p*u(A) + (1-p) * u(C) < p * u(B) + (1-p) * u(C). In other words, adding another outcome to both lotteries A and B does not affect their preference relation.

Equipped with such a consistent utility function, how should one make decisions? Quite clearly, as per definition, one should choose the option that has the highest utility value. However, this is often complicated by the fact that outcomes in the real world (especially in markets of scarce goods) are uncertain, i.e. one will not know exactly what one will receive (or be able to purchase). How can one combine this uncertainty into one's decisions? Following Bernoulli, von Neumann & Morgenstern show that first multiplying each potential outcome's utility by its corresponding probability and then summing all up will result in rational behaviour. Furthermore, as we shall see later, it can make (sometimes contentious) predictions about risk-seeking and risk-averse behaviour simply based on the shape of the utility curve. The formula looks as follows:

$$EU = \sum_{i=1}^{N} P_i \times U_i$$

Where P denotes the probability and U the utility of option i out of the set of N possible outcomes. Consider, for instance, being in a casino and having £10 worth of chips. One then plays a gamble which costs £5, which offers the chance to win £50. Thus, the two possible outcomes are a stack of chips worth £5 (lose, A), or £55 (win, B). We can plot these outcomes and their expected value (C) onto an existing, arbitrary utility curve as in Figure 1.

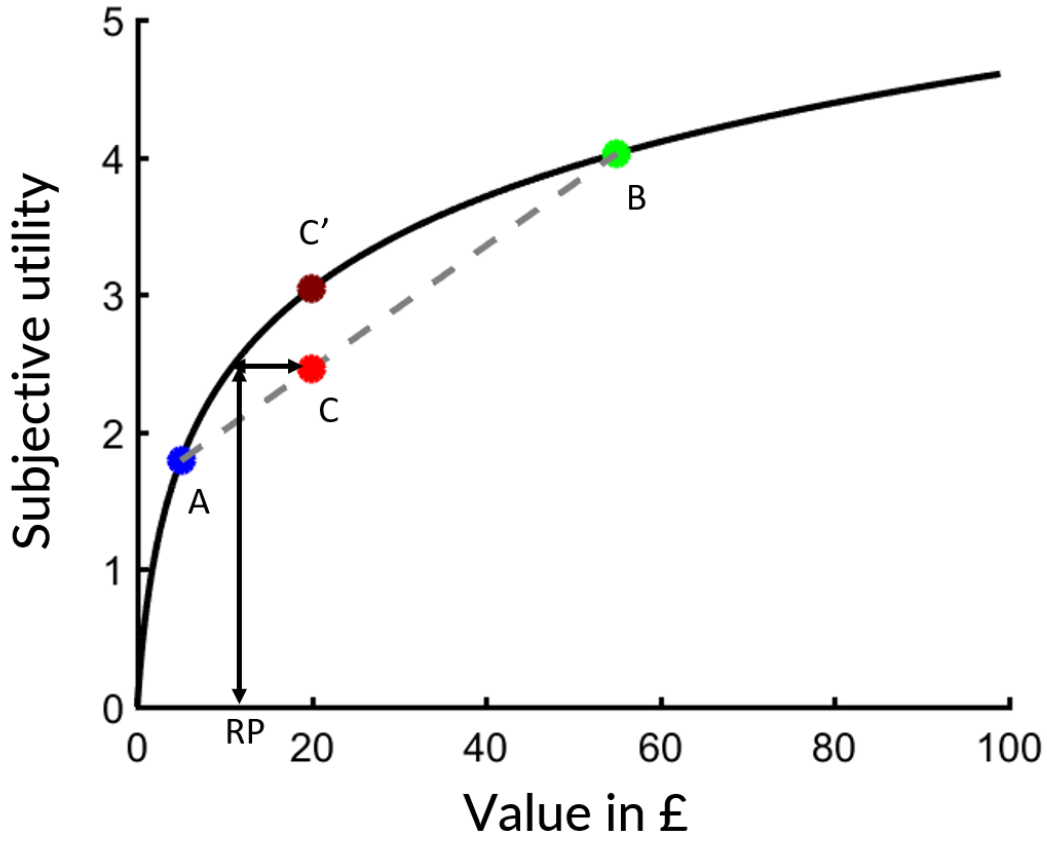


Figure 1. Diminishing marginal utility curve. This illustration depicts an individual's subjective utility curve over amounts in £. The two potential outcomes of the lottery are A and B. C represents the expected value of the lottery and C' maps this onto the individual's utility curve. As can be seen, C' has higher subjective utility than C and thus, the person would thus prefer a safe amount of equivalent £ value to this lottery. This would remain the case if the safe amount was lowered up to an amount labelled RP for risk premium. The difference between C and RP indicates how much money the individual would be willing to pay to avoid the risk of the lottery, e.g. if the lottery was a potentially hazardous incident such as a severe weather event.

The straight line connecting these outcomes represents the set of all possible probabilities relating to all possible gambles over amounts A and B. The true probability is marked with a red dot. We can then read off its expected value (x-axis), and the corresponding expected utility (the subjective equivalent) on the y-axis. Because C' on the person's utility curve lies above the utility of the gamble, it

indicates that this person would be risk averse: Here, it implies that if we offered this person a safe bet of anywhere between £20 (the gamble's EV) and around £10 (the EV's subjective utility) they would choose the safe bet rather than participate in the gamble just described. The difference between these two values is often called the risk premium and describes how much a person would be willing to pay in order to avoid an additional risk, or how much better a risky prospect has to be for a person to consider it over a safe prospect. For example, this is how much they might pay an insurer to shield them against the risk of losing their chip by playing this bet (or, losing their house in a hurricane).

One clear advantage of this approach is that it makes many different risk-bearing activities mathematically tractable (and intuitive in graphical representation). For example, if a person has a concave utility function as in the example in Fig. 1, then they are considered to be risk-averse, whereas a convex utility function results in risk-seeking preferences.

Non-linearity of subjective value and reference-dependence

In both cases, utility is non-linear over wealth – the amount of utility gained from an increase of one unit of wealth depends on the wealth itself. This is to say that gaining £1 if one has £2 is perceived as being different from gaining £1 if one has £1,000. Such non-linearity in the function mapping physical stimulation to subjective percept is commonplace in other domains, such as vision, mechano-sensation, or time estimation where potentially infinite signals from the environment need to be mapped onto a finite (capacity-limited) sensory range of the organism (Stevens, 1960). While non-linear utility over wealth has been well-recognized in economics (Bernoulli, 1954; Friedman and Savage, 1948; Markowitz,

1952), probabilities were long treated as if they were subjectively linear and thus identical to objective probabilities. Despite earlier efforts showing that this assumption may not be warranted (Edwards, 1962; Ramsay, 1926; Savage, 1971) and that reasoning involving risky lotteries systematically differed from the rational theory (Allais, 1953; Ellsberg, 1961) it was not until 1979 that researchers delivered a better descriptive model within the framework of expected utility (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). This framework was called Prospect Theory. It did not only inherit the non-linearity of utility from earlier models, but also treated probabilities as subjectively non-linear. For example, people tend to over-estimate low probabilities: people are more likely to invest in a lottery ticket than they should be given the low probability of success. The reverse is true for high probabilities which tend to be seen as lower than their objective counterparts. When we plot subjective probability against objective probability, the empirically derived function often follows an inverse S-shape as the ones depicted in Fig. 2 (Gonzalez and Wu, 1999; Prelec, 1998; Steiner and Stewart, 2016; Tversky and Kahneman, 1992).

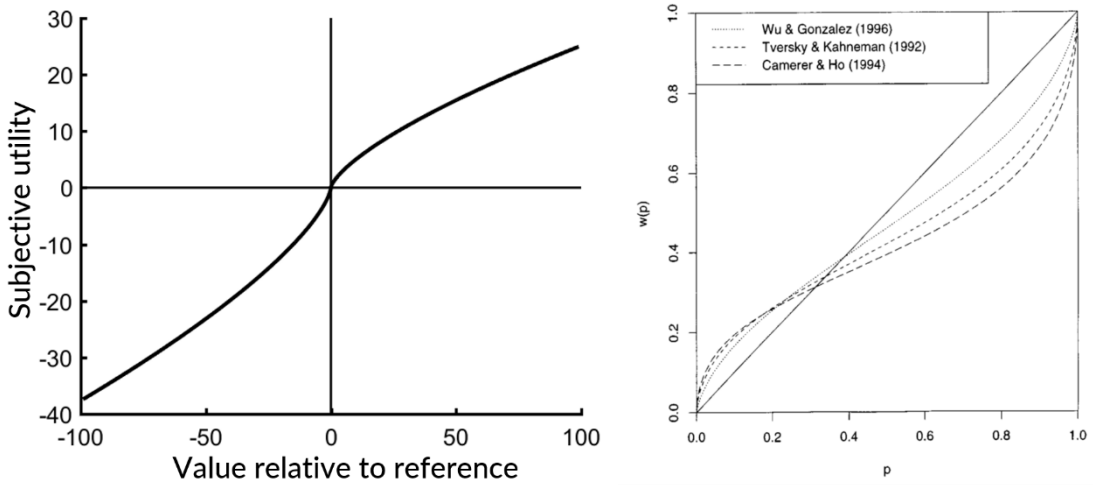


Figure 2. Prospect Theory subjective utility around a reference point and subjective probability. Commonly, the reference point is thought to be the status quo and deviations from it divide the space of possible outcomes into gains and losses. Typically, losses "loom larger", i.e. they are perceived as more punishing than a gain of equivalent size is perceived as rewarding. Many versions of the probability weighting functions exist, which all share the typical inverted S-shape, whereby low probabilities are over-weighted and high probabilities are under-weighted. Reproduced with permission from (Gonzalez and Wu, 1999).

One possible explanation as to why these distortions exist is that computation in neural systems is intrinsically noisy. In fact, it can be shown that under minimal assumptions probability distortions as in Fig. 2 lead to reward-maximization in a noisy system (Steiner and Stewart, 2016).

In addition to convincingly demonstrating non-linear subjective probabilities, Prospect Theory invoked the notion of reference-dependence in economic decision-making. Under this view, the agent does not integrate possible outcomes into its status quo, but rather views outcomes as *changes* to the status quo. For example, a wage worker may view the choice between two jobs with pay 50,000 and 60,000 not in terms of absolute salary, but as relative to their current income (e.g. 55,000). Despite being similar in salary, the lower-paid job will be seen as a loss compared to the status quo, whereas the higher-paid job as a moderate increase. In other words, the utility u(x) does not simply depend on value of x, but also on a reference value s, such that u(x) = v(x-s). In many scenarios this reference value will be identical to the status quo (as in the worker example), but may also refer to expectations (Kőszegi and Rabin, 2007, 2006), targets set in advance (Camerer et al., 1997; Lopes and Oden, 1999; Lopes, 1994), or background values

such as other possible outcomes within the same context (Bavard et al., 2018; Fouragnan et al., 2019; Rigoli et al., 2016, 2015).

Introducing non-linearities and reference-points can lead to perhaps surprising irrationality when comparing behaviour of such models with normative models (e.g. the expected utility model introduced by Von Neumann & Morgenstern, 1944). For example, preferences may reverse depending on either the framing (De Martino et al., 2006; Tversky and Kahneman, 1981), other (seemingly) irrelevant options (Li et al., 2018; Louie and Glimcher, 2012; Polanía et al., 2018; Rigoli et al., 2016; Webb et al., 2014), or the options' concurrent pairs in a stream of signals (Tsetsos et al., 2016). From the perspective of a utility-maximizing system, these reversals of preference may seem puzzling, because preferences need to be *consistent* to ensure good predictions of the utility function. However, one key feature of the findings noted above is that preferences may only be consistent within certain contexts or given the same reference and expectation. In other words, they may be *piecewise consistent*. For example, the precision needed to make optimal choices in domains important for survival may come at the cost of irrational choices when no particular goal is engaged and inconsistency is relatively costless. Indeed, animals are often assumed (and shown) to behave close to optimally in relevant domains such as food foraging or mate selection. This is the central idea of the influential marginal value theorem (MVT) which describes how long an animal should spend foraging for food in a given patch: exactly as long as the current patch's reward rate is higher than the average across patches adjusting for travel costs (Charnov, 1976; Hall-McMaster and Luyckx, 2019; Wajnberg et al., 2000).

What may be the root of this apparent inconsistency? First, it may arise simply from the fact that animals do not always maximize utility over arbitrary time horizons and that other objective functions (that which the animal optimizes instead) can better explain observed behaviour. This will be the focus of the remaining paragraphs in this chapter. Second, independently of the assumed objective function, one may gain insight from the study of learning and the limitations of biological organisms. Animals are necessarily limited in their capacity to process and transmit information and this alone implies that animals will never achieve the normative maximum and may thus sacrifice consistency *at all times* to achieve an overall closer fit to this maximum. Furthermore, one may gain insight into *why* animals may fall short of the normative models by observing how behaviour is learnt (and whether or not it has a social component). This largely biological approach to understanding decision-making including neural substrates will be discussed in the following chapter.

An abundance of objective functions

In practice, utility is often defined over probabilistic lotteries that yield a possible range of monetary gains or losses. Thus, it may seem surprising that similar amounts are treated differently according to seemingly arbitrary context variables such as experimenter-imposed descriptions or sampling schedules. What one can buy with £10 does not change with one's expectations so why does it elicit such different perceptions? This apparent discrepancy stems partly from the fact that the single quantity that governs much of our lives should, in principle, be perceived in proportion to its real value. However, in naturalistic scenarios such clear and intuitive *objective* utility is not to be found. Instead, animals need to maximize such

diverse goals as mating success, survival of offspring, personal survival, and maintaining internal states close to desired set points (e.g. body temperature). It is thus tempting to assume that the key variable that is being optimized is evolutionary fitness. However, this is often difficult to compute and relate to individual behaviours (Doebeli et al., 2017; Houston et al., 2007; Kacelnik and El Mouden, 2013). Instead, animals should be expected to behave rationally (and optimally) in situations that their ancestors are likely to have evolved in (McDermott et al., 2008). Thus, it is critical to use models which are "ecologically valid" in the sense that they tap into behaviours likely needed by the animal throughout its lifetime. The following paragraphs will review extensions of the utility framework to incorporate such biologically relevant constraints.

Probability maximizing functions

One class of possible functions posits that risky choices made by animals may be better described by maximizing the probability of success rather than the overall amount to be expected (Caraco, 1981; Srivastava and Schrater, 2015). For example, a foraging animal may choose riskier options to avoid impending starvation, but may prefer safer choices when it is already well fed. This kind of behaviour is known as the budget rule (Houston et al., 2014; Mallpress et al., 2015; McNamara and Houston, 1985; Stephens, 2008, 1981). Because of their nature as functions maximizing (or minimizing) the probability of an event, these descriptions naturally give rise to the prediction that risky behaviours are directly dependent on the internal state of the animal (Stephens, 1981) as well as on the overall quality of the environment (Dener and Kacelnik, 2016). In other words, they are explicitly context-dependent. Such context-dependence does not rule out EUT accounts of

behaviour: Indeed, if choices are assumed to achieve a certain target (such as maintaining resources above a survival threshold), preferences can be shown to be equivalent to expected utility (Bordley and LiCalzi, 2000). Nevertheless, they offer a clear path to better measurement (and by implication, prediction) of new behaviours based on previous observations. If one can identify the internal state variables modulating risky behaviour, one's predictions will provide a more complete description of behaviour than one based purely on revealed preferences without reference to context.

Despite inconclusive evidence about the predictions of the budget rule in animals (Kacelnik and El Mouden, 2013), the dependence on internal state is shared by a wider (and more successful) class of models reviewed below.

Homeostatic or target-based functions

The budget rule describes behaviour with respect to an aversive state (e.g. starvation), arguing that behaviour changes as a function of the distance to this aversive state. More generally, however, agents wish to attain certain goals by setting aspirations, while avoiding negative outcomes. Indeed, this is the central idea of objective functions based on the concept of homeostasis: It describes the fundamental process by which biological systems maintain important variables, such as blood glucose, within permissible ranges. This is not to say that this range is necessarily *fixed* – it may drift over time and itself depend on other features of the environment, such as ambient temperature (Carpenter, 2004). Such a model accomplishes three things: First, it posits a *setpoint* which the organism seeks to attain or stay close to. This setpoint need not be encoded by any particular

substance or neural correlate, but may itself result from opposing forces (Berridge, 2004). Nevertheless, its workings are observable in the frequency with which the organism occupies a range close to this setpoint, e.g. in blood glucose measured across long timescales. Second, it describes the pressure, or *drive*, of the organism to redress a shortfall with respect to its desired setpoint (Cannon, 1929; Hull, 1943). Third, and most importantly, it posits that utility, often measured by the rate or vigour of approach or avoidance critically depends on the current drive (Keramati et al., 2017; Keramati and Gutkin, 2014), or its *incentive salience* (Zhang et al., 2009). In this sense, utility may be defined by how much a decision changes the critical variable towards its optimal set point.

Predictive models

A separate class of models describes actions as minimizing "surprise". The organism is thus assumed to build a generative model of either its incoming perceptions (Rao and Ballard, 1999) or of both perception and action, with the latter concept known as the *free energy principle* (Friston, 2010; Pezzulo et al., 2015). The organism thus continually computes the error – or surprise – that its actions generate with respect to its model. Such prediction error processes are shared with other models, particularly the reinforcement learning model discussed in the following chapter. However, in contrast to other predictive models, the free energy principle goes further than this: It casts *inference* as the central guiding principle of animal behaviour. It not only provides a process by which this happens – Bayesian probability updates – but also casts precision in this process as the ultimate goal. Perhaps one key question in this context is what defines the form of the generative model, and by implication the predictions to be made. For example,

without any additional exploratory drives, a predictive system of just a handful of variables would generate less surprise than one containing many more variables. Thus, the model may be driven to predict very little with high precision. As a consequence, the high flexibility of the free energy principle may not constrain the model (and experiments) enough to provide conclusive evidence in favour of it (Gershman, 2019). Nevertheless, casting value-based decision-making as an active inference process and thus abstracting away the specifics of any particular choice situation may elucidate the fundamental processes by which biological agents make decisions, assuming that the fundamental currency is probability.

Conclusion

Which definition of utility – does it maximize reward, achieve set points, reach targets, or minimize surprise? – should be used is ultimately a question of convenience and purpose (Etzioni, 1986; Haslett, 1990). The criticisms levelled against any definition – even the most fundamental one, i.e. that utility is a tautology – may not cut deep for any particular use. Indeed, where its primary purpose is to describe and summarize behaviour, a model based on ordered preferences such as the von Neumann & Morgenstern model may be of great use despite some inaccuracies. However, as neuroscientists, our main goal is to *explain* behaviour and as such demands definitions of utility that include both a notion of the appropriate objective function (especially when talking about neural circuits) as well as an explanation of *how* a decision is made. From this point of view it is irrelevant whether one can define any function that consistently describes

aggregate behaviour, because one's goal is to find the intermediate parts and explain how they work together. Indeed, it would be of no use to an engineer to say that a factory minimizes loss of material in its production (which, presumably, is the case for most factories) without knowing how it does so and what shapes the material arrives in. Neither should a vision scientist be content with the statement that the retina processes electromagnetic radiation. Thus, decision neuroscientists usually ask *how* decisions are made. This problem naturally decomposes into two aspects: decisions based on learned value and decisions based on (near) instantaneous inference and computation. This will be the focus of the following chapter.

Chapter Two: Learning and neural representation of value

Part One: Learned and inferred value

Learning values through local reinforcement

Finding an answer to the question of *how* one chooses between apples and oranges may seem a daunting endeavour. One good starting point is to consider *changes* in behaviour based on reinforcement or other teaching signals. For example, pets will learn to perform tricks in exchange for food. Similarly, subjective tastes evolve over the course of development, often from initial dislike: one is not born with a preference for salty liquorice; it is an acquired taste.

This learning approach was taken by early behaviourists who studied the formation (and extinction) of associations between stimuli or stimuli and actions (Thorndike, 1898). The simple and intuitive observation that actions leading to positive outcomes are more likely to be repeated was first summarized as the "law of effect" and later into formal mathematical descriptions of the computational mechanisms involved in this learning. Arguably, the most successful accounts in this area are Reinforcement Learning (RL) models (Sutton and Barto, 2018).

RL agents seek to maximize an incoming signal (the reward) through behaviour. What this signal entails and where it originates is deliberately left ambiguous, with the theory chiefly concerned with algorithms that guarantee that learning converges on optimal behaviour rather than the precise content of these algorithms (Keramati and Gutkin, 2014; Schaul et al., 2015; Sutton et al., 2011). A thorough critique of this particular aspect of RL models is given in Chapter Six.

For convenience, many toy problems in this domain are framed as navigation problems: For example, how can one learn to find the way from Union Square (start) to Times Square (goal) if one has never been to New York? We can solve this by defining four actions (left, right, north, south) that the agent can take at every junction and defining each unique junction (e.g. 42nd and 5th Avenue) as a *state*. The basic process then evolves as follows: The naïve agent finds itself in a discernible *state* s at time *t*, and needs to pick an action available at this state. Initially, it will take some actions at random for want of a better strategy. Once the agent has reached its goal (in this case through random exploration) it needs to somehow update the value of each state it has visited in order to make better decisions in following runs. RL provides such algorithms based on the key simple notion that the agent should estimate the average value it receives from a certain behaviour, a process called *caching*. Essentially, the agent will thus learn how much reward it *expects* to obtain from the state it is currently in. Once such updates start coming in, the agent can continually update its policies based on these cached values without ever having to remember what exactly it did in previous runs. Many specific algorithms based on this idea exist in the literature, each with their own specific advantages and shortcomings. A thorough discussion of these is beyond the scope of this thesis, but a few examples can be found here (Mnih et al., 2015; Momennejad et al., 2017; Russek et al., 2017; Schultz et al., 1997; Sutton and Barto, 2018; Tesauro, 1994).

One particular algorithm, however - the Temporal Difference (TD) learning algorithm - shows striking parallels to dopaminergic signalling in midbrain areas

(Schultz et al., 1997, 1993), the neurobiological specifics of which I will discuss in the following chapter. The update the agent performs looks as follows:

$$\delta_t = R_t + \gamma V(s)_{t+1} - V(s)_t$$

$$V(s)_t \leftarrow V(s)_t + \alpha\delta$$

Where V(s) represents the agent's *estimate* of expected reward for a state *s* and R represents the reward signal received. The parameters $\boldsymbol{\alpha}$ governs the *learning rate*, essentially the amount by which an old estimate is adjusted when new evidence is available, and the parameter $\boldsymbol{\gamma}$, called the *discount factor*, governs how myopic (or: greedy) the agent behaves with respect to time: For example, a low discount factor leads to higher weight for more proximal events, whereas a discount factor of 1 takes into account all available timepoints in its learning.

Thus, the algorithm has three core ingredients: i) a *prediction* over rewards that varies with time, ii) two parameters that govern how much this estimate depends on the past ($\boldsymbol{\alpha}$) and on the expected future ($\boldsymbol{\gamma}$), and iii) a *prediction error* $\boldsymbol{\delta}$, which measures the (signed) inaccuracy of the prediction. Under a minimal set of assumptions, this algorithm can be shown to converge (over a large number of observations) to the optimal solution in many simple learning problems such as choosing the action yielding the shortest path to a goal (Sutton and Barto, 2018).

From the perspective of the behaving organism or artificial agent, knowing which action to take (or which cue to pay attention to) thus reduces to looking up the prediction in a table of all possible actions and feeding the prediction error back into all relevant predictions whenever an observation is made. However, an agent running on such an algorithm is destined to follow the short-term best course of

action without regard to long-term consequences and lacks flexibility when the same action can be performed in different situations leading to different outcomes, i.e. if its representation of states is mis-specified. This has led to this type of learning being called habitual learning as courses of action become stereotyped and are relatively difficult to change (Daw et al., 2006; Dolan and Dayan, 2013; Miller et al., 2016).

Learning with a model

In particular, this kind of behaviour does not require any specific *goals* to be pursued, but rather describes overall action propensities irrespective of the (biological) agent's intentions. Thus, one important piece of animal behaviour – purposive or goal-directed behaviour – is missing. For example, animals may learn to follow new, not previously encountered paths to a food reward (Redish, 2016; Tolman, 1948), or learn higher-order rules that generalize knowledge across contexts (Harlow, 1949; Wang et al., 2018). Evidently, these animals have learned something that goes beyond a simple stimulus-response or response-outcome mapping. For example, the agent learning to navigate New York City, may re-use its knowledge of the grid street layout when navigating in other American cities. The abstract content of this learning is often called a *model* or *cognitive map*, with the latter term gaining considerable traction over recent years (Behrens et al., 2018; Schuck et al., 2016; Stachenfeld et al., 2014). Such models enable the agent to make flexible choices at the expense of greater memory cost and longer computation time at a decision point.

Again, RL offers a wide range of algorithms to incorporate various models into its computations (Sutton and Barto, 2018). The key insight is that these models

endow the agent with the capacity to *plan* and *simulate*. For example, if the agent experiences that an obstacle blocking the shortest path to its goal has been removed, it can immediately update its policy to account for this change, unlike a model-free system, which would have to experience each intermediate state and gradually adapt to the newly-found best route.

How does the agent learn to represent its environment in an efficient way? One possible way to achieve this is for the agent to simply represent a second look-up table that stores *transitions* between states in addition to the value table. In effect, each time the agent moves from one state to the next, it can increment its expected occupancy of this state by one, thus learning that moving north from 37th street leads to 38th street (Daw et al., 2006; Doll et al., 2012; Sutton and Barto, 2018). The agent may do so in a way that is temporally abstract, for example instead of storing all possible transitions, it may simply store whether it expects a certain course of action to lead to a given future state over an arbitrary time horizon. This is the idea of the Successor Representation, which considerably reduces memory and computation cost at the expense of some flexibility compared to a full model (Gershman, 2018; Momennejad et al., 2017; Momennejad and Howard, 2018; Russek et al., 2017)

The propensity with which humans (and, recently, rodents) employ a model-based versus model-free action strategy has been studied using simple two-step tasks, where the agent learns to predict action-dependent, but unrewarded transitions between states followed by another, fixed transitions which leads to reward of varying magnitude (Daw et al., 2011; Miller et al., 2017).

Part Two: The neural representation of value

Biological systems need to efficiently represent incoming sensory information to choose adequate actions given their goals. This chapter will cover the current state of knowledge of how (and where) the brain represents value in a way that is conducive to effective behaviour. The first part will review neural correlates of reward, reward predictions and risk, whereas the second part will describe how these correlates are modulated by external context (e.g. framing, distribution of outcomes) and internal context (e.g. mutual inhibition, signal imprecision, metabolic or cognitive state of the agent).

General correlates of reward-related information

Cached values: predictions and prediction errors

Reward (or cost) signals are a ubiquitous feature of a distributed network of cortical areas (Bartra et al., 2013; Steinmetz et al., 2018; Stott and Redish, 2015), as well as of many midbrain areas containing the source neurons of many neuromodulatory pathways, such as dopamine (Schultz et al., 1997), serotonin (Liu et al., 2014), and noradrenaline (Bouret and Richmond, 2015). Figure 3 depicts some of the most commonly identified value-coding areas and the dopamine pathway involved (at least in part) in reward processing.

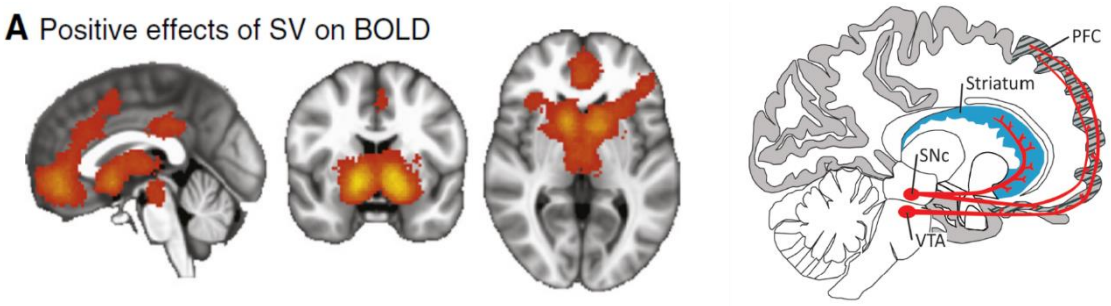


Figure 3. Correlates of subjective value and dopaminergic pathways. Positive subjective value outcomes (e.g. receipt of 5£) primarily engage the medial prefrontal cortex, the ventral striatum, and the orbitofrontal cortex. The two main dopaminergic areas, the substantia nigra (SN) and the ventral tegmental area (VTA) project primarily to the striatum and the prefrontal cortex. Figures reproduced with permission from (Bartra et al., 2013; Puig et al., 2014) – Puig et al., 2014 reproduced under CC-BY.

The system most readily associated with the notion of value is the dopaminergic system, though encoding of value is not its only function. Dopaminergic neurons are most abundant in the ventral tegmental area (VTA), which primarily projects to the ventral striatum and the prefrontal cortex (Morales and Margolis, 2017), and the substantia nigra (SN), which primarily projects to the striatum (Puig et al., 2014).

Work in primates in the 1990s convincingly demonstrated that dopamine neurons in the VTA encode the error relating to reward predictions (Schultz et al., 1997, 1993). In this classic paradigm, the monkey receives a cue (an auditory or visual stimulus), followed by a delay, and the delivery (or omission) of a juice reward. The recorded VTA neurons first reflected the amount (i.e. the reward value) of juice delivered. However, as the animal learned that the cue signalled delivery of a certain amount of juice, dopamine firing gradually shifted towards the time of the cue, with no discernible firing at the time of the (now expected) delivery of reward.

Instead, if the reward is now omitted, dopamine neurons will *reduce* firing at the time when the reward was expected, Fig. 4.

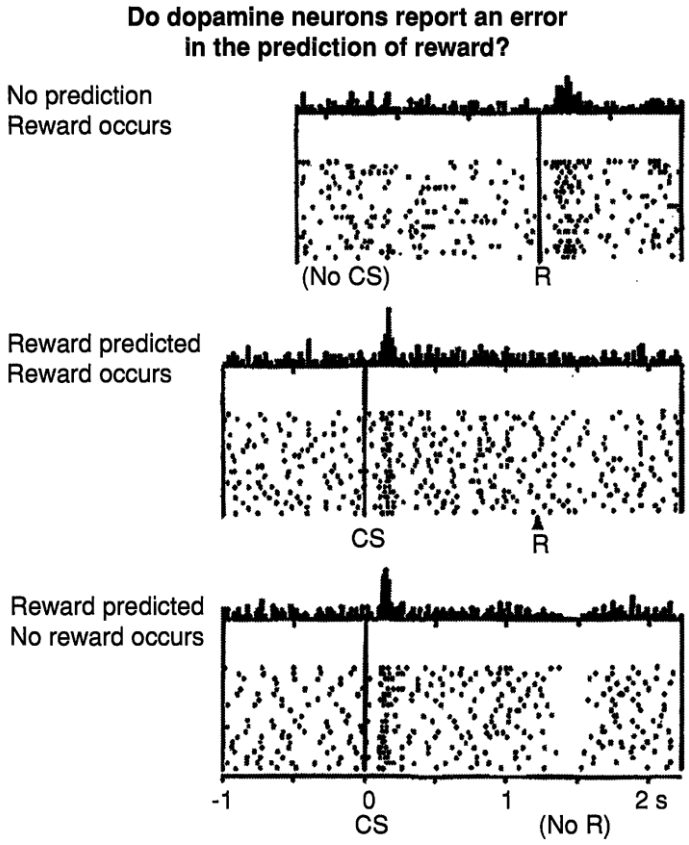


Figure 4. Dopaminergic reward prediction errors in the VTA. This figure depicts the firing rates of example dopaminergic neurons in the VTA to unexpected stimuli (top panel), expected reward (middle panel), and unexpected reward omissions (bottom panel). The neurons first fire at the time when an unexpected reward occurs, slowly adjusting their firing to the time of the reward predicting cue, and reduce their firing at the time when a reward was expected, but omitted. Reproduced with permission from (Schultz et al., 1997).

Later work showed that this reward prediction error signature is remarkably homogeneous across dopamine neurons (Eshel et al., 2016) and that it encodes the actual expected *size* of reward in an adaptive manner (Tobler et al., 2005). However, the precise role of dopamine in the prediction of reward is more complicated. For instance, RPEs can be modulated by the animals internal state (Cone et al., 2016), depends on whether actions are executed accurately or

withheld (Guitart-Masip et al., 2014; Syed et al., 2016), and may also reflect other (potentially reward predictive) features of the environment (Engelhard et al., 2018). Although transient dopamine firing has been shown to be sufficient for conditioning (Tsai et al., 2009), it remains a topic of debate whether the dopamine signal is best understood as signalling "value" or a more general associative learning signal, which does not itself confer value to cues (Sharpe et al., 2019a).

The pattern of RPEs has also been reproduced in many human fMRI studies, where these signatures commonly appear in the ventral striatum, specifically the nucleus accumbens (Bartra et al., 2013; Rutledge et al., 2010), which is one of the downstream targets of VTA dopaminergic neurons (Morales and Margolis, 2017).

The abundance of RPE signals begs the question of where (and how) the prediction itself is encoded in the neural system. It is here where the neocortex plays a central role. The predictions themselves can come in many shapes – for instance, one may predict the consequences of a particular action, the incentive value associated with a passively viewed cue (such as a picture of a slice of cake), or predict just the *identity* of an outcome without its valence.

The value of particular actions is perhaps also most closely related to the striatum (particularly its dorsal part) as well as the anterior cingulate cortex (Rushworth and Behrens, 2008). Expected values of viewed objects – e.g. measured via choice preferences in an independent task or through willingness-to-pay for the object – elicit strong responses in the ventromedial prefrontal cortex (Bartra et al., 2013; Knutson et al., 2005; Lebreton et al., 2009; Plassmann et al., 2007). Furthermore, one study in macaque monkeys demonstrated that neurons in

the (central) orbitofrontal cortex encode the stimulus identity – in this case the flavour of juice – as well as the offered *subjective* value (Padoa-Schioppa and Assad, 2006). Strikingly, however, many OFC neurons employed yet a different frame of reference: the value of the *chosen* option in a way that was unrelated to the particular action that needed to be executed (Padoa-Schioppa and Assad, 2006; Rushworth and Behrens, 2008). The presence of both value identity and choice value make the prediction that neuron firing should change if the encoded stimulus has been devalued, e.g. by pairing it with an aversive stimulus, and that OFC should be critical for accurate choices between "goods" (Padoa-Schioppa and Conen, 2017). There is now strong evidence that the OFC is indeed sensitive (and critical) for devaluation (Bradfield et al., 2015; Gottfried et al., 2003; Howard and Kahnt, 2017; Sadacca et al., 2018; Sharpe et al., 2015), but some recent findings have called into question the OFC's key role in representing "value" (Gardner et al., 2018), instead, they argue that the OFC represents the associative structure of the environment, e.g (Schoenbaum et al., 2011). I will discuss this in more detail in a later paragraph.

This contrasts somewhat with evidence from humans, where chosen value is most strongly associate with the ventromedial part of the PFC (Bartra et al., 2013; Levy and Glimcher, 2012, 2011; Plassmann et al., 2007), although it is at times difficult to establish the correct homology between monkey and human brains, and especially between rodent and human brains (Neubert et al., 2015; Wallis, 2011). The vmPFC encodes *chosen* value as a seemingly domain general signal (Lebreton et al., 2009; Levy and Glimcher, 2012), reflecting both gains and losses (Bartra et al., 2013; Plassmann et al., 2007). These chosen value signals are often thought to

reflect a decision signal which evolves from mutual inhibition of competing pools of neurons (Hunt et al., 2012; Hunt and Hayden, 2017; Wong and Wang, 2006).

How would an agent choose between goods of equal value if all it has is the value signal itself? Clearly, such choice requires representation of further information, allowing the agent to break the ensuing deadlock between choice options (aside from letting it become resolved through noise). Thus, over and above the expected reward, an efficient agent should also encode the *uncertainty* or *risk* around this expectation. For example, as discussed above, animals should adjust their risk preferences depending on their internal state (Caraco, 1981; Stephens, 1981) and should learn differently depending on whether incoming information (e.g. the reward received) provides a reliable or a noisy estimate of the true underlying distribution (Behrens et al., 2007; O'Reilly et al., 2013). A key area representing such uncertainty is the anterior cingulate cortex (ACC), particularly its dorsal part (Kolling et al., 2016; Meder et al., 2017; Rushworth and Behrens, 2008). The ACC had earlier been described as one of the sites of cognitive control, which resolve conflict among competing actions, as, for instance, in the Stroop task (Botvinick et al., 2001; Yeung et al., 2004). More recent evidence, however, suggests an integral role of the ACC in switch-stay decisions, such as those in foraging, where one decision is made against the background of other available options. Here, the dACC signals decisions away from the current context (or "foraging patch") (Hayden et al., 2011; Hunt and Hayden, 2017; Kolling et al., 2012). Nevertheless, debates persist about the best explanation for such patterns of activity in the dACC (Ebitz and Hayden, 2016; Shenhav et al., 2016, 2014, 2013). However, the balance of evidence seems lopsided in favour of complex switch-stay

decisions which appear more readily quantifiable than cognitive control itself (Kolling et al., 2016). Thus, in addition to the ACC's long intrinsic timescale – a measure of the time horizon over which an area integrates – at the apex of a cortical hierarchy (Bernacchia et al., 2011; Murray et al., 2014) and its representation of temporal distance (Balaguer et al., 2016; Kolling et al., 2014; Shidara and Richmond, 2002), dACC seems ideally placed to represent decision uncertainty.

A closely related concept to uncertainty is *confidence*, which essentially describes the *subjective* inverse of uncertainty. It has been shown, for example, that activity across the ventromedial prefrontal cortex not only scales with expected reward, but also with a how confident participants were in their value estimates (De Martino et al., 2013; Lebreton et al., 2015).

Context-dependence in value encoding

Choices do not usually happen in a vacuum. Instead, they are embedded into contexts of various kinds. For example, researchers have studied how both neural encoding and biases in responses depend on other concurrent information entering the senses. For instance, a common feature of neural circuits is that activity in one pool of neurons inhibits the activity of another pool of neurons (Carandini and Heeger, 2012; Heeger, 2017). This mutual inhibition can serve a number of functions, such as enhancing contrast between two stimuli, de-correlating sensory inputs (Heeger, 1992), or normalizing firing rates to accommodate constant precision across wide ranges of sensory inputs (Rustichini et al., 2017). In its most

general sense, neuronal firing rates are thus always relative to other neurons in the circuit and, by implication, other sensory inputs. While most of these observations stem from research of the visual system, models utilizing such a *relative* view of neuronal coding have recently been successful in accommodating both previously puzzling behavioural effects in economic decisions (Li et al., 2018; Louie and Glimcher, 2012; Polanía et al., 2018; Tsetsos et al., 2016; Webb et al., 2014), as well as neuronal firing patterns (Louie et al., 2011; Machens et al., 2005; Wong and Wang, 2006) and their macroscopic aggregates (Hunt et al., 2012). One key consequence of this seemingly ubiquitous mechanism is that economic behaviour may violate the axiom of irrelevant alternatives such that the previously preferred option in a binary comparison is now rejected after introduction of a third (but overall worse) alternative (Louie et al., 2013). However, under the additional assumption that neuronal coding is noisy, this mechanism may actually lead to overall optimal decisions (Tsetsos et al., 2016) when applied to local comparisons, such as when comparing two options attribute by attribute.

Efficient coding of value

We can extend this principle of relative firing in the temporal domain, where neuronal plasticity may adapt to various *expected* features of the stimulus distribution so as to represent the relevant information in the most efficient and effective way possible. The key problem here is that each neuron can only fire within a limited range of rates implying that a possibly infinite range in the physical world needs to be mapped onto a finite range in neural circuits. One way to achieve

this while maintaining discriminatory power is by employing mutual inhibition (as seen above). However, capacity limitations raise the more general question of how occurrences in the environment should be mapped onto their neuronal representations. One answer to this question – known as the efficient coding hypothesis – posits that neuronal firing rates should represent the cumulative distribution function (CDF) across possible stimuli as illustrated in Fig. 5.

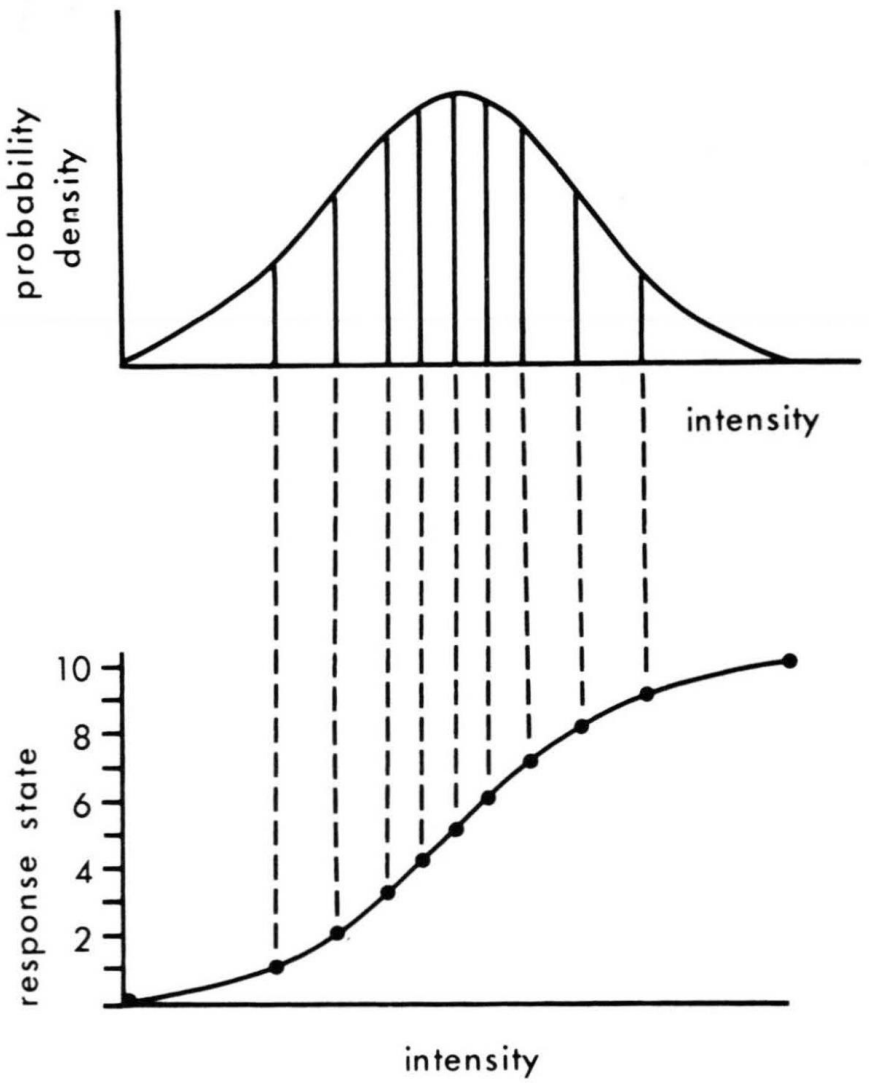


Figure 5. Illustration of stimulus probability density and its relationship to efficient neural encoding. Efficient coding posits that neurons should use their finite firing rates in a way

that removes redundancies and makes best use of the available range. In particular, neurons should be most sensitive to changes in stimulus features at points where many stimuli are likely to occur, i.e. around the peak of the top probability in the top panel. This results in a curve resembling the cumulative density function (CDF) of the stimulus distribution. Reproduced with permission under CC-BY-NC-ND from (Laughlin, 1981).

This ensures that neurons are maximally sensitive at the points in stimulus space that are most likely to be occupied, or at which many distinct stimuli occur close together (Barlow, 1961; Laughlin, 1981). This enhances the ability to discriminate stimuli where it matters at the cost of poor discrimination at very unlikely sensory events. The efficient coding hypothesis, together with its cousin "adaptive coding", has been instrumental in recent advances in explaining some counter-intuitive effects such as the finding that stimuli close to the category-boundary are often perceived as being further away from the boundary both in perceptual (Cheadle et al., 2014) as well as economic decisions (Li et al., 2018; Polanía et al., 2018). Furthermore, distortions of value representations such as those proposed by Prospect Theory have been described as reflecting the CDF of values in the real world – this theory is known as decision by sampling (Stewart et al., 2006; Vlaev et al., 2011).

However, one formidable obstacle to efficient representation lies in the noise inherent in neuronal coding. A neuron's response will not be exactly the same in any two instances, for example the same firing rate may occur at differently spaced interval, latencies may be larger or smaller, and firing rates may be enhanced or inhibited due to inputs from other neurons not participating in the representation. This means that from the perspective of a down-stream neuron, the same incoming signal will be ambiguous: the down-stream neuron will be

uncertain about the underlying cause of its input. The system will thus seek to reduce this uncertainty by systematically distorting inputs in order to allow the down-stream neuron to better represent incoming information as it relates to relevant behaviour. For example, in a binary discrimination task, stimuli that fall very close to the category boundary risk being confused with the other category simply due to noise in the system. Thus, it may be advantageous to distort information in a way that minimizes such risk. It has been shown that both magnitude (Luyckx et al., 2018; Spitzer et al., 2017; Wei and Stocker, 2017) and probability (Steiner and Stewart, 2016) distortion (as opposed to linear encoding) can be understood in this way. Such distortions are not optimal in the sense that they maximize overall success of the agent (for example reaching 100% accuracy in a task), but they make the agent behave in a way that is as good as possible given the constraints of the system (such as its inherent noise and limited capacity).

Aside from these forms of context arising largely from the actual neural *implementation*, the brain needs to represent value signals in a general, abstract form so as to allow it to draw inferences from previous experience in new situations. The last part of this introductory chapter will look at these in detail and distinguish "value through inference" and "active construction".

Value through inference, planning, and active construction

Any value representation needs to be flexible to accommodate changing circumstances that may render cached values useless. Thus, a good model of what generates value (inference) or of the internal states which contribute to valuation

(active construction) may significantly improve the efficacy of behaviour. For example, endowing the mind with computational primitives, such as rings, lines, or trees enables the agent to simply learn new projections onto these primitives and re-use the policy relevant for each primitive (Gershman et al., 2017; Tenenbaum et al., 2011). The core underlying principle, abstraction, may thus be expected to underpin neural activity in at least some brain areas. Indeed, abstract codes have been found to increase in prevalence in the prefrontal cortex along a dorsal-posterior to ventral-anterior axis (Donoso et al., 2014; Koechlin et al., 2003; Koechlin and Hyafil, 2007; Koechlin and Summerfield, 2007), with the most anterior part of the PFC, the frontopolar cortex (aPFC), representing the most abstract rules underlying a given task. In support of this, studies have shown that the aPFC reflects the value of alternative courses of actions (Boorman et al., 2011; Rushworth et al., 2011) and maintains information over long timescales (Mansouri et al., 2017).

Abstract representations have the advantage that they can be re-used in new contexts, thus significantly reducing the need to re-learn, but they may come at the cost of slower learning (Botvinick et al., 2019). Indeed, abstract representations may be unreliable themselves at some stages of learning, necessitating mechanisms that arbitrate between the reliability of the abstract model and the cached value signals (Daw et al., 2006; Doll et al., 2012; Koechlin, 2016).

One critical feature of a good model is that it allows the agent to plan its actions into the future through mental simulation. In the context of RL, this is often assumed to happen via “tree search”, where the agent starts from the current state

and simulates all possible paths from this state (perhaps until the goal is reached). This planning can quickly become computationally intractable and lead the agent to "prune" the tree such that it does not search through the entire set of possible futures (Keramati et al., 2016). More generally, however, the agent should build a representation of *task space* in the sense that it builds a model of the currently relevant environment which it can mentally traverse in the absence of actual behaviour. Two brain areas are key to such a representation: the hippocampus and the orbitofrontal cortex (together with the vmPFC). For example, the hippocampus forms abstract representations of the temporal relationships between stimuli (Schapiro et al., 2013; Stachenfeld et al., 2014), "replays" experienced trajectories (Mattar and Daw, 2018; Momennejad et al., 2018; Schuck and Niv, 2019), and "pre-plays" possible trajectories at decision points (Redish, 2016). However, the hippocampus is not the only region where such information can be found. Instead, the OFC may underpin abstract cognitive maps, especially when they relate to value information (Schuck et al., 2016). For instance, OFC underpins devaluation behaviour, essentially a one-step look-ahead which helps animals avoid aversive situations without having to experience them (Balleine and Dickinson, 1998; Gottfried et al., 2003; Sharpe et al., 2019b, 2019a).

However, what good is a map without a destination? Before setting out on the best path towards a given goal, the agent will first have to select one out of the many possible avenues. This is not simply a decision between what generates the highest rate of rewards as in many hierarchical and multi-objective RL scenarios (Barreto et al., 2016; Moffaert et al., 2014; Roijers et al., 2013; Vamplew et al., 2017), but rather a selection of which *reward function* is appropriate to use. For

example, if one wants to value water when thirsty, but devalue all other goods, the brain needs to dynamically adjust these associated values – it needs to actively construct the value based on current needs and select one out of the many possible rewards functions to guide behaviour. At present, we know relatively little about how the brain accomplishes this feat. O'Reilly et al. (2014) distinguish between this *goal selection* process and a *goal engaged* process, which largely covers the computational models discussed so far. The authors also attempted to map important aspects of this distinction (e.g. representation of action outcomes, possible goals, working memory maintenance) onto the medial and ventral parts of the PFC. This map summarizes much of the research presented in the introduction and is thus reproduced here:

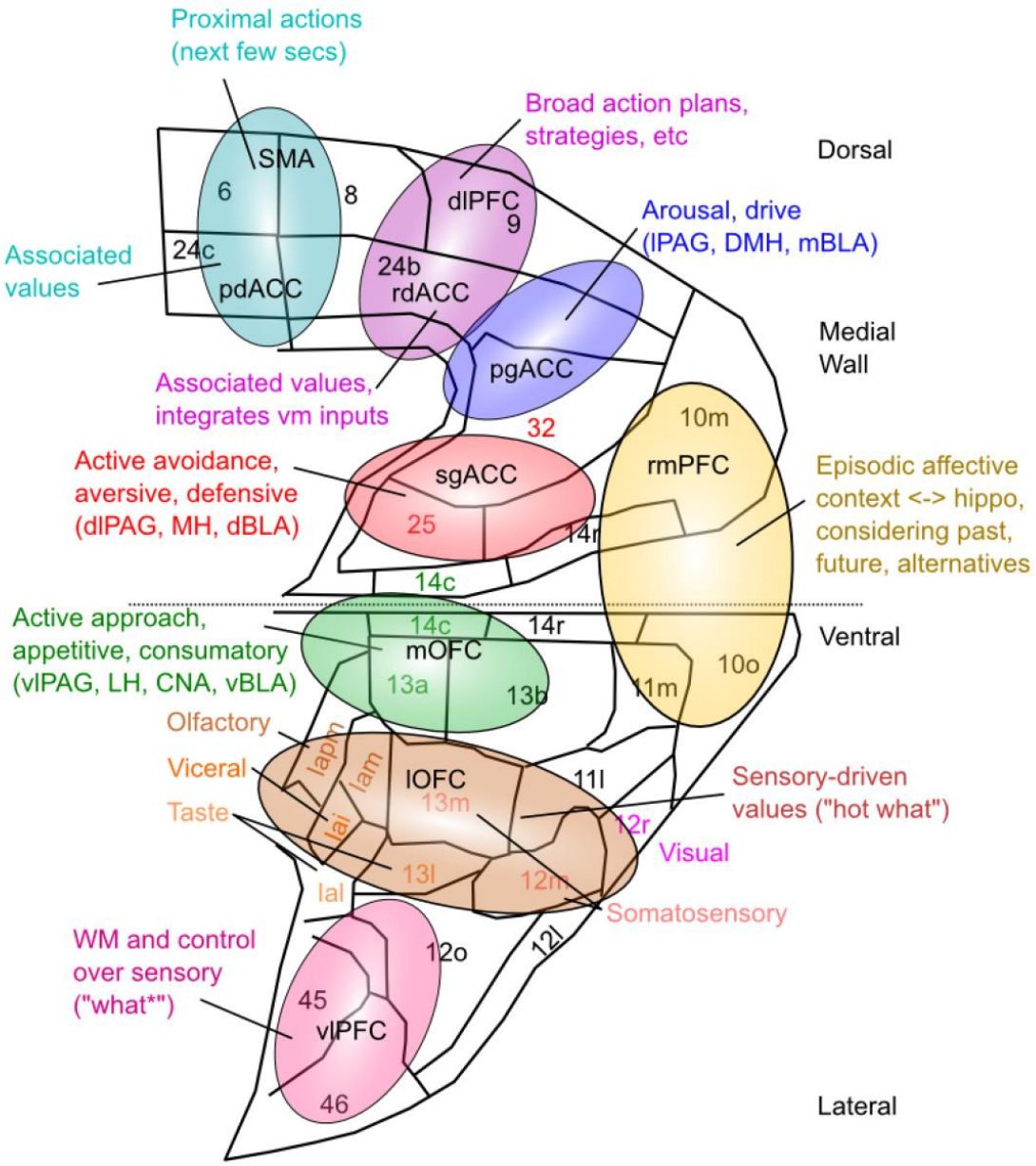


Figure 6. Summary of prefrontal areas and their involvement in goal-directed decision-making. Abbreviations: vlPFC: ventrolateral prefrontal cortex; lOFC: lateral orbitofrontal cortex; mOFC; medial OFC; rmPFC: rostromedial PFC; sgACC: subgenaul anterior cingulate cortex; pgACC: pregenual ACC; rdACC: rostral dorsal ACC; pdACC: posterior dACC; SMA: supplementary motor area; dlPFC: dorsolateral PFC. Reproduced with permission from (O'Reilly et al., 2014).

Similar distinctions between the frontal brain areas have also been drawn by others (Mansouri et al., 2017; Roy et al., 2012). Especially the rostromedial part of the PFC (rmPFC) is ideally placed to participate in goal selection as it receives input

from all major association areas in the cortex and is able to maintain information over long timescales (Mansouri et al., 2017). Why is this an important feature of goal selection? Goals are usually defined of incommensurable, distinct features, such as water content (thirst) and prestige (social standing). Thus, any region attempting to arbitrate between these disparate goals, needs to receive the relevant information and needs a mechanism by which it can enhance the relevant goal and inhibit the irrelevant one, a process which may be called on-line devaluation. Specifically, this means that value, as is commonly assumed, is not additive across dimensions, but is the product of *interactions* between value dimension. Furthermore, each value dimension will itself depend on time, for example by varying degrees of hunger over the course of a day.

At the point of writing – to the best of my knowledge – there were no studies attempting to look at the encoding of value in fMRI in such multi-dimensional tasks, where the rewarding value depended on the on-going accumulation across these dimensions.

In the following experimental chapters, Chapter Three will lay the groundwork by establishing a code for the on-going accumulation of rewards over time in the rmPFC. Chapter Four will generalize this to a task where rewards are accumulated along two dimensions and will discuss how the medial PFC, especially the perigenual ACC and the rmPFC, arbitrate between these two distinct dimensions. Chapter Five will then look at the aforementioned *interactions* between value dimensions and how this affects encoding in the medial and lateral PFC. Lastly, Chapter Six will discuss the relevance of these findings (and much other recent research) for research involving *value* or *reward*, and will argue that tasks

inspired by homeostatic reward functions are a promising future avenue for research in this area.

Chapter Three: Ventromedial prefrontal cortex encodes a latent estimate of cumulative reward

Abstract

Humans and other animals accumulate resources – or wealth – by making successive risky decisions. If and how risk attitudes vary with wealth remains an open question. Here, humans accumulated rewards by accepting or rejecting successive monetary gambles within arbitrarily-defined temporal contexts. Risk preferences changed substantially toward risk aversion as rewards accumulated within a context, and BOLD signals in the ventromedial prefrontal cortex (PFC) tracked the latent growth of cumulative economic outcomes. Risky behavior was captured by a computational model in which rewards prompt an adaptive update to the function that links utilities to choices. These findings can be understood if humans have evolved economic decision policies that fail to maximize overall expected value, but reduce variance in cumulative outcomes, thereby ensuring that resources remain above a critical survival threshold.

Introduction

Animals accumulate resources through repeated encounters with the world. In an uncertain environment, each encounter carries potential costs and benefits. For example, approaching a food item might expose a foraging animal to predation. The average resources accumulated over a fixed time period will be maximized by engaging with encounters (or "prospects") that have net positive expected value (i.e. those for which the likely benefits outweigh the likely costs) and avoiding all others. Thus, on average, a stockbroker will maximize weekly profit by investing in shares with positive return, even in a volatile stock market. However, humans and other animals are highly sensitive to the outcome variance, or *risk*, associated with the choice alternatives (Arrow, 1971; Bernoulli, 1954; O'Neill and Schultz, 2010; Platt and Huettel, 2008; Pratt, 1964; Schultz et al., 2011). When asked to judge everyday economic scenarios, humans often avoid risky prospects, including those with net positive expected value (Tversky and Kahneman, 1981). Classical theories argue that attitudes to risk can be accounted for by nonlinearities in the function that maps economic outcomes onto internal estimates of value, or *utility*, while more recent accounts have extended this framework by applying subjective decision weights to prospects instead of face-value probabilities (Kahneman and Tversky, 1979; Lopes and Oden, 1999).

However, humans and other animals are not always risk-averse. Risk preferences may reverse if the potential gains and losses involved are relatively small (Prelec and Loewenstein, 1991), when outcomes are framed as losses rather than gains (De Martino et al., 2006), or when primary, rather than secondary reinforcers are involved (Hayden and Platt, 2009), and may vary across the lifespan (Paulsen et al.,

2011) as well as across species (Hayden and Platt, 2009; Platt and Huettel, 2008). Critically, attitudes to risk can also change dynamically according to the abundance of resources in the environment, and the current level of wellbeing (Losecaat Vermeer et al., 2014; McDermott et al., 2008; Stephens, 2008; Xue et al., 2010; Yamada et al., 2013). For example, in resource-scarce environments, where safe hunting options may yield average levels of return that are insufficient for survival, animals typically adopt risky strategies (Abrahams and Dill, 1989; Kolling et al., 2014; Symmonds et al., 2010). One interpretation of these findings is that animals may have evolved decision policies that promote survival, rather than simply seeking to maximize long-run average value of outcomes. An animal wishing to maximize its chances of survival should take risks when resources are scarce (because there is little to lose) but become risk averse when ample resources have been accumulated, because safe choices maximize the confidence that cumulative resources will remain above a fixed threshold (such as zero) (Mallpress et al., 2015; McDermott et al., 2008; Wolf et al., 2007). However, the so-called budget rule describing this switch in risk sensitivity has yielded inconsistent evidence in the animal literature (Kacelnik and El Mouden, 2013). In humans, one recent study observed increased risk-taking in order to achieve an externally-imposed target, with BOLD activity in the medial prefrontal cortex scaling with the pressure to accept risky choices (Kolling et al., 2014). Another showed violations of dynamic consistency in sequential choices, with the target value encoded in medial PFC (Symmonds et al., 2010). However, both studies involved target values below which all resources were lost. It remains unknown how humans track reward accumulation in the absence of such researcher-induced targets.

In psychology and economics, the relationship between resource level, or "wealth", and risk attitudes remains controversial. For instance, Prospect Theory posits that changes from the status quo matter more than the status quo itself. However, if the reference point – the key improvement of Prospect Theory over Expected Utility Theory – is not defined as the status quo, but rather as an "aspiration" level, the coding of gains and losses can reverse (Kahneman and Tversky, 1979). For instance, receiving $5 will feel like a gain to someone who expected $1, but like a loss to someone who expected $10, even though their final asset positions are identical. Evidence from the field is likewise inconsistent: For example, stock market participation increases with household wealth (Hong et al., 2004), whereas troubled companies tend to take higher risks than their more affluent counterparts (Bowman, 1982; Kliger and Tsur, 2011). Moreover, in laboratory experiments, the "house money effect" showed that risk-seeking is higher after a prior gain (Thaler and Johnson, 1990), and analysis of sequential decisions suggests that rewards often beget further risky choices (Hayden and Platt, 2009), but can also induce risk-aversion (Gehring and Willoughby, 2002), whereas thirsty monkeys become less risk-averse with approaching satiety (Yamada et al., 2013). Multiple factors, including potentially uncontrolled economic factors in field studies (e.g. market liquidity) or differences in the evaluation of primary and secondary rewards in lab-based studies involving monkeys and humans, complicate the interpretation of these findings, and the descriptive relationship between risky behavior and reward accumulation thus remains unclear. Understanding the determinants of risk attitudes in humans and other species is an ongoing challenge in neuroscience, psychology, behavioral ecology and economics (Arrow, 1971; Caraco, 1981;

Hayden and Platt, 2009; Kacelnik and Bateson, 1997; Kahneman and Tversky, 1979; Kolling et al., 2014; Platt and Huettel, 2008; Rangel et al., 2008; Rudorf et al., 2012; Rushworth and Behrens, 2008; Schultz et al., 2011).

Here, we asked how human attitudes to risk vary with the context provided by their current resource levels – or "wealth" – and how risky behavior is modulated by these variables at the neural level. If risk attitudes are to vary with resource levels, animals must be able to keep track of the cumulative reward they have obtained in a given behavioral context. To date, however, studies of the biology of economic decisions have largely focused on the neural response evoked by the expectation of a momentary outcome, or its subsequent receipt (Bartra et al., 2013; Boorman et al., 2009; Daw et al., 2006; De Martino et al., 2009; Lim et al., 2011; Padoa-Schioppa and Assad, 2006; Pessiglione et al., 2006; Philiastides et al., 2010; Plassmann et al., 2007). For example, many studies ask humans to repeatedly choose between consumer products (e.g. food items) or abstract symbols that predict monetary reward, offering a randomly selected chosen outcome at the end of the experiment. Much less is known about how estimates of cumulative reward are encoded over time in human brain signals. Here, we designed a gambling task that, in conjunction with functional neuroimaging, allowed us to investigate how brain activity varies with cumulative reward over a discrete temporal context, and to measure how risk attitudes change as resources are accumulated. We found that regions of the ventromedial prefrontal cortex (vmPFC) that encode momentary outcomes also tracked the accumulated level of resources encoded over each temporal context.

Finally, we used computational simulations to understand these data, comparing the ability of both standard econometric accounts (such as Prospect Theory) and novel models to capture human behavior. Surprisingly, the model that provided the best quantitative account of resource-varying risk attitudes was inspired by work in perceptual decision-making, and posits that the weight (or gain) associated with decision information is adjusted dynamically according to the local context (Cheadle et al., 2014). In the current task, the model proposes that the function that maps subjective utility onto a risky choice adapts as resources accumulate. Our theoretical explanation for changing risk attitudes with wealth thus concurs with recent proposals that the coding of sensory signals and economic value can be understood by shared principles of normalization and adaptive gain control (Louie et al., 2013; Soltani et al., 2012).

Results

On each trial, human volunteers (n = 20) decided whether to accept or reject a risky gamble. The relative likelihood of winning or losing (reward probability), and the amount that could be won or lost (reward magnitude), were fully disclosed by a central display (Fig. 1a). 'Reject' choices led to a small but certain loss. On each of five blocks of 75 trials (scanner runs), decisions were made in successive temporal contexts (n = 5 within each scanner run) that were signaled by the color of information on the screen. Participants were told that their cumulative winnings in each context would be entered into a "pot" of the corresponding color (visible at the top of the screen), and that upon completion of the experiment, one of the pots in each scanner run would be chosen pseudo-randomly in a lottery procedure

and awarded as a financial bonus for participation. Upon completion of a context, the amount entered into the "pot" was displayed beneath the corresponding colored disc and remained visible for the remainder of the scanner run. However, cumulative rewards associated with an ongoing context were not signaled to the participants, but had to be estimated from the recent trial history. Contexts were of variable length (range 3-22 trials) but the number of remaining trials in the context was signaled to participants via a "countdown" signal on each trial (see Fig. 1a). The reward-maximizing policy for this task is simply to accept gambles with positive net value and reject the others, ignoring the temporal context. For more details of the task and lottery, see Methods.

Behavioral data. Participants elected to gamble on 54.9% (SD = 11.9%) of trials, with a mean response time of 961 (SD = 169) ms. We used a logistic regression approach to identify predictors of human decisions to accept the risky option (i.e. to gamble; see Fig.1b). As predicted by economic theories, the expected value (EV) of a gamble was a strong positive predictor of acceptance (t_{19} = 17.67, $p < 1 \times 10^{-12}$) in regression model 1 (RM1). Over and above this however, cumulative reward – that is, the sum of all outcomes obtained thus far in the temporal context – was a negative predictor of the tendency to gamble (t_{19} = -4.06, $p < 0.001$, RM1). In other words, the more participants' estimate of accumulated reward grew within a context, the less likely they were to choose the risky option (Fig. 1f). However, when controlling for several additional variables in regression model 2 (RM2, see methods), it became clear that this was not simply driven by changing preferences over time, as the overall trial number within a context (t_{19} = 0.003, $p > 0.99$, RM2), or trials elapsed within the scanner run of 75 trials (t_{19} = 0.76, $p > 0.45$, RM2), had

no impact on risk preferences. Nor was the tendency to gamble influenced by proximity to the forthcoming context or by any interactions thereof (all p-values > 0.14, RM2). Thus, decisions depended on economic factors alone – including latent estimates of the current level of resource, or "wealth" – rather than the passage of time elapsed across the experiment. Importantly, there was no effect of accumulated rewards over the 5 contexts (t_{19} = -1.06, p > 0.30, RM2), indicating that participants treated each context as a unique behavioral episode. In addition, there was no significant interaction between wealth and scanning runs elapsed, indicating that the effect of wealth was stable over the five scanning runs (t_{19} = 0.36, p > 0.72, RM2), Fig. S1.

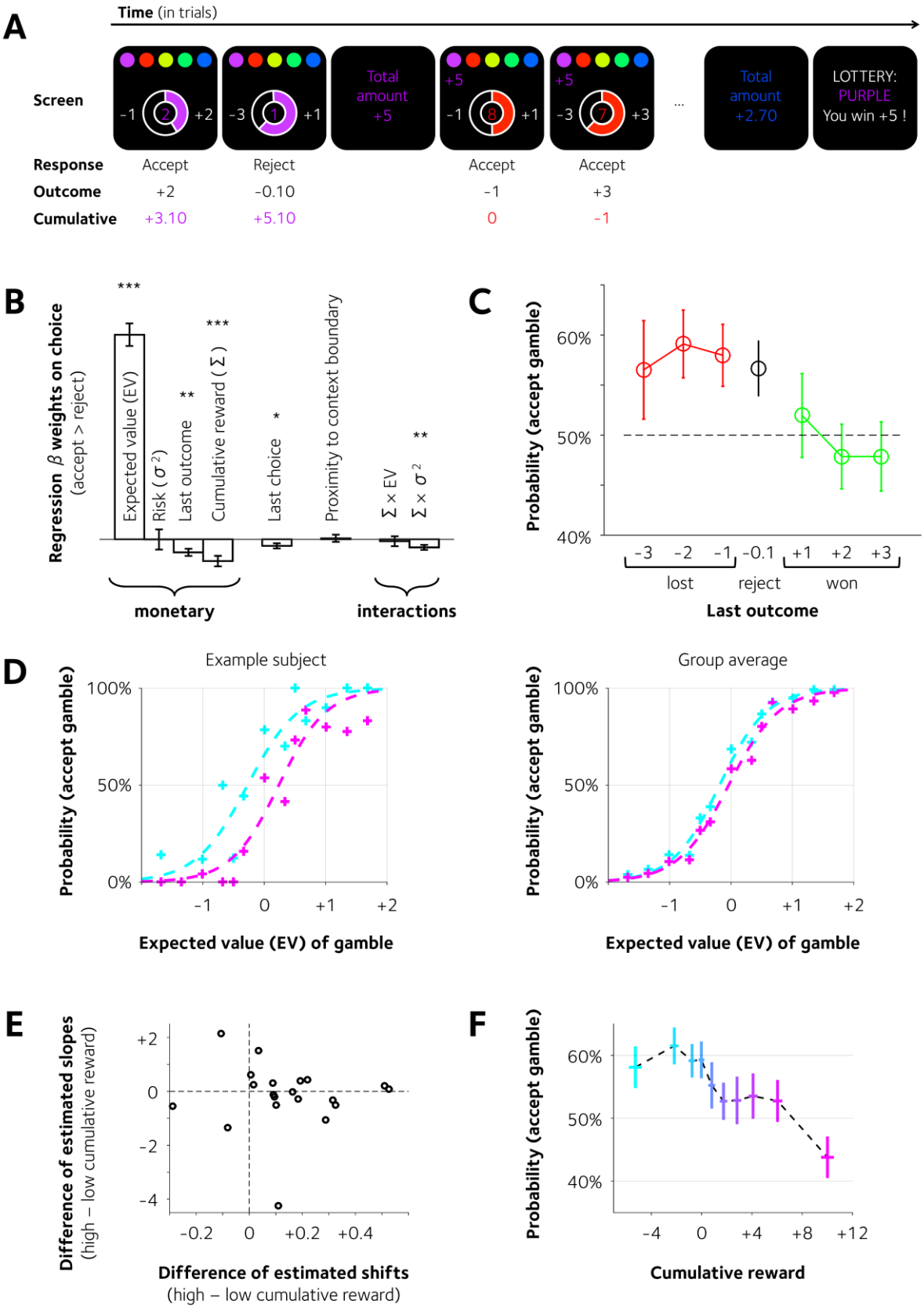


Figure 1 Task and behavioral results. (a) The task with an example sequence of responses, outcomes and their combination into cumulative reward. (b) Logistic regression of choice

in RM 1, where *** is $p < 0.001$, ** is $p < 0.01$, and * is $p < 0.05$ for a two-tailed t-test against zero with 19 degrees of freedom. Bar height is mean, and error bars are standard error of the mean (SEM). (c) Effect of most recent choice and outcome on accept probability on current trial. Circles are across-subject mean with SEM overlaid. (d) Example subject data (left panel) and group average of accept probability (right panel). Crosses correspond to mean accept probabilities for a given EV for below-median cumulative (light crosses), and above-median cumulative (dark crosses). Dashed lines correspond to best fitting sigmoid. Note that some subjects have sigmoids which are more shifted to the left (risk-seeking on average), or vice versa. (e) Differences in estimated sigmoid shifts and slopes for high versus low cumulative reward. Shift and slope differences are expressed as the estimates for high (above median within-participant) cumulative reward minus the estimates for low cumulative reward. Each circle represents one participant. (f) Accept probability as a function of cumulative reward (over the ten deciles). Error bars are SEM.

Overall, the risk ('outcome variance'; equation (10)) associated with each gamble did not influence choices ($t_{19} = -0.005$, $p > 0.99$, RM1). However, we observed an interaction between risk and cumulative reward, such that participants became more risk-averse with reward accumulation, as displayed by a tendency to select the safe option ($t_{19} = -3.14$, $p < 0.01$, RM1). The risk-aversion bias following reward accumulation was also observed at the level of successive trials, when mean responses were compared as a function of the immediately preceding outcome (Fig. 1c). Unlike humans and monkeys gambling for liquid reward (Hayden and Platt, 2009), our participants eschewed a "win-stay, lose-switch" strategy, and were instead less (not more) likely to gamble on the trial following a positive outcome ($t_{19} = -3.54$, $p < 0.005$, RM1).

In a subsequent analysis, we plotted how choices mapped onto the expected value of each gamble, separately for trials where accumulated reward was higher or lower (within-participant median split). Psychometric functions were shifted towards "safe" choices following higher reward accumulation, as indicated by a nonparametric test on the inflection point of a best-fitting sigmoid function

(Wilcoxon signed-rank test: $p < 0.005$). By contrast, there was no difference in slope ($p > 0.65$) for the psychometric functions (Fig. 1d-e). Taken together, these behavioral data show that human attitudes to risk varied dynamically during resource accumulation, with risky choices giving way to safe choices as wealth gradually accumulated. In other words, cumulative rewards (across an arbitrarily defined temporal context) are a critical factor determining attitudes to risk in human economic decisions.

BOLD correlates of momentary and cumulative reward. We began by measuring how the brain responded to the receipt of momentary reward, correlating BOLD signals with the monetary outcome received (GLM 1; Fig. 2a, Table S1). As anticipated, this analysis strongly implicated structures including the ventral striatum (left: peak (x,y,z) -14, 12, -6; right: peak 14, 12, -6) and a rostral portion of the medial prefrontal cortex; for consistency with the past literature we call this the ventromedial prefrontal cortex (vmPFC; sub-peak: -2, 52, -2). Note, however, that the peaks in striatum and vmPFC belong to one large cluster and are reported to aide understanding of the cluster's spatial extent. These reported peaks, and all others reported throughout the results section were situated within clusters of voxels that survived correction for multiple comparisons using a cluster-wise false discovery rate (FDR) approach (Chumbley and Friston, 2009; Genovese et al., 2002).

However, in order for risk preferences to vary with accumulated wealth in our task, the brain must have access to a latent variable encoding the current level of resources. Next, thus, we asked whether the vmPFC BOLD signal tracks estimates of accumulated rewards over the course of each temporal context. In GLM 1, we

thus also included a competing regressor encoding the sum of all outcomes obtained across the temporal context. This regressor encoded cumulative rewards up to and including the immediately preceding trial, but not the current trial, and was thus orthogonal to the predictor based on momentary outcome.

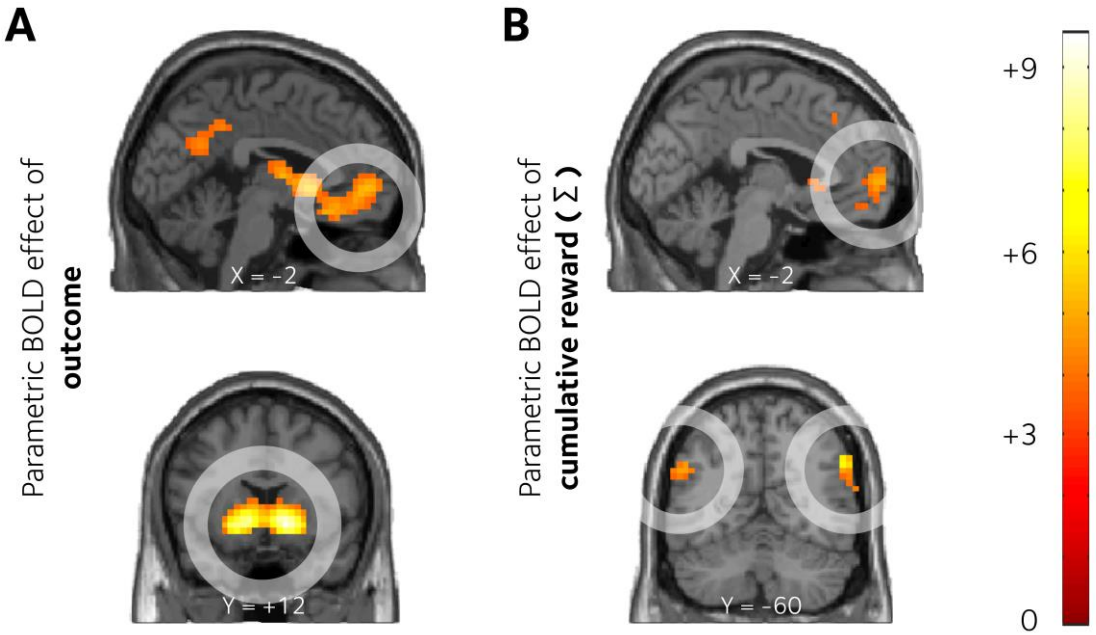


Figure 2 BOLD effects of outcome and cumulative reward. (a) Sagittal and coronal slices for the outcome contrast depicting the clusters in the vmPFC, and the ventral striatum (global peak). (b) Sagittal and coronal slices depicting the clusters vmPFC and angular gyrus (global peak). All images are on a common color scale with threshold at $p < 0.001$, unc.

This regressor also captured unique variance in the vmPFC BOLD signal (Fig. 2b), with a peak in a slightly more dorso-frontal region of the ventromedial prefrontal cortex (-6, 64, 2), as well as in the striatum (left peak: -14, 24, -2; right peak: 10, 20, 2). Both of these clusters survived FDR correction for multiple comparisons. The portion of the vmPFC that responded to current trial outcome and that which responded to accumulated reward were overlapping (Fig. 2a-b). Indeed, voxels within the vmPFC that responded to outcome at a threshold of $p < 0.001$

(uncorrected) were also on average sensitive to cumulative reward (t_{19} = 3.18, p < 0.005), even when ROIs were defined in a leave-one-out fashion over participants to avoid potentially circular inference. We also observed responses to accumulated rewards in the left angular gyrus (peak -54, -60, 30), and unexpected areas such as the right occipital lobe (38, -92, 14), a full list of activations is described in Table S2.

Computational model of risk attitude adjustment. Risk preferences during economic judgments are often characterized using Prospect theory (Kahneman and Tversky, 1979), a descriptive model of how losses, gains and probabilities are weighted before being combined to form a subjective internal estimate (expected utility). In Prospect Theory, the expected utility of accepting a gamble U(accept) is computed by combining the output of two transfer functions *w*() and *v*() that respectively map objective probabilities and relative monetary gains (g^+), losses (g^-) onto corresponding subjective quantities, as follows:

$$U(accept) = w(p) \times v(g^+) + w(1-p) \times v(g^-) \quad (1)$$

The expected utility of rejecting the gamble is related to the certain fixed loss (g^0) that is incurred on these trials

$$U(reject) = w(1) \times v(g^0) \quad (2)$$

The decision-relevant quantity, ΔU, is then computed as the difference between the accept and reject utilities:

$$\Delta U = U(accept) - U(reject) \quad (3)$$

Decisions are simulated by passing ΔU through a softmax choice function with slope s and bias b:

$$p(gamble) = \frac{1}{1+e^{-s[\Delta U - b]}} \qquad (4)$$

However, the key question for our purposes is how the changes in risk attitudes as a function of reward accumulation observed in behavior can be reflected in the subjective values, or utilities. Specifically, Prospect Theory postulates a mechanism by which each prospect is coded as a gain or loss relative to a reference point. In this study, we implicitly defined the reference point by framing the information on the screen as a positive or negative outcome. Such a model, however, would predict no change in risk attitudes as a function of reward accumulation – contrary to our behavioral finding. Here, we compare six models which are all based on Prospect Theory, with different conceptualizations of the reference point (summary Fig. 3a): Firstly, a "static" model in which decision-makers followed our implicit framing and represented the gamble as a deviation from the status quo. Secondly, a "wealth-dependent" model in which participants compare final wealth positions. Participants would thus compute the prospects relative to the status quo wealth (R):

$$g^+ = R + gain$$

$$g^- = R + loss$$

$$g^0 = R - 0.1 \qquad (5)$$

In this model, the reference R affects the perception of the monetary incentives of both accept and reject options prior to their transformation into a subjective utility signal, Fig. 3a.

Thirdly, three "curvature" models in which one of the parameters of the Prospect Theory function (the curvature of v(), the loss aversion parameter lambda, or the probability parameter of w()) at trial t were determined by a linear function of status quo wealth:

$$\theta_t = \theta_0 + \delta \times R_t \qquad (6)$$

where θ stands in for the parameter in question (from equations 12 and 13), and δ is the update rate.

Finally, we formulate an "adaptive gain" model which builds on a general framework for adaptive decision-making that is most often used to account for perceptual decisions (Cheadle et al., 2014). Here, the gains and losses g^+ and g^- map directly onto the gain and loss offered on each trial, without being normalized by R. However, monetary outcomes on each trial t drive an update of the bias term (b from equation (4)), such that it moves towards R:

$$b_{t+1} = b_t + \alpha \times reward_t \qquad (7)$$

where α is an update strength parameter. This adaptive gain model (Cheadle et al., 2014) thus updates the inflection point b towards the right (higher utility) when rewards are obtained from risky gambles, thereby reducing the probability that further risks will be run, and vice versa for losses, Fig. 3b. Hence, one can construe the reference point in this model as a threshold defined in utility space. Equivalently, the bias represents an additive increase in value of the reject option.

Here, we compared these six Prospect Theory models by fitting them to human responses and comparing their likelihoods using Bayesian model selection (Stephan et al., 2009) with leave-one-out cross-fitting over the 5 independent scanner runs. Bayesian model selection estimates the "exceedance" probability of a candidate model being the best-fitting model out of a selection of models by comparing not just the summed log-likelihoods, but also the pattern of fit across the cohort. Hence, a model with better fit in the summed log-likelihood will only have a high exceedance probability if it also fits best for a majority of subjects in the cohort. Cross-validation with separate training and test data reduces both overfitting and mitigates the necessity of correcting for any increases in model complexity due to the differing number of parameters in each model variant.

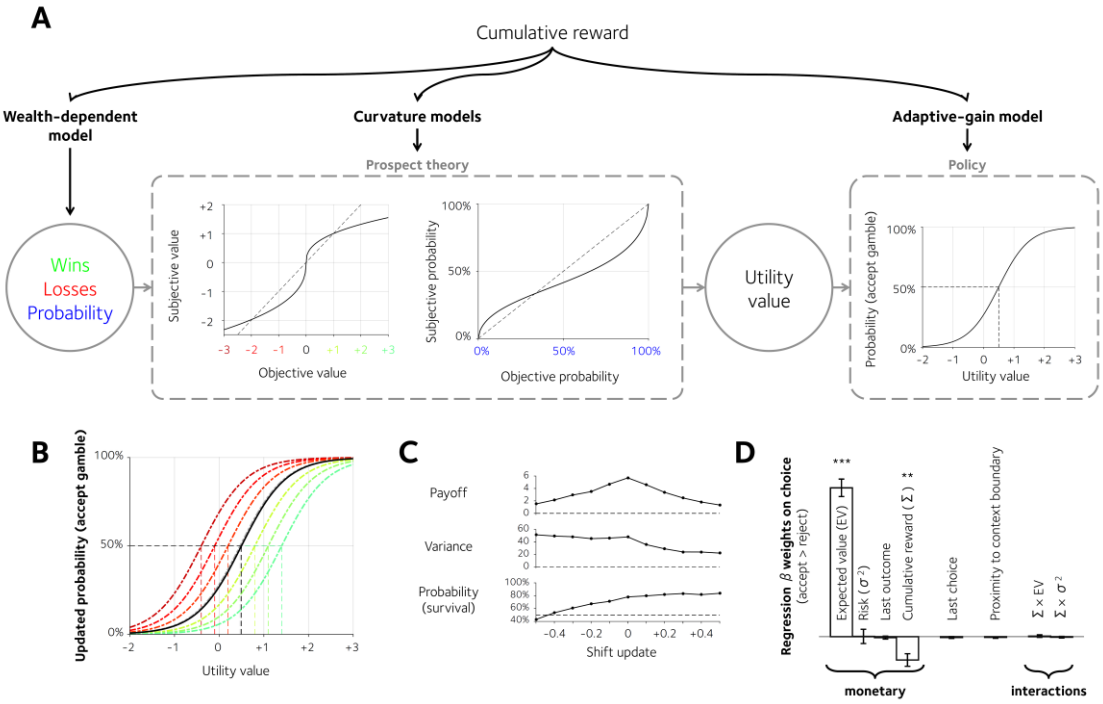


Figure 3 Computational models of behavior. (a) Depicts the rationale of Prospect theory, where wins, losses and probabilities are passed through nonlinearities and then combined into a utility value, upon which the decision policy is based. The arrows show a subset of models (labels from main text) and where cumulative reward changes their computation. (b) Adaptive gain model. The indifference point of the sigmoid is shifted from its initial

value (black) towards the right for incurred gains (green), towards the left for incurred losses (red) and for the certain loss incurred from reject choices (gray). (c) Policy simulation of a dynamic agent for an example sequence of 21 trials. Upper panel depicts the average payoff at the end of the sequence, the middle panel the variance of the payoffs across simulations, and the lower panel the probability of incurring payoffs above zero. All panels as a function of update parameter α. (d) Logistic regression of the choices predicted by the adaptive model in RM1. Bar height is mean, and error bars are SEM, *** is $p < 0.001$, and ** is $p < 0.01$ for a two-tailed t-test against zero with 19 degrees of freedom.

The adaptive gain variant of Prospect Theory fit better than its five competitors (see Table 1 for model comparison and mean parameters) on both the summed log-likelihood and the exceedance probability with $p = 0.96$ versus the second-ranked "static" model with $p = 0.02$. It also fit better than other model variants with (i) where R was calculated by weighting accumulated rewards via an exponential function, or (ii) a model postulating the reference point as an "aspiration level" which increased as participants progressed through the sequence such that participants aspire to higher wealth by the end of a context, Table 1, and Methods.

Model	-LL	Exceedance p	β	λ	γ	bias	slope	α	Others
Adaptive gain	2344.9	0.9611	0.69	1.46	1.42	0.22	7.75	0.017	
Static with α = 0	2394.2	0.0221	0.53	1.39	1.17	0.36	11.67	0	
Change in β	2370.0	0.0065	0.61	1.44	1.40	0.23	9.28		0.005
Change in λ	2407.1	0.0048	0.69	1.36	1.28	0.23	8.54		0.027
"Leaky" wealth-dependent	2421.3	0.0040	0.63	1.44	1.34	0.25	8.75		0.051
Change in γ	2570.9	0.0010	0.77	1.34	1.43	0.13	7.58		0.004
Wealth-dependent	2862.7	0.0003	0.99	1.04	1.15	-0.11	3.22		
Aspiration model	2930.8	0.0002	0.94	1.43	1.14	0.07	3.73		0.68

Table 1 Comparison and mean parameters of computational models based on Prospect Theory

Additionally, when we performed logistic regression on the choices predicted by the models, we found that the adaptive gain inspired model recreated the effect of accumulated reward (Fig. 3d). In contrast, the model with second-best fit – the "static" PT model – did not capture this aspect of the data. In summary, participants dynamically adjust their decision policies via a roving function that maps utility onto choices, rather than rescaling estimated utility according to the current level of wealth.

To identify candidate brain regions encoding decision information, we computed the log-likelihood ratio of choosing to gamble, which we here call "decision value". It reflected an unbounded estimate of choice probability, as follows:

$$DV = \log\left(\frac{p(gamble)}{1-p(gamble)}\right) \qquad (8)$$

Previous studies have demonstrated that during reward-guided decisions, BOLD signals (Boorman et al., 2009) and neuronal firing rates (Padoa-Schioppa and Assad, 2006) encode the relative value of a chosen and unchosen options. We were thus interested in how the neural representation of the DV depended on participants' choices, searching for voxels that encoded DV (a signed estimate of the value of gambling) differentially during accept and reject decisions. This DV × choice interaction (GLM 2) yielded negative correlations in the dorsomedial PFC (dmPFC; -2, 32, 42), as well as in the insula (left peak: -34, 20, -2, right peak: 34, 20 -6) and angular gyrus (54, -52, 42). These data are shown in Fig. 4a (similar results were

obtained when we used raw choice probability values, bounded between 0 and 1). Specifically, dmPFC and insula correlated positively with DV on reject trials, and negatively on accept trials, consistent with previously described encoding of the value of the unchosen option in these regions (GLM 3) (Boorman et al., 2009; Tsetsos et al., 2014) (see Table S3 and Fig. 4b). Of note, we obtained similar findings in GLM1 and GLM2 even when controlling for reaction time - as an additional, non-orthogonalized regressor - suggesting that our effects are not secondary to more general effects of choice difficulty.

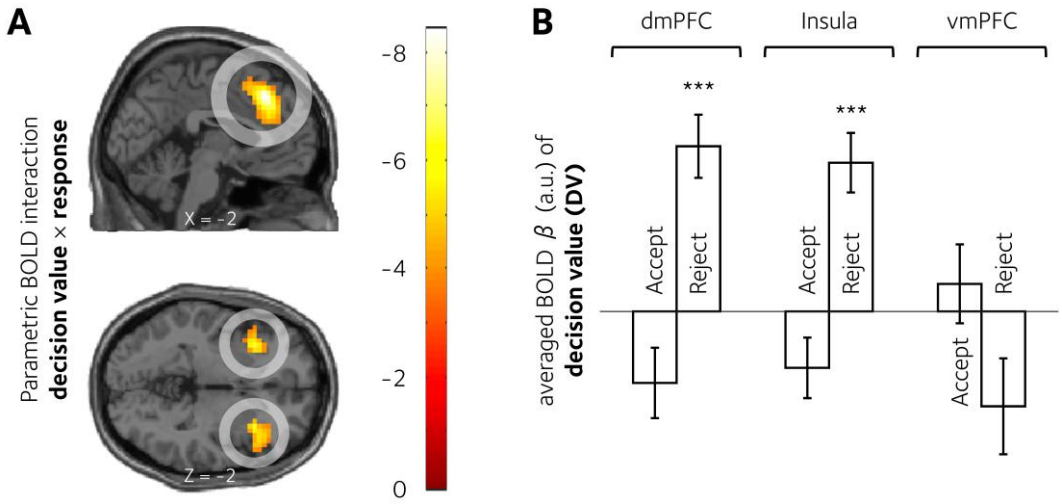


Figure 4 BOLD effects of DV by choice interaction. (a) Sagittal and axial slices depicting the dmPFC (global peak) and insula clusters on a common color scale. Images are thresholded at $p < 0.001$, unc. from GLM 2. (b) Encoding of DV (BOLD beta) in the dmPFC, insula, and vmPFC depending on choice from GLM 3. Data from these three regions was extracted from the ROI described in the Methods. Bar height is across-participant mean, and error bars are SEM. *** is $p < 0.001$ for a two-tailed t-test against zero with 19 degrees of freedom.

Simulating choices as a function of model parameter alpha. In order to investigate the potential costs and benefits of the dynamic policy adopted by our participants, we conducted a simple simulation. We computed likely mean and variance of

cumulative outcomes as a function of different values of α, the update parameter in our adaptive model that controls how sharply risk aversion should grow (or fall) with wealth (Fig. 3a-c; see Methods). These simulations show that although an "optimal" agent (with $\alpha = 0$) will on average maximize expected value over the course of the experiment, this comes at the expense of considerable variability in overall cumulative outcomes. By contrast, averaged over simulations, an agent with $\alpha > 0$ obtains a lower mean cumulative reward tally, but is less likely to experience a resource level that falls below the "survival" threshold (i.e. zero). The same holds for suboptimal agents who exhibit Prospect Theory preferences. In other words, humans and other animals may pursue economic decision policies – and adopt risk attitudes – that lead to satisfactory but suboptimal levels of cumulative outcome, in order to avoid the risk that resources drop below a perilous critical level.

BOLD correlates of temporal structure. Our task featured a unique temporal structure, which required participants to keep track of a trial's position within each behavioral episode. Although proximity to a new episode did not influence risk attitudes, neural data suggested that participants were sensitive to the passage of time within each episode. Specifically, the regressor encoding the proximity to context termination (from GLM 1) captured positive BOLD responses in several cortical areas, including the anterior cingulate cortex (ACC; 10, 28, 38), supplementary motor area (SMA; 6, 8, 62), right dorsolateral PFC (22, 56, 6) – areas which have been implicated in the decision-making circuitry - as well as in the occipital lobe (global peak: 10, -72, 2), Fig. 5, Table S4.

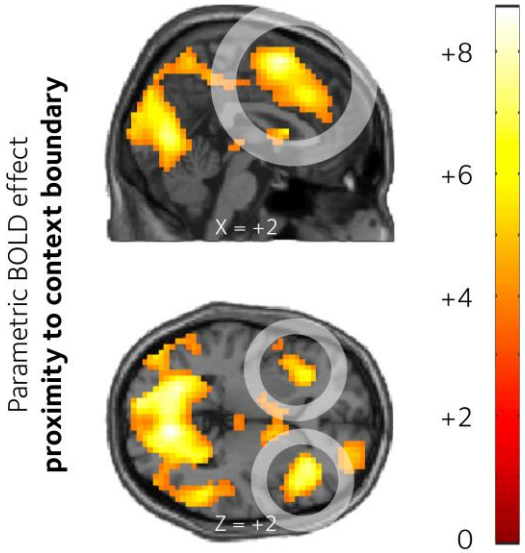


Figure 5 BOLD effect of proximity. Sagittal and coronal slices depicting clusters (circles) found for the DV by choice interaction in Fig. 4. Images are on a common color scale, and thresholded at $p < 0.001$, unc.

Discussion

Humans decisions are sensitive both to the expected value of an economic prospect, and to the level of risk (or outcome variance) that it entails. Typically, humans are averse to risk, and this phenomenon can be accounted for with the classic framework provided by Prospect Theory, in which losses "loom larger" than gains due to a steeper subjective loss function. Here, we observed that humans making successive risky gambles become more averse to risk as they accumulated rewards within a discrete temporal context. This behavior was best captured by a computational simulation in which cumulative rewards engender a dynamic shift in the transfer function that maps subjective values (utilities) onto choices. Of note, this account fared better in describing human choices and the pattern of BOLD activations than classic Prospect Theory models, which propose instead that status

quo wealth modulates the subjective valuation of monetary outcomes, rather than acting on their subsequent mapping to decision signals.

The budget rule and loss aversion

Why does risk aversion increase with cumulative rewards? One ecologically plausible explanation for this phenomenon is that economic decision policies have evolved not simply to maximize total expected value, but rather to minimize the probability that cumulative rewards will fall below some threshold value. This policy will promote survival in naturalistic settings, where death ensues when resources fall towards zero. For example, animals that forage over the diurnal cycle may have evolved risk attitudes that maximize the likely lowest level of resources that will be obtained during the day, in order to provide the best chance of surviving through the night (Caraco, 1981; Kolling et al., 2014). Consistent with this explanatory framework, behavioral ecologists have noted that asymmetric, convex value functions such as those proposed by Prospect Theory predict that risk aversion should naturally increase as resources grow further from the survival threshold (McDermott et al., 2008). Evidence for this resource budget rule has been inconsistent, however (Kacelnik and El Mouden, 2013), and the precise mapping of resource level onto evolutionary fitness is a subject of debate (Houston et al., 2014). Similarly, it has been noted that taxi drivers set themselves a target income for the day beyond which they stop working (Camerer et al., 1997; Kőszegi and Rabin, 2006). The policy that humans adopt here achieves a similar feat: it minimizes the chance of running a loss at the end of the episode, *as if* participants

had aspired to achieve a net positive outcome (Lopes and Oden, 1999). This model would be captured by the "wealth-dependent model", in which positive wealth pushes the gambles into the domain of gains (less risk-taking), while negative wealth predicts more risk-taking. Here we also tested a variant of the "aspiration model", which changed the aspiration level according to the position in the sequence to avoid the assumption that participants aspire to the same discrete level of wealth at every point in time. In contrast, however, the adaptive gain model does not impose a fixed threshold or reference point in the space of "final wealth" but rather describes how each individual outcome affects subsequent risk attitudes. Intuitively, this can be achieved by the policy that humans adopt here: to shift from risk-seeking (when resources are low, and there is little to lose) to risk aversion (as resources grow, and a stable status quo wealth has been achieved) over the course of a behavioral episode.

Prospect Theory and the reference point

In Prospect Theory, outcomes are evaluated relative to a reference point that encodes prospects as either changes in status quo wealth or an aspiration wealth, leading to differing subjective utilities (and consequent choices) with alternate framings of a risky prospect (Tversky and Kahneman, 1981). Prospect Theory deliberately leaves the reference point vague to facilitate a wide range of applications. However, to our knowledge, the reference point always changes the value function in the classical formulation of Prospect Theory. For instance, a recent account describes Prospect Theory as "efficient perceptual distortion",

where the reference point is the mode of the distribution of expected outcomes (Woodford, 2012). An alternative, however, is that the status quo does not influence how outcomes are evaluated, but instead drives a decision bias that reduces the frequency of risky relative to safe choices, given equivalent subjective utility. Here, we pitted these accounts head-to-head using statistical model comparison techniques, and found that the latter provides a better explanation of human choices. While the curvature models (especially the one based on v() and the loss aversion parameter, lambda) could in principle capture the same qualitative effects, they fared worse in quantitative model comparison. Our findings are thus consistent with a general framework that has suggested that internal decision variables – including estimates of both sensory signal strength, and economic value – are subject to an adaptive mechanism that confers time-dependent normalization (Carandini and Heeger, 2012; Cox and Kable, 2014; Parducci, 1965; Rangel and Clithero, 2012; Summerfield and Tsetsos, 2015). One simple description of this mechanism is provided by a model in which the function by which internal variables map onto choices adjusts dynamically to control behavior. This account is inspired by earlier work which described how context modulates decision signals in a psychophysical setting (Cheadle et al., 2014).

The neural underpinnings of reward accumulation

Analysis of BOLD signals obtained during the task shed further light on the mechanisms by which cumulative resources influenced risky choices, and helped us chart their neural underpinnings. Previous research has extensively explored the

neural mechanisms by which subjective values and choice variables are encoded during decisions about risky prospects, such as monetary gambles (De Martino et al., 2009; Hare et al., 2011; Knutson et al., 2005; Lim et al., 2011; Plassmann et al., 2007; Tom et al., 2007). However, past studies have overlooked the question of how brain responses vary as rewards are accumulated over time. Two particularly relevant studies have either imposed a priori target levels (Kolling et al., 2014), or did not provide choice feedback to participants, which renders analyses of reward accumulation over time impossible (Symmonds et al., 2010). Here, we show that the ventral portion of the medial prefrontal cortex (vmPFC) encodes not only momentary value signals, as previously demonstrated, but also the cumulative tally of rewards obtained over the past history defined by a discrete temporal context. This occurred even though ongoing resource levels were not signaled to the participant within each context, suggesting that the vmPFC encodes a latent, internal tally of cumulative wealth. This finding is consistent with a view of the vmPFC as tracking outcomes as they pertain to long-run goals, rather than to shorter-term hedonic outcomes (Tsetsos et al., 2014).

Structuring the world in time

In our task, the temporal context was defined in such a way that it was irrelevant for the reward-maximizing policy, which was simply to accept gambles with net positive EV, irrespective of contextual factors such as cumulative rewards, or the time elapsed within a context, or over the course of the experiment. However, both risk attitudes and accompanying neural signals varied with cumulative rewards over

a distinct temporal context. Indeed, although time elapsed within a context did not influence decisions, BOLD signals in diverse cortical regions varied substantially as each context unfolded. It may be that humans have a strong tendency to break down an ongoing, unstructured episode into discrete temporal "chunks", and to monitor resource levels within these chunks, as proposed by hierarchical models of reward-guided learning (Botvinick et al., 2009; Holroyd and Yeung, 2012).

Notably, in our experiment, the cumulative reward tally was signaled to the participant at the end of each temporal context, motivating participants to monitor their cumulative outcomes up to the point at which they were revealed. Previous studies have shown that signals in dmPFC and underlying cingulate gyrus track proximity to an interim goal state where rewards or feedback are provided(Hayden et al., 2011; Ribas-Fernandes et al., 2011; Shidara and Richmond, 2002). In one particularly relevant study, dmPFC tracked the "risk pressure" that increased as participants struggled to reach an experimenter-determined reward threshold, below which cumulative rewards were lost (Kolling et al., 2014). Our study shows that in the absence of an externally imposed "survival" threshold, humans nevertheless pursue a wealth-accrual policy that is strongly modulated by the need to conserve constant positive cumulative reward levels.

Methods

Participants.

Twenty healthy volunteers (7 men, mean age: 23.25 years, SD = 3.4) with no history of psychiatric or neurological disorders participated in the study. Participants gave informed consent prior to scanning, and the study was approved by the ethics committee of the University of Granada. Participants' compensation was 25€, plus a performance-based bonus.

Task

Each trial offered the choice between accepting and rejecting a monetary gamble (see Fig. 1a). A central pie-chart indicated the probability of winning by a colored area (see context manipulation below), and the probability of losing as a black area. The possible monetary losses and gains were displayed to the left and right of the screen, respectively. Gain and loss magnitudes ranged from 1€ to 3€, and -1€ to -3€. Probabilities used were 1/3, 0.5, and 2/3. Gambles were sampled uniformly at random from these values and were always mixed (gain and loss). Accept choices led to the gamble being played out, while reject choices led to a small certain loss of 0.10€.

Context manipulation

The 75 trials in each of the five scanner run (375 trials in total) were allocated to successive "contexts", each associated with an unambiguous display color. Each context was of variable length (3-22 trials). Whilst a participant was in a given context, the pie-chart indicating reward probability would be shaded in the current context's color. Additionally, a countdown in the center of the screen conveyed the

number of trials (including the current trial) remaining until the end of that context. The countdown font color also corresponded to the current context's color. The sequence of contexts in each session was displayed as 5 colored disks on the top of the screen. Whenever participants concluded a context, their cumulative earnings of that context were displayed on a 'bonus screen' for a duration of 2s. These earnings would then be displayed below the corresponding disc, for the remainder of the scanner run.

Trial sequence. Stimuli were presented for a duration of 3s. Within the first 2s, participants could respond by pressing a button with their left or right hand. Correspondence between response (accept or reject) and hand was counterbalanced across participants. Failure to respond resulted in a penalty of 1€. Immediately after their response, trial outcome was displayed in the middle beneath the pie chart and was visible for the remainder of the trial duration. No feedback about accumulated rewards was given until the 'bonus screen', which informed participants after the conclusion of context about the accumulated rewards within that context. All intervals between screens were drawn from a uniform distribution between 3s and 7s.

Reward at the end of a run: lottery. After each run, one context was selected for payment in a lottery. Unbeknownst to participants, this procedure was pseudorandomized so as to avoid very high gains and very high losses, which would be incompatible with local ethical guidelines. Thus, only contexts (60.2%) in which cumulative earnings ranged from -4€ to +6€ were used for the lottery and chosen with uniform probability. Participants did not report suspicion that not all contexts entered the lottery procedure. At the end of the experiment participants were paid

the sum of all five lotteries, on average 13€ (SD = 5.64€, range: 2.50€ to 24.10€). No participant finished with negative wealth.

Training. Directly prior to fMRI scanning, participants completed one run of the task, but no rewards were paid out for this session and all analyses are based on the fMRI data only.

Behavior: Data analysis

The following regressors were transformed before entering analyses:

Accumulated rewards. For each trial, rewards within its corresponding context were summed up to, but not including, the current trial. Accumulated rewards were always zero on each first trial within a context, and orthogonal to trial outcome (r = -0.02).

Proximity to next color. The proximity regressor was constructed as the reciprocal of the number of trials to the next context, thereby bounding values between zero and one.

Expected value (EV) was calculated using the following formula:

$$EV(x, p, y, 1-p) = \ px + (1-p)y \qquad (9)$$

where x corresponds to a possible gain, y to a possible loss, and p to the probability to win.

Risk as outcome variance was calculated as follows:

$$Risk(x, p, y, 1-p, EV) = p(x-EV)^2 + (1-p)(y-EV)^2 \qquad (10)$$

Logistic regression of choice. Participants' choices were analyzed at the single-trial level, using a logistic regression to predict the probability of choosing to gamble. The estimated beta weights across the cohort were then used for a two-tailed one-sample t-test against zero. Regression model 1 contained the following regressors: EV, risk, accumulated rewards, last outcome, last choice, proximity to context boundary, EV by accumulated rewards interaction, and risk by accumulated rewards interaction. This model is presented in the main text in Fig. 1b. Regression model 2 added several controls to RM1: accumulated rewards as a running sum of the previous eight trials (across contexts), accumulated rewards across the whole scanning run, number of trials elapsed in scanning run, number of trials elapsed in context, the amount of trials in the current context, the financial bonus from the previous run ("last lottery"), number of scanning runs elapsed, and interactions of proximity by EV, proximity by risk, and accumulated rewards by scanning runs elapsed. None of these additional regressors were significant (all $|t_{19}| < 1.38$, $p > 0.18$, Fig. S1).

Behavior: Model specification and comparison

All models shared the following features of the PT utility functions (Kahneman and Tversky, 1979), *w*(p) and *v*(x):

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma}+(1-p)^{\gamma})^{1/\gamma}} \quad (11)$$

$$v(x) = x^{\beta}, for\ x \geq 0 \quad (12)$$

$$v(x) = -\lambda(-x)^{\beta}, for\ x < 0 \quad (13)$$

In all models, utility was mapped onto choice using equation (4) from the main text. The function approximates a deterministic step function for $s \rightarrow \infty$ with the step at *b*.

Wealth-dependent Prospect Theory model with exponential leak. Working memory capacities might limit an accurate representation of accumulated rewards, so we tested another model, in which accumulated rewards were calculated as the exponentially weighted sum (EWS) of all previous outcomes. It computes utility with equation (5), and substitutes $R = R_{EWS}$ in the computation of g^+, g^-, and g^0 at trial t. A parameter zeta governed the weight of the most recent outcome as follows:

$$R_{EWS}(t) = \zeta \times reward(t-1) + (1-\zeta) \times R_{EWS}(t-1)$$

$$R_{EWS}(1) = reward(0) = 0 \qquad (14)$$

Adaptation in Prospect Theory parameters. The curvature models, described in equation (6) change one of the Prospect Theory parameters β, λ, or γ as a linear function of accumulated rewards.

Adaptive model. The adaptive gain model updates the inflection point *b* of the sigmoid mapping utility onto behavior from trial to trial according to equation (4). A positive update parameter shifts the indifference point to the right when participants incurred a gain and vice versa for losses (and the reverse for negative parameters). Thus, participant will become risk-averse after a series of gains and more risk-seeking after losses. On each first trial within a new context, the indifference point was reset to its initial value as estimated by the model. This procedure is warranted, since participants treated each context as a new episode

and there was no effect of accumulated rewards across contexts in RM 2 (t_{19} = -0.22, p > 0.82).

Aspiration model. We tested another model, in which participants aspire to a certain level of wealth depending on their position in the sequence. Gains and losses are added to current wealth and then compared to the aspiration level:

$$value = g^{+} + R - c \times trialcount \qquad (15)$$

Model comparison. Parameters were fit separately for each participant in a cross-validation framework on the basis of four out of the five scanner runs and then tested on the left out run, such that each run functions as independent data set once. MATLAB's simulated annealing toolbox (Kirkpatrick et al., 1982) was used to find the optimal parameters which maximized log-likelihood, given by:

$$LL = \sum_{i=1}^{n} \ln\left(p(choice \mid \theta)\right) \qquad (16),$$

where *p(choice* | $\boldsymbol{\theta}$*)* is the predicted probability of making the same choice as the participant given parameter set $\boldsymbol{\theta}$. Non-response trials (n = 56 out of N = 7500) were excluded from all models. $\boldsymbol{\theta}$ was constrained to:

$$0 < \beta, \lambda, \gamma \leq +5,$$

$$-1 \leq b, \alpha \leq +1,$$

$$0 < s \leq 20,$$

$$0 < \zeta \leq 1,$$

$$-10 \leq c, \delta \leq +10 \qquad (17)$$

Behavioral models were compared using SPM's Bayesian model selection function. The function's input was an n-by-k matrix - where n is the number of participants and k is the number of models -, which contained the log-likelihoods of each participant for each model. A model was deemed better than its competitors when its summed log-likelihood across participants and its exceedance probability - as estimated by SPM's function - were higher than all others.

Simulation

We simulated how the update strength α influences the mean and variance of cumulative earnings within an episode. An agent imbued with variable degrees of α played through N = 10,000 simulated runs (75 trials each) of the task. For simplicity, the agent based its choices on a trial's EV, which is a special case of PT utility with linear transfer functions. The indifference point was set to EV = -0.1, where both accept and reject choices yield the same expected payoff, while the slope was set to 10^{16}, corresponding to a step-like function. The range for α was set to $-0.5 \leq \alpha \leq 0.5$, where $\alpha = 0$ corresponds to a 'static' agent that is optimal with respect to reward maximization.

FMRI: Data acquisition

MRI data were acquired on a 3T Siemens scanner. T1 weighted structural images were recorded directly prior to the task using an MPRAGE sequence: 1 × 1 × 1 mm^3 voxel resolution, 176 × 256 × 256 grid, TR = 1900ms, TE = 2.52ms, TI = 900ms. Each fMRI image contained 32 axial echo-planar images (EPI) acquired in

descending sequence with a voxel resolution of 3.5mm^3 isotropic, slice spacing of 4.2mm, TR = 2000ms, flip angle = 80°, and TE of 30ms. 1700 EPI were recorded per participant, with the exception of three participants (1478, 1610 and 1644 images), resulting in a scanning time of about 57 minutes. Data were pre-processed using SPM8 (Wellcome Trust, London). Data were corrected for motion artifacts using the ArtRepair toolbox (http://cibsr.stanford.edu/tools/human-brain-project/artrepair-software.html). As EPI acquisition used a descending sequence, images were corrected for slice time acquisition with the middle slice (at TR/2 = 1s) as reference to minimize interpolation errors (Sladky et al., 2011). Scans were realigned to the first scan within each session. The anatomical scan was co-registered to the mean of all functional images. Anatomical scans were normalized to the standard MNI152 template brain. The functional EPI images were resampled to 4x4x4mm, then normalized and smoothed with a full width half maximum Gaussian kernel of 8mm.

FMRI: Data analysis

The five scanning sessions were concatenated and constants included in the GLMs to identify runs and account for differences in mean activation and scanner drift. Late events (less than 16s before end of run) were removed to avoid fitting a stimulus across two runs. Stimulus onsets were incremented by 1s to account for slice timing correction to the middle slice, which occurred at TR/2 = 1s after stimulus onset (Sladky et al., 2011). Microtime onset was set to the first slice. All GLMs used stick regressors with zero duration locked to the onset of the gamble (including the parametric modulation of outcome, because feedback followed approximately ~1s after choices). Similar results were obtained using regressors of duration 3s (length of trial). We also included the 6 motion parameters derived

from pre-processing as nuisance regressors, as well as the temporal and dispersion derivatives of the standard hemodynamic response function (HRF). Automatic orthogonalization was switched off. Data were analyzed with SPM12. All contrasts were constructed as simple t-contrasts with first-level t-maps as input. We only report clusters that fell below an FDR-corrected p-value of 0.05, as reported by SPM's standard results table with a setting of cluster extent to 10 voxels or more and a voxel-wise threshold of $p < 0.001$. In addition, we report all main results using the statistical non-parametric mapping toolbox (SnPM13; `http://warwick.ac.uk/snpm`) in Fig. S2. The settings for the SnPM analysis were the defaults: 5,000 permutations, variance smoothing of 8mm isotropic (as recommended for degrees of freedom smaller than 20), cluster forming threshold of 0.0001 (a Z-value of 3.71), and cluster-wise family-wise error rate of 0.05. Data were visualized using the xjview toolbox for SPM (http://www.alivelearn.net/xjview).

General Linear Model (GLM) 1 included: choice (accept > reject), trial outcome, accumulated rewards, and proximity to next context, all as parametric modulators time-locked to stimulus onset. The goal of GLM 1 was to identify correlates of accumulated rewards over and above correlates of trial outcome.

GLM 2 included: the GLM1 regressors, DV as given by equation (8) in the main text of the best-fitting behavioral model, and an interaction of DV and choice. Here, DV was calculated after fitting to the entire dataset. The goal of GLM 2 was to identify candidate regions for the late stage process by which risk preferences change with accumulated rewards.

GLM 3 split trials into accept and reject trial conditions, and estimated the betas separately for each condition: trial outcome (for accept trials only), accumulated rewards, proximity to next context, and DV. We used GLM 3 to investigate the pattern of activations identified in the DV by choice interaction from GLM 2.

All models also included the context outcomes - the 'bonus screens' - with parametric modulation of context outcome time-locked to bonus screen onset.

Region of interest (ROI) extraction. ROI were extracted in a leave-one-out procedure: Each participant's mask was estimated on the basis of all other participants, while leaving out the current participant. ROI masks were then constructed by taking all significant voxels (with $p < 0.001$, unc.) that fell within the target region (given by xjview's database), e.g. within medial PFC. From GLM 1, we extracted an ROI for trial outcome in the vmPFC. From GLM 2, we extracted two ROI from the DV by choice interaction contrast: one in the dmPFC, and another encompassing the bilateral insula.

Chapter Four: A network for computing value equilibrium in the human medial prefrontal cortex

Abstract

Humans and other animals make decisions in order to satisfy their goals. However, it remains unknown how neural circuits compute which of multiple possible goals should be pursued (e.g. when balancing hunger and thirst) and how to combine these signals with estimates of available reward alternatives. Here, humans undergoing functional magnetic resonance imaging (fMRI) accumulated two distinct assets over a sequence of trials. Financial outcomes depended on the minimum cumulate of either asset, creating a need to maintain "value equilibrium" by redressing any imbalance among the assets. BOLD signals in the rostral anterior cingulate cortex (rACC) tracked the level of imbalance among goals, whereas the ventromedial prefrontal cortex (vmPFC) signalled the level of redress incurred by a choice, rather than the overall amount received. These results suggest that a network of medial frontal brain regions compute a value signal that maintains value equilibrium among internal goals.

Introduction

Canonical models in psychology, economics, and machine learning assume that decisions are made in order to maximise expected reward. One popular view posits that the brain evolved dedicated structures that represent the value of stimuli or actions. For example, in non-human primates and rodents, neuronal firing rates in the orbitofrontal cortex scale with the value of primary reinforcers (Padoa-Schioppa and Assad, 2008; Rich and Wallis, 2016; Strait et al., 2014). In humans, BOLD signals in the orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (vmPFC) code for the value of diverse assets including food items (Hare et al., 2011, 2009), money (De Martino et al., 2006; Frydman et al., 2014) and social cues (Lin et al., 2012). This has been interpreted as revealing a domain-general value code or "common neural currency" in this region. Representing diverse prospects on a common value function may allow organisms to estimate the value of multidimensional prospects and to make decisions among otherwise incommensurable goods (Chib et al., 2009; Kable and Glimcher, 2009; Kim et al., 2011).

However, in natural environments, survival depends on the minimum level of any one of a number of competing internal needs. For example, a hungry animal needs food more than water, whereas the reverse is true for a thirsty animal. As the world changes dynamically, distinct internal assets (e.g. satiety or hydration) are continuously being depleted and replenished, so that equilibrium among internal resource levels needs to be monitored and maintained in neural circuits for valuation and choice (Cannon, 1929; Keramati and Gutkin, 2014; Korn and Bach, 2015). A related framework for considering reward-guided choices, thus is that

animals strive to satisfy multiple needs in parallel, maintaining a balance among relative asset levels, such as satiety, warmth, or reputation (O'Reilly et al., 2014). This class of model requires that value is coded on multiple distinct axes, each pertaining to a currently relevant goal (rather than being represented on a single monolithic value function) that mapping from these axes to choices occurs rapidly and flexibly (Valentin et al., 2007), and that axes interact and compete for action selection (Hunt and Hayden, 2017). In the current work we asked how the brain evaluates which of multiple possible goals should be pursued at any one time, and how it dynamically updates goal values according to time-varying resource levels.

This question has been hitherto unaddressed because most previous studies have used paradigms (e.g. bandit or food choice tasks) that involve maximisation of a single asset, rather than tasks that require the balancing of multiple assets. Nevertheless, there is reason to predict that the OFC/vmPFC may play a role in evaluating stimuli in the context of internal goal states (e.g. satisfy hunger or quench thirst). Firstly, animals with OFC lesions continue to choose actions that lead to devalued outcomes, as if they were failing to maintain or follow currently active goals (Bouret and Richmond, 2010; Burke et al., 2008; Izquierdo et al., 2004). Secondly, humans and primates with OFC lesions fail to integrate multiple attributes of a stimulus when making decisions, implying that this structure plays an active role in constructing composite value estimates from available stimulus dimensions (Fellows and Farah, 2007; Gläscher et al., 2012; Izquierdo et al., 2004; Wallis, 2011). Finally, vmPFC tracks the internal state given by cumulative satiety or wealth over a sequence of trials, a critical requirement for computing which assets may be most needed (Juechems et al., 2017; Tsetsos et al., 2014).

We reasoned that the medial OFC/vmPFC and interconnected areas of the medial prefrontal cortex might allow animals to compute ongoing levels of competing assets, and dynamically adjust choices to both replenish immediate needs and guard against future scarcity. We tested this using a task in which participants were encouraged to maximise the minimum of two assets (animals in a virtual zoo). We report evidence for a new frame of reference for human neural value signals, with rACC encoding the imbalance among assets (or goals), and the vmPFC coding how this imbalance is redressed by a choice.

Results

Participants (n = 21) performed a "virtual zookeeper" task in the fMRI scanner. The task involved managing a zoo during contexts of variable length (10-20 trials) that housed two assets (lions and elephants). On each trial, participants were offered the opportunity to expand their zoo by a variable number of animals (e.g. 4 lions or 2 elephants). Subsequently, trialwise reward was offered in proportion to the minimum number of animals in the zoo (e.g. if they had a total of 12 elephants and 8 lions in the zoo, they received 8 points) creating a pressure to redress any imbalance in the assets. To mimic the autocorrelation in natural environments, we created "streaks" in which numerically more of one species (4-6 uniform) were offered than the other (1-3 uniform); these numbers reversed with probability 0.3 on each trial. This manipulation ensured that the optimal policy was neither to always choose the more plentiful animal, nor to always satisfy immediate needs (redress the imbalance), but to vary these strategies according to the quality of the offer, the need to redress, and the time remaining in the block (see Fig. 1a).

Human choices are driven by both offer values and internal state. Participants received 84.3% of the maximum attainable cumulative reward as calculated by numerical simulation (see Methods; **Fig. 1b**). This was significantly better than chance (mean: 59.6%, Wilcoxon sign-rank test, $p < 0.001$), better than a fixed strategy of always choosing the higher offer ("greedy"; mean: 74.3%, sign-rank test, $p < 0.001$), and better than a model that used a fixed strategy of always satisfying immediate needs, assuming perfect memory for the tallies of animals in the zoo ("redress"; mean: 81.8%, sign-rank test, $p < 0.03$).

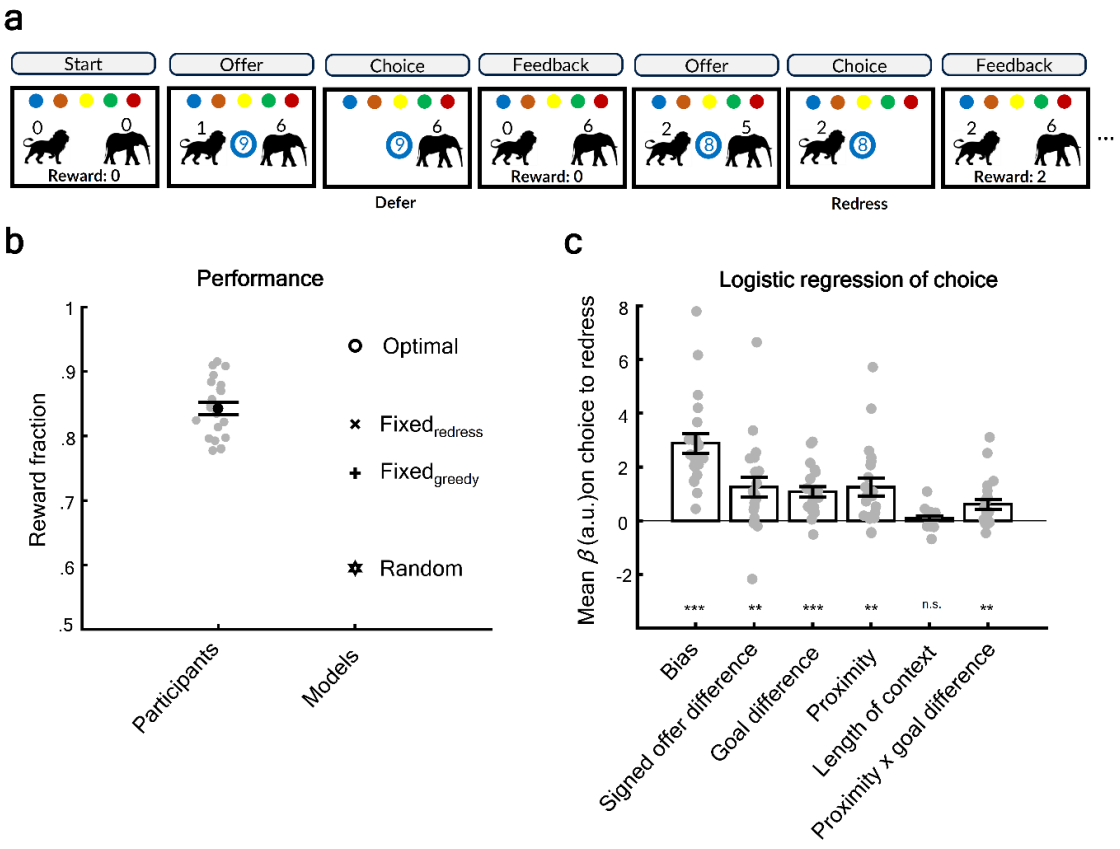


Figure 1. Task and behavior. (A) Task with 2 example trials from the first block in a scanner run. The number of animals on offer ("offer" screen), the number of animals chosen ("choice" screen) and the number of animals in the zoo ("feedback") are shown numerically above a picture of a lion and elephant. Coloured circles (top) indicate the zoos (blocks) for the current run, with the central circle coloured according to the current block. The central number indicates the number of trials (including the current trial) remaining before the end of the block, i.e. the "countdown". Trialwise reward was indicated on the feedback screen

as the minimum of the two assets. Background screens were grey during testing, but are shown in white here for illustration. (B) Black circle indicates average earnings as fraction of maximum possible; black error bars are standard error of the mean (SEM) and grey circles show the data from individual participants. Other markers indicate averaged reward fractions for various models. (C) Beta coefficients from a logistic regression of various predictors on choice to redress. "Proximity" was coded as 1 over the number of trials left in a context, "length of context" reflected the total number of trials in a context (10-20). Bar height indicates mean beta, error bars are SEM, grey circles represent individual participant data. Significance was assessed using a two-tailed t-test against zero. *** is $p < 0.001$, ** is $p < 0.01$, * is $p < 0.05$.

These data suggest that participants were not simply maximising their receipt of animals, or myopically seeking to balance their assets, but rather that choices were jointly driven by offer quality and relative needs. To test this contention more formally, we defined choices as either "defer" or "redress", where the latter trials were those on which participants chose to acquire more animals of the species with the lower current tally (mean = 80.85%, SD = 10.78%). We then used a logistic regression model to identify the factors that predicted the decision to redress or defer. Although participants were strongly biased to choose to redress ($t_{20} = 7.74$, $p < 1 \times 10^{-6}$), this tendency was stronger when the (signed) offer difference was higher, i.e. when more of the currently minimal asset was offered ($t_{20} = 3.36$, $p < 0.005$), and also stronger when there was greater need to redress, i.e. when the level of imbalance among the two assets (or "goal difference") was higher ($t_{20} = 5.61$, $p < 1 \times 10^{-4}$). In other words, participants were neither greedily choosing the highest offer, nor slavishly satisfying immediate needs, but rather trading off the relative offer values and their internal needs or goals when making decisions. This was also evident in a regression of reaction time (Fig. S4b). We also verified that a simple bias to redress was not the optimal policy in our task; indeed, those

participants with an overall stronger tendency to redress tended on average to reap lower rewards (r_{21} = -0.36, p > 0.94, one-sided test).

The dACC encodes offer difference and the rACC encodes goal difference. Behavioural analyses indicated that decisions to redress depended on both the offer difference (which asset was offered most plentifully?) and the goal difference (how much pressure was there to restore asset imbalance?). Hence, in our first neural analysis (GLM1) we asked whether these quantities are encoded in blood-oxygenation-level dependent (BOLD) signals at the time of choice, i.e. time-locked to offer onset. We observed a dissociation whereby more dorsal regions of the ACC responded to offer difference and more rostral regions of the ACC responded to goal difference. Specifically, a correlate of goal difference (asset imbalance) was observed in rACC (BA24; peak: 10, 40, 14 [x,y,z], cluster-false-discovery-rate-q (FDRq) < 0.01; Fig. 2c and Table S2), accompanied by a cluster in the ventral occipital lobe (peak: -34, -48, -18, cluster FDRq < 0.005) and in some surrounding regions of dorsomedial PFC. By contrast, a correlate of offer difference was observed in dACC (BA32), extending into the overlying dorsomedial prefrontal cortex (dmPFC) (see Fig. 2b; peak: -6, 28, 38, global peak: -14, 28, 50, cluster - FDRq < 1 x 10^{-11}). Additional clusters were found in the bilateral insula (left peak: -34, 20, -2; right peak: 46, 20, 2; cluster – FDRqs < 0.02), bilateral inferior parietal lobule (iPL), and angular gyri (left peak: -46, -56, 30; right peak: 50, -56, 30, cluster – FDRqs < 0.005). We additionally included the signed offer difference in the regression, as in the behavioural analysis above (i.e. most– least needed asset) but did not observe significant clusters that encoded this quantity signed by needs.

One potential concern is that these effects are secondary to time on task. However, neither unsigned offer difference nor goal difference were strongly correlated with reaction time (RT; r = 0.046, r = 0.006, respectively), and we found the same neural effects when controlling for RT in GLM1 (Fig. S1), so we think it unlikely that this finding reflects the confounding influence of time-on-task. Rather, these data suggest that distinct regions of the ACC encode the two factors that determine whether a need should be addressed: the relative quality of the rewards on offer (dACC), and the relative balance of internal needs (rACC).

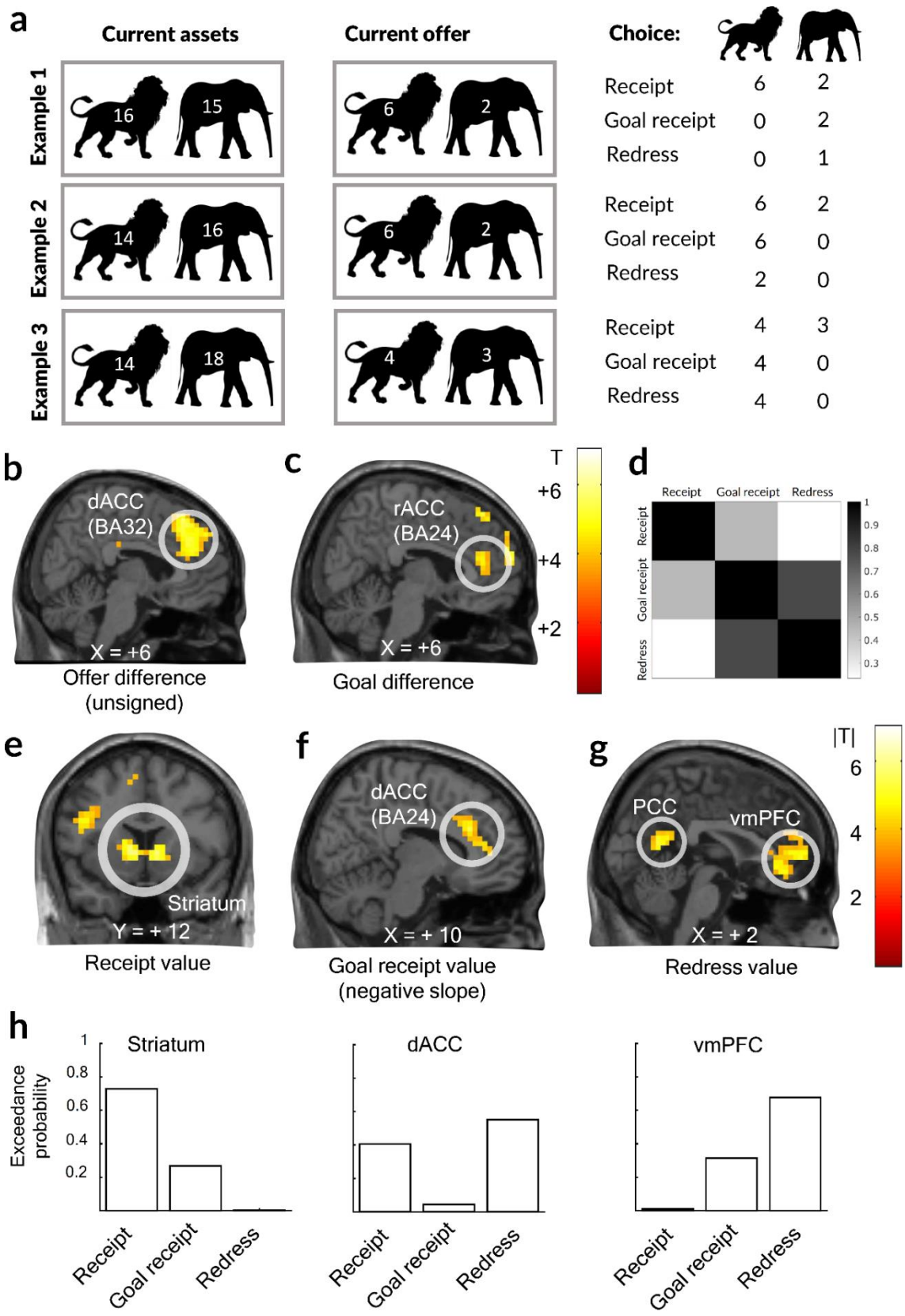


Figure 2. Representation of choice-relevant task features and value codes. (A) Illustration of the different value codes described in the paper for three example trials. The "current assets" panel depicts the state of assets before participants made a choice, whereas the

"current offer" depicts the available choices. The panel to the right shows the values of the three value codes used for analyses depending on choice: Receipt value reflected the increase in the overall number of animals achieved by choice. Goal receipt value codes the receipt value in the frame of reference of the goal. Finally, redress value encodes the realized redress between the goals or, equivalently, the increase in trial reward. It is computed as min (receipt value, goal difference) if participants chose according to their needs, or zero otherwise. (B) Positive correlation between offer difference (high minus low) and BOLD signals in the medial PFC, sagittal slice at x = 6. (C) Correlation matrix for the three value codes. (D) Positive correlation of goal difference with voxels on the same slice as in B). (E) Encoding of receipt value in voxels rendered on a coronal slice at y = 12. (F) Encoding of goal receipt value on a sagittal slice at x = 10 (G) Encoding of redress value on a sagittal slice at x = 2. Voxels were thresholded at $p < 0.001$, uncorrected. (H) Results of Bayesian model selection. Bar height corresponds to exceedance probability when testing which of the three value models fit best in a random effects analysis across subjects within each ROI (striatum, dACC, vmPFC). No error bars can be computed for exceedance probabilities.

vmPFC encodes a value redress signal. Previous research has reliably demonstrated that vmPFC encodes decision outcome but left open the question of how current needs may impact reward encoding. Our task allowed us to dissociate three possible frames of reference for the outcome signal that might be computed on each trial: the total number of animals received independent of need (receipt value), the number of animals received in the frame of reference of need (goal receipt value), and the level of redress among asset imbalance that was incurred by a choice (redress value). To illustrate, consider a participant with 10 lions and 12 elephants in their zoo who receives 3 more lions. The receipt value (and goal receipt value) would be 3, whereas the redress value would be 2. However, if the participant chose 2 more elephants, the receipt value would be 2, and the goal receipt value and redress value would be zero (see Fig. 2a for an illustration). By design, receipt value and redress value were only weakly correlated ($r = 0.23$, $R^2 < 0.06$; correlation matrix Fig. 2d) and thus dissociable; however, goal

receipt value and redress value were inevitably more correlated, because they were identical whenever the choice did not fully restore the asset imbalance ($r = 0.78$, $R^2 < 0.61$).

We first used a standard fMRI analysis approach, allowing these three value regressors compete for variance in the BOLD signal at the time of choice (GLM2). Receipt value positively activated clusters in the bilateral dorsal striatum (caudate nucleus), extending into ventral striatum (peak: -6, 12, 2; cluster – FDRq < 1×10^{-6}; Fig. 2e). Goal receipt value was negatively correlated with an area in the dACC (BA 24; peak: 14, 28, 26; cluster – FDRq < 0.05 Fig. 2f). This cluster lay between the dACC cluster that responded to offer difference and the rACC cluster that encoded goal difference, as if the ACC variously encoded the offer, the need, and the receipt in the frame of needs –three quantities required to compute the level of balance or imbalance among assets. Finally, redress value activated a cluster in the vmPFC (peak: 6, 52, 2, cluster – FDRq < 0.001; Fig. 2g) and the posterior cingulate cortex (peak: -6, -60, 14, cluster-FDRq < 0.005), two regions that are often found to coactivate in value-based decision tasks (Bartra et al., 2013) (see Tables S3-5 for full results). This association between redress value and vmPFC signals was also observed when we coded redress value relative to unchosen redress, in line with the previous observation that the vmPFC codes for the value of a chosen relative to an unchosen alternative (Boorman et al., 2009; Rushworth et al., 2012).

We were initially concerned about the reliability of this effect given the non-negligible correlation between receipt value and redress value. However, when we adopted the conservative approach of first orthogonalising this relative redress

value with respect to the other two predictors, ensuring that any effect of redress value must be over and above the other two value codes, we observed the same pattern of activations in dorsal striatum, dACC and vmPFC (Fig. S2, Tables S6-8). Thus, whilst mindful of the challenge of dissociating correlated regressors in BOLD analysis, we think the most probable explanation is that vmPFC encodes the redress among asset imbalance in our task.

One reasonable alternative hypothesis is that the vmPFC is simply encoding the monetary receipt on each trial. We note that redress value does not equal the overall monetary reward received on a given trial, but rather encodes the expected *increase* in reward conditional on the choice. Nevertheless, we used a non-circular ROI approach to test whether the vmPFC simply coded for the monetary reward observed on each trial by testing whether it encoded the level of the lower (trial reward-determining) asset at the time of choice, but found that this was not the case (t_{20} = 0.02, p > 0.98; GLM3). This leads us to conclude that vmPFC encodes the level of redress incurred by a choice, and not simply hedonic reward incurred on any given trial.

One interesting feature of these data is that they seem to suggest that vmPFC, dACC and caudate code for the value signal in distinct frames of reference: the caudate codes the level of receipt independent of needs; the dACC codes the level of receipt in the frame of reference of needs; and the vmPFC signals the degree to which a choice restores value equilibrium. To test this claim, we used a more stringent model fitting approach in which each predictor (parametric value signal) was entered alone into the GLM to estimate which provided the best fit to BOLD responses in each ROI. We compared models using a Bayesian model selection

method that relaxes the assumption that all participants are explained by a single model, allowing inference at the random effects level (Stephan et al., 2009). We compared models based on the exceedance probability, which estimates the likelihood that one model outperformed all other models in the set. It is thus a continuous measure indicating the confidence one may place in a model being better than the other ones that were tested.

For each model we computed the log evidence (as approximated by the model's log likelihood fit) in each ROI (caudate, dACC, vmPFC) selected from the contrasts previously described for GLM2 using a non-circular, leave-one-subject-out approach. The pattern we observed corroborates our previous finding that the caudate, dACC, and vmPFC exhibit different responses to these value codes, but added further nuance: Caudate was best fit by model 1 (containing receipt value; exceedance probability (ep) > 0.73), whereas vmPFC was best fit by model 3 (redress value; ep > 0.67). Within the dACC, however, no model showed a clear advantage over the others, although model 3 was numerically best (ep > 0.55; Fig. 2h). This analysis thus confirmed our contention from GLM2 that striatum and vmPFC were best described by two different value codes. The pattern of results in dACC, however, was less clear, indicating that it may encode a wide range of different values relevant for a decision (Kolling et al., 2016).

Together, thus, we find that the dorsal striatum and vmPFC fulfilled dissociable roles in value encoding in our task. The dorsal striatum encoded the number of animals received irrespective of needs (receipt value). In contrast, the vmPFC encoded the redress in imbalance incurred by a given choice, i.e. the extent to which internal needs were returned to a balanced state by a given choice. The

ACC, by contrast, seems to form a hub that brings together information about the offer, the needs and the change in asset levels incurred by a choice.

Computational model and optimal policy. The analyses above leave open the question of how an agent should behave in order to maximise reward in our task. One way to approach this question is to use dynamic programming (DP) to compute the policy that will maximise expected future return over each zoo. DP searches through possible future states and chooses a reward-maximising action according to the expected future return, assuming optimal subsequent choices. Because it accounts for possible future offers up to the horizon defined by the end of the zoo, the DP algorithm will sometimes choose the immediately less needed option in order to maximize overall return. For instance, where an offer of the asset not immediately required is generous, it may be sensible to nevertheless choose that asset, to guard against future scarcity when the offer probabilities reverse. This policy will be less useful at the end of the block, because such future benefits are curtailed when the zoo ends. Our use of a DP model does not imply a commitment to dynamic programming as a plausible algorithmic account of the computations that humans performed during the task. Rather, we adopted this approach to explore the normative policy for maximising reward.

Armed with this computational tool, we first computed the upper bound on reward shown in Fig. 1b, as reported above. Next, we asked whether humans, like the model, considered the potential time horizon when choosing to defer an immediate redress in asset imbalance. We found that they did: as for the DP model, proximity

to the end of a zoo (defined as 1/countdown) positively predicted redress responses (t_{20} = 3.74, p < 0.005) and that this tendency was stronger when the goal difference was greater (t_{20} = 3.13, p < 0.01). However, this was not driven by the overall length of the zoo alone, which failed to predict redress probability (t_{20} = 1.26, p > 0.22). Next, we evaluated choices made by the DP model using the same logistic regression approach. Despite some similarities in the betas obtained between humans and the DP model, the latter exhibited a stronger impact on *p(redress)* of both offer value and proximity to the end of the zoo (Fig. S3a), but predicted the general trend in participants' choices (Fig. S3b).

These findings might be explained if humans adopt a strategy that is approximately optimal but myopic i.e. has insufficient search depth when computing expected future return. We thus reimplemented the DP model, but allowed the planning horizon (in steps) to vary as a free parameter, as well as the subjective reversal probability, and policy terms that allow for bias and choice variability. When we fit the output of this "myopic" DP models to participants choices using maximum likelihood estimation with five-fold cross-validation (one fold per scanner run), we found that, on average, participants estimated the reversal probability to be approximately 0.44 (SD = 0.33) and planned only 7.5 trials ahead (SD = 6.04), significantly lower than the theoretical value of 20. The overall model log-likelihood was -2047.2, which was lower than both the logistic regression model with proximity (7 parameters; LL = -2111.1) and without proximity (4 parameters; LL = -2189.2), and the DP model was strongly favoured by Bayesian model selection (exceedance probability = 0.99).

This best-fitting (but myopic) DP model was also able to reproduce several qualitative features of the data. First, it captured how the average probability of redress varied with proximity to the end of the block (Fig. 3a). Second, it captured how the average need to redress (i.e. tolerance for disparity among the two assets in the zoo) varied as a function of time for blocks of different length, Fig. 3b. By contrast, a fully optimal DP model is more tolerant of larger goal difference values in the middle portion of the block, indicative of a longer time horizon for planning.

In conclusion, the DP analyses revealed that although humans planned for future needs, they were suboptimal in two ways: firstly, they did not plan ahead sufficiently, and secondly, they over-estimated the reversal probability.

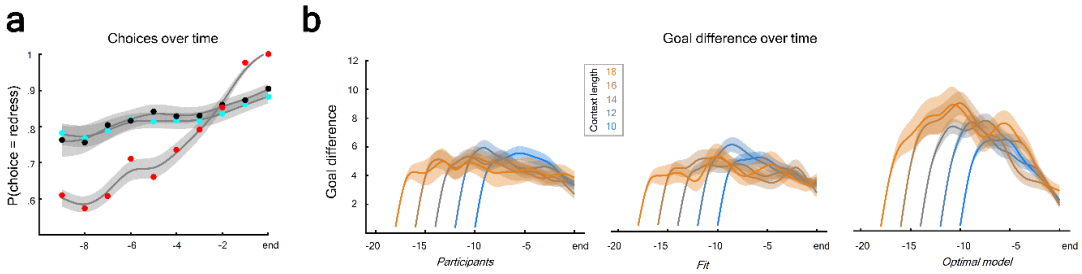


Figure 3. Comparison of optimal policy model and participants' behaviour. (A) Plot of redress choices against countdown toward the end of a block for optimal DP model (red dots), human participants (black dots), and best fitting "myopic" DP model (turquoise dots). Circles are mean and shaded areas are SEM around the spline-smoothed means (for display purposes only). (B) Plot of goal difference against time for blocks (zoos) of different length. The three subpanels show data averaged across participants (left) the simulated fit of the myopic model (middle), and the optimal model (right panel). Solid lines are mean across participants/models, shaded areas are SEM. Curves are spline-smoothed for clarity of viewing. Blocks with 20 trials were excluded from this analysis, as these represented less than 1% of all blocks.

Neural coding of time-varying pressure to redress. One algorithmic account that would be consistent with both this description of human performance and the neural data above is that the brain keeps track of the goal difference (pressure to redress, as a function of the two asset tallies) and codes pressure to redress in a way that varies with proximity to the end of the block. To test this view, we returned our attention to the neuroimaging data, and asked whether BOLD signals covaried with the interactions of the two asset tallies and proximity, i.e. whether cortical regions reflected a time-varying pressure to redress. In GLM3 we found that the rACC (BA 24) region covarying with goal difference (i.e. the quantity that participants need to keep in balance) also encoded the interaction of the higher asset and proximity (negatively; t_{20} = -2.16, $p < 0.05$), but not of the lower asset and proximity (t_{20} = 1.33, $p > 0.19$). Thus, encoding of the higher asset (and by extension goal difference) decreased over time in this region (Fig. 4b). We also observed that a more posterior region in the dmPFC (GLM1; peak: -36, 16, 50, cluster – FDRq < 0.001), as well as bilateral angular gyri (left peak: -50, -48, 34; right peak: 54, -44, 38; cluster – FDRqs < 0.001), and dorsolateral PFC (left peak: -34, 20, 46; right peak: 30, 4, 62; cluster – FDRqs < 0.001, Table S9) encoded the main effect of proximity to the block end, independent of asset difference, as if they were tracking the available time remaining to redress (Fig. 4a).

Finally, we asked whether single-trial activity (from GLM4) in the same rACC ROI (BA 24) had an impact on behaviour. We constructed a logistic regression model in which choices to redress were predicted by the two goal tallies, the rACC activity, and the interactions of rACC with the goal tallies. Choices to redress were negatively predicted by the lower tally (t_{20} = -4.75, $p < 0.0002$), and this effect was

accentuated when rACC signals were stronger (t_{20} = -2.11, p < 0.05), whereas choices to redress were predicted positively by the higher tally (t_{20} = 6.69, p < 1 x 10^{-5}), and this effect was also stronger when rACC was more active (t_{20} = 2.09, p < 0.05; Fig. 4c). No main effect of rACC activity was observed (t_{20} = -1.72, p > 0.10). Thus, high rACC activity guided choices based on the current need to redress.

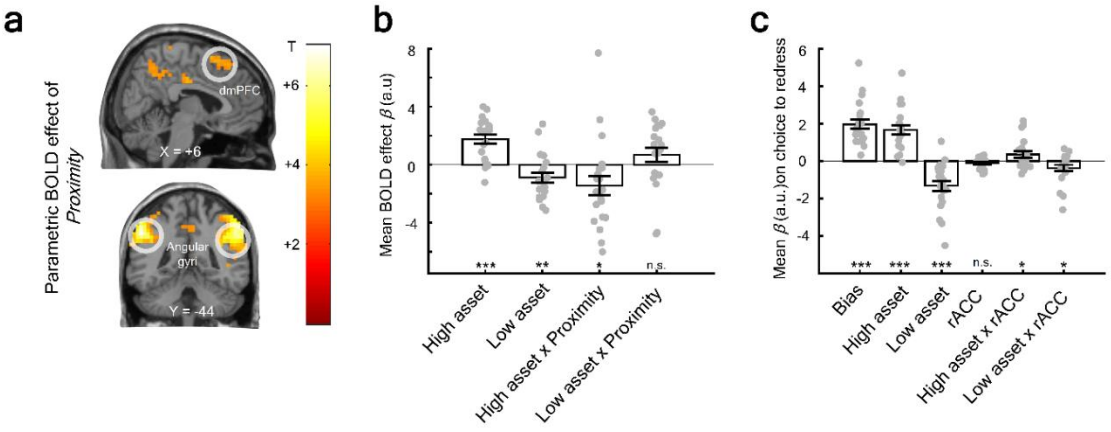


Figure 4. Neural effects of trial sequence. (A) Main effect of proximity on voxels rendered on a sagittal slice at x = 6 and a coronal slice at y = -44 from GLM1. Voxels were thresholded at p < 0.001, uncorrected. (B) Higher and lower asset encoding in leave-one-out ROI in rACC (BA 24; ROI defined from goal difference contrast in Fig. 2). Bar height indicates mean neural beta, error bars are SEM, grey circles represent individual betas. (C) rACC activity predicted choice to redress on a trial-by-trial basis. Bars are mean parameter estimates from logistic regression, error bars are SEM. Significance was assessed using a two-tailed t-test against zero. *** is p < 0.001, ** is p < 0.01, * is p < 0.05.

Discussion

We asked humans to perform a value-guided decision task requiring them to balance levels of two possible assets ("value equilibrium"). We found that they did so strategically, basing their choices about whether or not to redress or defer on the value of current offers, knowledge of value imbalance, and the time horizon available for future redress. In other words, humans make choices that actively

arbitrate among current goals, and plan for possible future scarcity. We argue that considering this alternative frame of reference for the neural coding of value sheds new light on the function of key brain areas involved in reward-guided decision-making.

Previous studies have identified human BOLD responses that code for the economic value of offers and outcomes in various frames of reference, observing for example that the vmPFC codes positively for the value of a chosen option and negatively for an unchosen option, with the reverse pattern observed in the dACC (Bartra et al., 2013; Boorman et al., 2009). Other, less well-established frames of reference for value signals have been reported, for example that value is coded relative to an average or default stimulus in either the vmPFC (Cox and Kable, 2014; Lopez-Persem et al., 2016) or dACC (Boorman et al., 2013), or that these regions respectively signal the value of a proximal choice, and the need to switch to a new context of known average value (Hayden et al., 2011; Kolling et al., 2014, 2012). However, the foundational premise that has underpinned these (and related) past studies is that an outcome can be indexed on a single value function that denotes the change in overall reward or wellbeing that is incurred by its receipt.

Our approach is rather different. We begin with the assumption that agents strive to maintain equilibrium among multiple different classes of asset, because depletion of any one asset can be detrimental to survival even if levels of other assets are high (O'Reilly et al., 2014). Thus, a sated and hydrated animal can still die of cold, or a human who has accumulated strong social capital can still fall into ill health through personal neglect. This intuition suggests a novel frame of reference for the

neural coding of value, in which the brain keeps track of the relative disequilibrium among goals (which indicates how pressing it is to maximise the minimum asset) and the degree to which any one choice redresses this imbalance (which is a proxy for how that choice promotes survival, at least over the short term). In support of this view, we find that BOLD signals in distinct brain regions vary with these quantities. For example, the rACC encodes the need to redress imbalance among assets, and the vmPFC signals the extent to which choices restore value equilibrium. In other words, in our task, values are coded with reference to internal needs: the relative levels of depletion or replenishment of a multiplicity of different assets.

Turning to the detail of our neural findings, the current work offers new insights into the function of three regions previously implicated in value-guided choice: the striatum, vmPFC and ACC. Previous studies have variously implicated these regions in computing the value of stimuli, actions and goals (Dolan and Dayan, 2013; Doll et al., 2012), with vmPFC and ACC typically coding inversely for value in an at least partially redundant fashion and the vmPFC and striatum seemingly coding for the value of proximal rewards and goals in a perplexingly overlapping fashion (Bartra et al., 2013). By contrast, our work suggests some unusually clear dissociations among the quantities coded by these regions. Firstly, the results we report dissociate value encoding in the striatum and vmPFC, with the former responding to outcomes independent of goals and the latter responding to the level of redress among goals incurred by a choice. Secondly, we observed a dissociation among ACC signals, with more dorsal regions signalling the quality of an offer and more rostral regions

indicating the need to restore imbalance among goals, with the latter signal modulated by the available time remaining to do so.

Dissociable value codes in vmPFC and striatum. On each choice, the striatum responded to the magnitude of the chosen offer (i.e. the number of animals received), but unlike cortical regions, it did so in a fashion that was insensitive to internal needs – the signal went unmodulated by estimated cumulative assets. Strikingly, these outcomes do not readily relate to reward received on a given trial, but simply indicate a change in one's overall internal state. Previous studies have suggested that where a single asset (e.g. money) is to be maximised, both vmPFC and striatum seem to code the long-run average value of an action, and the value computed via tree search through future states or outcomes (Bartra et al., 2013; Doll et al., 2015, 2012). These common value signals in cortical and subcortical regions may reflect the fact that animals use a mixture of model-free and model-based information when making decisions (Daw et al., 2011; Dolan and Dayan, 2013; Doll et al., 2012). However, our work instead suggests that cortical systems compute values in the frame of reference of ongoing internal needs, whereas the striatum codes for overall receipt independent of those needs. It should be noted, however, that the peak of the striatal cluster identified here was slightly more dorsal than commonly identified for hedonic value (Bartra et al., 2013), although it also included voxels in the ventral striatum.

Importantly, however, our task allowed us for the first time to dissociate outcomes *per se* from their impact on current goals. We found that unlike in previous tasks, vmPFC did not simply mirror the code employed by the striatum, but rather coded for the redress in goal difference – that is, the extent to which any given choice

restores the imbalance among internal needs. This signal remained the best explanation of vmPFC signals even when we partialled out the actual outcome received in number of animals, or the trialwise monetary reward that motivated participants to perform the task. We note that in our task, as in a natural environment where wellbeing is determined by the minimum of multiple possible assets, redress value is inevitably identical with the change in reward or hedonic value. This presumably explains the ubiquitous observation that vmPFC correlates positively with reward outcome when a single asset is being optimised, and why this signal is often observed to be modulated by the local context provided by average value (Cox and Kable, 2014; Padoa-Schioppa, 2009; Yamada et al., 2018). Further research is needed to more fully distinguish our account of the vmPFC from other models that emphasise encoding of trial-wise change in reward (measured on a single or multiple attributes). However, our task may provide new insights into *how* the vmPFC may compute hedonic value. Choice outcomes (i.e. the number of animals received) had to be related to the goal difference to compute value. Thus, our findings differ on two accounts: First, vmPFC did not simply reflect the value of a trial, but rather the *between-trial* increase in value. Second, it could only do so by utilizing the goal difference encoded in the rACC. We note, however, that our design did not allow us to differentiate vmPFC activity immediately preceding a choice from activity post-choice as participants were allowed to respond at any time after stimulus onset.

These findings build on our previous work implicating the human vmPFC in coding an internal, unsignalled representation of cumulative assets in a way that maximises a specific goal (maintaining net positive aggregate reward), over and above any

momentary hedonic value that arises from choices (Juechems et al., 2017). However, the current work additionally implies *why* the brain keeps track of cumulative assets – because this allows any imbalance among internal needs to be redressed by future choices. This helps understand a perplexing paradox in the decision literature – why is it that neural signals in the vmPFC code ubiquitously for rewards, but lesions to the vmPFC incur only subtle deficits where decisions involve maximisation of a single asset (Noonan et al., 2010)? The vmPFC outcome signals ubiquitously observed in fMRI studies may reflect the tracking of redress value to one of multiple potential goals – obtaining food, or money, or maximising accuracy – rather than computing the relative strength of one offer over another. Indeed, in studies of navigation the vmPFC signal also ramps up over multiple steps that are made towards a destination, presumably because each step redresses the distance between current and goal state (Balaguer et al., 2016; Howard et al., 2014). Thus, we would expect patients with vmPFC lesions to arbitrate effectively between offers of differing value but to fail to allocate decisions strategically over the long term, precisely the pattern that is observed in both humans and monkeys (Gläscher et al., 2012; Noonan et al., 2010). Furthermore, patients with lesions in the medial PFC tend to search for information within individual options, failing to integrate over options and attributes (Fellows, 2006).

A**CC employs multiple frames of reference to arbitrate between goals.** We also observed that two distinct areas of the ACC encoded the current offer (dACC) and internal state (rACC). Firstly, a cluster in the rACC (BA24) responded to the current imbalance among assets, i.e. the extent to which one asset needed to be replenished in order to match another. This cluster did not simply reflect how this

imbalance evolved up to the current trial, but exerted influence on participants' choices indicating that participants actively sought to maintain equilibrium between assets. Previous studies of reward-guided decisions have emphasised that the dACC codes positively for the relative value of an "unchosen" option– i.e. that which was foregone when a choice was made (Boorman et al., 2009; Daw et al., 2006) – or that it signals the need to switch away from a current context towards a new, potentially richer, source of rewards (Kolling et al., 2016, 2012). Secondly, a cluster in the dACC reflected a generic disequilibrium in the offers irrespective of needs. Interestingly, a cluster that lay in the intermediate region between the dACC and rACC seemingly employed multiple frames of reference for value encoding. It has been recognised previously that the ACC employs multiple frames of reference and encodes variables on several timescales (Kolling et al., 2018, 2016; Meder et al., 2017). Our framework provides an intriguing explanation about *why* this may be the case: The ACC integrates offers with goals to inform the best course of action in a given context. In order to do so, it must reflect i) the overall magnitude of offers, ii) how a choice will change the cumulative assets, and iii) how several assets relate to one another. Its role in updating all relevant cumulative assets explains why dACC is often found to encode an unchosen (i.e. a counterfactual) value (Boorman et al., 2013), while its role in relating assets to one another fits with the finding that dACC encodes the value of switching away from a currently active context (Kolling et al., 2012). Together, our findings argue for a critical role for the ACC in computing which of multiple possible goals should be pursued at any one time.

The choice of whether to satisfy an immediate need, or to build up less pressing resources to offset future scarcity, is a ubiquitous problem for humans and other animals (Hurly, 1992; Mullainathan and Shafir, 2014). Critically, this choice depends on the time horizon over which choices can be made: an optimal agent with knowledge of the time horizon will tolerate an asset imbalance early in the block, where it can be later redressed if opportunities change, but respond to pressure to redress later in the block. Human participants were less tolerant of imbalance than the optimal model, which can be explained if they are somewhat myopic in their planning (Frydman et al., 2014; Kolling et al., 2018). Note, however, that this is akin to, and indistinguishable from a parameter discounting future rewards in favour of short-term gain. Nevertheless, reversal probability, planning horizon and an alternative discount parameter would all predict that participants chose to redress more than was optimal, especially at the early stages of the block. Consistent with this view, the rACC region that coded for goal difference did so in a fashion that varied with proximity to the end of the block. However, the direction of this effect was somewhat surprising to us, in that goal difference encoding decreased with time. It is possible that participants had a bias to redress towards the end irrespective of the size of the goal difference – thus, goal difference encoding may not be needed to drive choice towards the final few trials in a block. Nevertheless, this finding requires further corroboration in future studies.

We acknowledge one important limitation of our approach: working in the restricted environment of the laboratory, where varying the true internal needs of the participant is technically challenging if not impossible, we instead adopted a task in which participant payment depended overtly on asset balancing. Thus,

whilst we would like to argue that animals intrinsically seek to maintain value equilibrium in the natural world, in our task they were required to do so in order to maximise a single asset (money). Given that we were unable to vary the need for primary reinforcement, our task is thus perhaps a closer simulacrum of the uniquely human task of maintaining equilibrium among more complex needs, such as when a busy academic balances the goals of conducting impactful research and providing effective teaching. We also acknowledge that an alternative computational framework exists for understanding how animals might solve a task such as ours, namely that the brain searches forward through a tree of possible future states in order to calculate which course of action will reap the greatest long-term reward – indeed, this is the normative framework that we used as a benchmark for comparing human performance. Although our findings do not explicitly rule out this possibility, it is not clear to us why, if humans were not tracking equilibrium among goals, the brain would code for goal disequilibrium (rACC) and redress (vmPFC).

Our findings may more generally have implications for understanding health and wellbeing in humans. We assume that wellbeing is related to the minimum among current needs, and find evidence that neural circuits for valuation and choice have evolved to maintain a balance among needs via goal-directed choice. Disorders of valuation and choice, such as depression, might be associated with failures of a system that attempts to maintain value equilibrium such as an exaggerated representation of goal difference, or a failure to update internal asset estimates when they are redressed.

Methods

Participants

Twenty-two healthy volunteers with no history of psychiatric or neurological disorder participated in the study. One participant was excluded from analyses due to scanning equipment failure, leaving N = 21 (5 male, mean age = 23.6, SD = 2.77). All participants gave written informed consent and the study received approval from the ethics board of the University of Granada. Participants were compensated for their time with €25 plus a performance-based bonus (mean = €7.15, SD = €1.42, range: €4.52 to €11).

Task details

Participants performed a task that involved managing a virtual zoo which housed two types of animals: lions and elephants. Each zoo began empty and animals were offered over a block lasting a variable number of trials (10-20), before a new zoo began. Participants encountered 5 zoos on each of 5 scanner runs, with each run lasting a total of 65 trials. On each trial, participants were first shown a fixation cross (150 ms), followed by a "choice" screen (2.5 seconds) in which they were offered a variable number of lions or elephants. This was indicated by an arabic number above a drawing of the animal on the left and right of the screen. There were always a high (4-6) offer of one animal and a low (1-3) offer of the other. The assignment of high or low offers to each animal was autocorrelated over trials, reversing with a probability of 0.3, which led to "streaks" in which one animal could be harvested in more abundance. Participants indicated their choice by pressing a button with their left or right index finger (for options on the left and right,

respectively). The side on which each animal was offered varied randomly across trials. Once a choice was made, the unchosen option disappeared from the screen, leaving only the chosen option ("response" screen), and this lasted the remainder of the trial (2.5s minus reaction time), followed by a blank screen for a uniformly jittered interval (1.5-4.5 seconds). Subsequently, a feedback screen (1 second) indicated the number of animals of each species in the zoo (the "assets") as a number above the respective drawing. Trial-wise reward was indicated below the drawings and was directly proportional to the lower asset (e.g. it was 5 if there were 10 elephants and 5 lions in the zoo). Inter-trial intervals were drawn uniformly between 2s and 6s (1-3 time of repetition; TR).

Throughout the run, five coloured discs just below the top of the screen were shown, representing in the 5 zoos ordered from left to right in time. Each zoo had a colour, and the colour of the current zoo was indicated by a central circle that appeared on the "choice" and "response" screens. Colours were drawn randomly on each run and were only included to aide participants in grouping trials into discrete blocks. During the "choice" and "response" phases of the trial, a central number was displayed within the circle, also in a colour matching the colour of the current zoo. This number counted down the trials still remaining in the current zoo (including the current trial). At the conclusion of a block, its cumulative earnings were displayed on a "bonus" screen and this number was shown underneath the corresponding coloured disc on all subsequent trials. A lottery at the end of each run displayed to participants which of these bonuses was chosen as extra payment. Each block had equal probability of being selected in the lottery.

Training and instruction. Participants completed one run immediately before scanning to familiarize themselves with the task. They were told that if lions were currently the high offer, they were more likely to be the high offer on the following trial, but that this was not certain and this contingency could reverse unpredictably. However, they were not told the exact reversal probability (0.3). They were told that on some occasions it was more beneficial to defer choosing according to their current needs as this could yield better long-term payoff. Importantly, participants were not given specific instructions to maintain the goal difference at a certain level (e.g. as close to zero as possible). They were also told that they would receive the sum of the lottery values across runs after the end of the experiment.

Analysis of behavior

All analyses were carried out using MATLAB 8.6 (R2015b) and custom in-house code. Choices were coded as "redress" when participants chose the offer corresponding to the currently lower asset in their zoo and "defer" if they chose the other option. All behavioural results discussed in the main text are based on this dichotomy (but see SI for behavioural analyses that classed choices as "greedy" with respect to the offers on the screen; Fig. S4a). Proximity was coded as 1/countdown.

Logistic regression

The dependent measure for the logistic regression was the binary variable 1=redress, 0=defer. We fit the models to individual subjects using z-scored regressors in five-fold cross-validation (one for each scanner run). We then averaged the betas across runs for each subject and used a two-tailed one-sample

t-test to determine whether the group mean of each parameter differed significantly from zero. We excluded trials where the goal difference was zero from analysis (n = 1071 out of N = 7150 trials, of which almost half were first trials in a new block), because participants' choice would be trivial (choose the higher offer) and the signed offer difference would be undefined in these cases.

The first logistic regression model contained the following regressors: a constant term, signed offer difference (need option minus not-needed option), absolute asset difference, and overall length of block. The second logistic regression model added proximity and the interaction of goal difference and proximity.

An additional model testing for the influence of rACC activity on behaviour differed from these models by splitting the asset difference into high and low asset which allowed us to test interactions of these terms with rACC activity.

Model fits were compared using five-fold cross-fitting via Matlab's fitglm function, summing log-likelihoods across participants and trials. In addition, we performed Bayesian model selection to compute the exceedance probabilities for each model, which measures whether a given model was more likely to be the true model than all competitor models (Stephan et al., 2009).

Dynamic Programming model

We developed a Markov model to derive the optimal solution for the task. The model's state space was given by the goal difference, the number of trials left in the block, the two offers in the frame of reference of the lower asset, and the reversal probability (constant for all models, except when fitting to participants). The model was optimized using the following dynamic programming formula

$$Q(s,a) = R(s,a) + \sum_{s\prime}^{S} P(s'|s,a) \times \max_{a\prime} Q(s',a')$$

Where *s* and *s'* are the current and successor state, *a* and *a'* are the actions taken in each state, Q(s,a) is the value of taking action *a* in state *s*, and P(s'|s,a) is the probability of transitioning from state *s* to state *s'* by choosing action *a*, and R(s,a) is the reward obtained in the current state. Thus, the model chose the action which maximized the sum of immediate reward (R(s,a)) and the expected value of the best action in all subsequent states. As the task had a finite horizon, we do not include a discount factor. We fit the model with different reversal probabilities (from 0 in increments of 0.05 to 1, where 0.3 was the true value) in order to compare it to participants' behaviour. As the model computed the expected value for both choosing to redress and choosing to defer, we computed the difference between the two and passed it through a sigmoid to allow for noisy decisions. The output of this sigmoid – p(redress) – was then fit to participant's behaviour with a weighting parameter and intercept using maximum likelihood estimation. Maximum possible reward was calculated by permuting across all possible combinations of choices in each sequence and finding the highest reward score.

FMRI Data Acquisition and pre-processing

MRI data were acquired on a 3T Siemens scanner. T1 weighted structural images were recorded directly prior to the task using an MPRAGE sequence: 1x1x1mm^3 voxel resolution, 176x256x256 grid, TR = 1900ms, TE = 2.52ms, TI = 900ms. Each fMRI image contained 32 axial echo-planar images (EPI) acquired in descending sequence with a voxel resolution of 3.5 mm isotropic, slice spacing of 4.2mm, TR = 2000ms, flip angle = 80, and TE of 30ms. 1850 EPI images were recorded per

participant, resulting in a scanning time of around 62 minutes. Data were pre-processed using SPM12 (Wellcome Trust, London). As EPI acquisition used a descending sequence, images were corrected for slice time acquisition with the middle slice (at TR/2 = 1 s) as reference to minimize interpolation errors (Sladky et al., 2011). Scans were realigned to the first scan within each session. The anatomical scan was co-registered to the mean functional image. Anatomical scans were normalized to the standard MNI152 template brain. The functional EPI images were then normalized, resliced to 4x4x4mm resolution, and smoothed with a full width half maximum Gaussian kernel of 8mm.

FMRI data analysis

Data were analysed using SPM12 and custom scripts. Data from the five scanning sessions were concatenated and constants identifying each run were added manually to the GLM in order to account for potential differences in mean activation and scanner drift. Stimulus onsets were incremented by 1s to account for slice time correction to the middle slice, which occurred at TR/2 = 1s after stimulus onset (Sladky et al., 2011). All results reported came from GLMs in which the canonical haemodynamic response function was convolved with a delta function (i.e. with zero seconds duration) time-locked to event onset (e.g. offer or feedback screen). The six vectors of motion parameters derived from pre-processing were included as nuisance regressors. Automatic orthogonalization was disabled. Group-level contrasts were constructed as simple t-contrasts using subject-level contrast images as input. Throughout this paper, we report only clusters that survived false-discovery-rate-correction (FDR) for multiple

comparisons (Chumbley and Friston, 2009; Genovese et al., 2002) unless otherwise noted. This avoids the potentially inflated false positive rates at the cluster level (Eklund et al., 2016). The FDR was calculated using a minimum cluster extent of 10 voxels and cluster-defining threshold of $p < 0.001$, uncorrected. These settings have been found to perform well in most cases (Eklund et al., 2018). Data were visualized using the xjview toolbox for SPM (http://www.alivelearn.net/xjview).

Our analyses are based on 4 GLM models. All 4 GLMs included two predictors of interest (onset of the choice screen, onset of the feedback screen) and two nuisance predictors (trials on which the goal difference was zero, and all no-response trials). GLM1 additionally included the following parametric regressors time-locked to the offer (i.e. at the time of choice): congruency (whether the high offer was of the currently needed type), response hand (left vs. right), signed and unsigned offer difference, sum of the offers, proximity, and goal difference. For completeness, we also included the following parametric modulators on trials with zero goal difference: response hand, unsigned offer difference, offer sum, proximity. Finally, goal difference (post-choice) was included as parametric modulator time-locked to the feedback screen. GLM2 included parametric modulators for the receipt value, goal receipt value, and redress value, as well as proximity, all time-locked to the onset of the offer. For completeness, we also included chosen value and proximity on trials with zero goal difference, and goal difference and cumulative value at feedback. GLM3 included parametric modulators corresponding to the two zoo tallies (up to that trial) their interactions

with proximity, and their increases, as well as proximity. These modulators were time-locked to the onset of the choice screen. These regressors were z-scored to allow comparisons between their beta values. For completeness, we also included chosen and foregone receipt value on trials with zero goal difference, as well as goal difference and total sum of assets at feedback. GLM4 modelled each trial and feedback screen as a single event without any parametric modulation.

Bayesian model selection. We tested three variations of GLM2 which only included one of the three value codes – receipt value, goal receipt value, redress value. All other settings were identical to GLM2. We estimated these models in the striatum, dACC, and vmPFC ROIs based on the contrasts reported for GLM2 extracted using a leave-one-subject-out approach and Matlab's fitglm function to obtain the log likelihood of the model. We then summed the log likelihoods across voxels as an approximation to the model's log evidence. We used a standard Bayesian Model Selection approach in SPM described in Stephan et al. (2009). This approach uses a random effects Variational Bayes algorithm which iteratively estimates the probability of a model given the data. We can then estimate the exceedance probability – a measure of the probability that a given model outperformed all others in the comparison. Unlike other model selection approaches, this does not assume that all subjects were generated (best fit) by the same model. We repeated this procedure for all three regions.

Region of interest extraction. ROI were extracted using a leave-one-subject-out procedure to avoid potentially circular inference. A mask was created for each

participant using the first-level contrast images from all other participants (i.e. leaving out the current participant). ROI were then constructed by taking all significant voxels ($p < 0.001$, uncorrected) within the relevant region (approximately given by xjview's database, e.g. within medial PFC). We extracted one rACC ROI from the "goal difference" contrast from GLM1. In addition, we extracted one vmPFC ROI from the "redress value" contrast, a striatal ROI from "receipt value", and a dACC ROI from "goal receipt value", all from GLM2.

Chapter Five: The representation of irrelevant item attributes across the prefrontal cortex

Abstract

The utility of an item can change drastically depending on one's goals. Often the new value is determined by a different set of item attributes which the decision-maker averages across. When attributes are saliently signalled, researchers often find a slight positive influence of irrelevant attributes as if the decision-maker preferred items with adequate value across possible goals and contexts. Here, however, we describe a new behavioural effect, where irrelevant attributes negatively impact choices, as if participants preferred items with attributes of unequal magnitude. This effect may reflect an integration of opportunity costs into decisions and was captured well by a normalization model operating across attributes. In preliminary analyses of the neural data, we found the dorsal anterior cingulate cortex and the frontopolar cortex to jointly represent the irrelevant and relevant attributes and may support reasoning about opportunity costs across contexts.

Introduction

Humans and animals choose between options based on their utility (Von Neumann and Morgenstern, 1944). However, how this utility is computed based on continually changing goals and obligations remains an open question. Researchers have begun to study this using tasks in which salient item attributes either have to be considered in isolation (Flesch et al., 2018; Mante et al., 2013), or where different attributes need to be combined in order to judge the utility of an item (Chau et al., 2014; Davis et al., 2012; Pelletier and Fellows, 2018), usually finding that currently irrelevant attributes exert a significant positive, albeit small, influence on behaviour. Often, however, especially in social and economic transactions, an item carries attributes that are neither overtly signalled, nor a tractable combination of features. For instance, consider a dinner party host selecting the wine to be served. Even though the available wines represent a fixed set of options, the host may first consider which wine their party would appreciate most, but in a second step may consider whether the same wine may be better served at a future party with different guests. Thus, the host will first compute the utility of the wines given their guests, but also engage in "counterfactual" reasoning to compute the utility of the same item in future contexts. Thus, if future guests are more likely to value a particular wine, it is perhaps best saved for this occasion. While such counterfactual reasoning is commonplace in our daily lives, we know relatively little about the impact of individual item attributes on behaviour or the neural underpinnings of this phenomenon.

Previous research on counterfactual reasoning has mainly focussed on unchosen options, or "opportunity costs", where the agent encodes the second-

best option. The value of an unchosen option as been found to reliably engage the dorsal anterior cingulate cortex (dACC) and the lateral frontopolar cortex (Boorman et al., 2011, 2009; Hayden et al., 2011, 2009; Kolling et al., 2012; Rushworth and Behrens, 2008). Furthermore, monitoring which of several potential alternative courses of action should be undertaken instead has been proposed as a central function of the anterior prefrontal cortex (Badre, 2008; Buckley et al., 2009; Mansouri et al., 2017), where representations of such alternatives may follow a dorsal-posterior to anterior-ventral gradient along the PFC, with more anterior areas representing events and associations more distal in time (Donoso et al., 2014; Koechlin, 2016; Koechlin and Hyafil, 2007; Koechlin and Summerfield, 2007).

One particularly relevant recent study showed that in a task where only two out of a set of three options were available for choice, the dACC encoded the value of the currently unavailable, counterfactual, option and was causally related to translation of this value into choice on later trials when the option became available. It did so in conjunction with the hippocampus, which underpinned the retrieval of the counterfactual value from memory (Shadlen and Shohamy, 2016). However, previous studies, including the one just discussed, tend to focus on how different attributes are combined into a context-dependent utility judgment, or can be compared with other, potentially unavailable or inferior items on this uni-dimensional scale (Chau et al., 2014; Fouragnan et al., 2019; Li et al., 2018; Polanía et al., 2019)

Here, we took a different approach, where we trained participants on a two-dimensional value space, where each dimension represented the value of an item to two potential buyers, thus emulating commonplace economic transactions.

Importantly, utility was unrelated to stimulus features and which item attribute was currently relevant depended on the context provided by different markets. First, we describe a novel behavioural effect whereby the irrelevant attributes exert a negative influence on choices, as if participants engaged in reasoning about opportunity costs. In a preliminary analysis of the neuroimaging data, we asked whether the same brain areas implicated in counterfactual value (dACC, frontopolar cortex, hippocampus) would also signal the value of the irrelevant attribute, finding this to be the case. This is remarkable especially because a given attribute stayed irrelevant for a block of several trials, significantly reducing the potential utility of a counterfactual value representation. However, representing the full two-dimensional value space may enable reasoning about opportunity costs, interfering with choice in the current context, but leading to better allocation of goods overall. Thus, dACC and frontopolar cortex may underpin flexible decision-making based on memory by maintaining irrelevant information over long timescales.

Results

Human volunteers (n = 26) performed a virtual "car salesman" task in which they had to learn offers of two potential buyers (María and Pablo) for a set of 16 antique car models. The value of the cars was arranged so that it lay on a 2D grid (Fig. 1a). Arrangement of specific items (cars pictures) on this grid was randomized across participants. The buyers' preferences were orthogonal to each other such that, on average, knowing the offer of one buyer carried no information about the other buyer's offer for the same type of car. Participants first learned these values outside the scanner where they viewed a car and indicated who they wished to sell to.

Importantly, participants received counterfactual feedback such that they were able to learn the offers of both buyers regardless of their choice. Whilst in the fMRI scanner, participants first performed another block similar to the learning phase (the "refresher block"), but this time without feedback and with the incentive to sell to the highest bidder. Subsequently, on each of six runs, participants were placed in blocks of fifteen trials of three different markets: One, where only María was available to buy cars, another one where only Pablo was available (the "single markets"), and a third market in which they did not know who would buy the car, thus incentivizing them to choose based on the average or maximum value (the "mixed market"). On each trial, participants saw two cars and had to indicate via button press which car they wished to sell given the current market (indicated on the screen throughout the trial). Thus, participants had to apply new sets of rules to the values they had just learned.

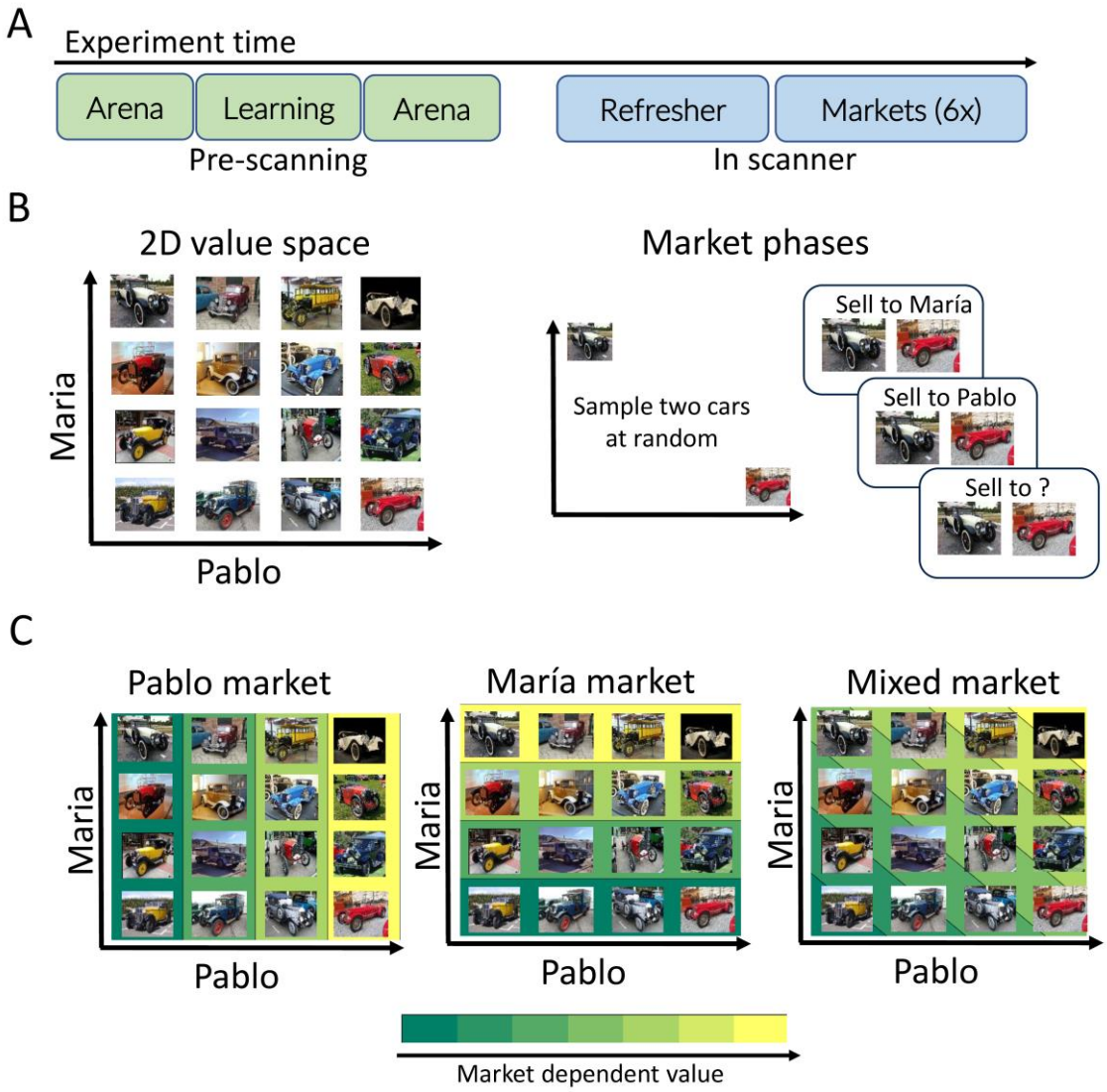


Figure 1. Value space and task design. A. Phases of the experiment, see Methods for details. B, Cars were arranged on a 2D grid across two potential buyers and participants learned the offer values associated to each car. In the scanner, two cars were sampled at random without replacement and participants indicated their choice based on the current market. For instance, in the María market, participants should choose the car on the left, whereas they should choose the car on the right in a Pablo market and be indifferent between the two in the mixed market, where choices should be based on the average value. C. Distribution of value across the three different markets.

Human choices are driven both by relevant and irrelevant value. Participants performed well on average, attaining 82.42% accuracy (SD = 8.2%; reaction time (RT) = 1.42s) in the single markets, as well as 95.2% (SD = 5.2%, RT = 1.27s) in the

refresher block, indicating that participants had accurately learned the value for each individual car. We employed logistic regression to test whether the currently irrelevant value (e.g. Maria's offers in a Pablo market) exerted any influence on participant's choices. For this purpose, we fit a model predicting whether participants made a correct choice. First, the absolute value difference for the relevant dimension had a clear influence on participants' choices (t_{25} = 2.92, p < 0.008), whereas the irrelevant value did not (t_{25} = 1.41, p < 0.172). Strikingly, however, the interaction of the two terms, had a robust *negative* influence on choice (t_{25} = -7.04, p <$3x10^{-7}$). This meant that on trials with high value differences for both buyers, participants were less likely to make a correct choice (Fig. 2a). To further probe this relationship, we analysed choices in a left-versus-right frame of reference. In this analysis, we excluded all trials where the relevant value difference equalled zero (20.26% of trials). In the previous regression, these choices were accounted for through the intercept as any choice on these trials was counted as *correct*. Again, we found an influence of the relevant value difference (t_{25} = 11.61, p < $2x10^{-11}$). However, there was now a negative influence of the irrelevant value difference (t_{25} = -2.61, p < 0.015), and no interaction (t_{25} = -0.15, p < 0.883). This relationship is illustrated in Fig. 2c. This indicated that participants chose the car on the left side less often if it was valuable to both buyers; as if they preferred cars where the two buyers disagreed about the appropriate price.

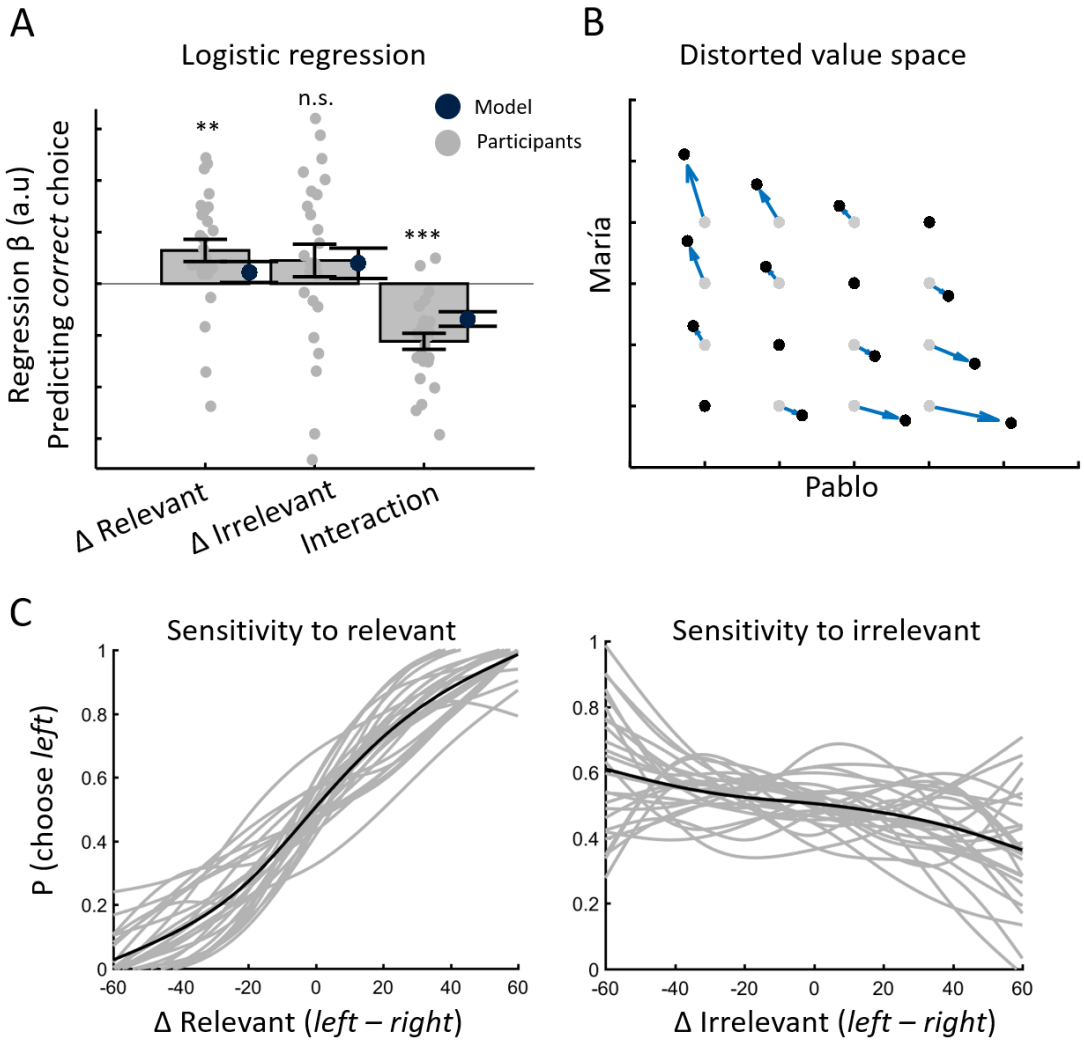


Figure 2. Logistic regression of choice and distortion of value space after normalization. A. The logistic regression predicted the likelihood with which participants made a correct choice (i.e. choose the higher-valued car in the current market), based on three regressors: the absolute value difference on the relevant dimension, the absolute value difference on the irrelevant dimension, and the interaction between the two. A simple, one-parameter, normalization model fit to choices reproduced a similar qualitative pattern. B. The original value space (grey dots) was significantly distorted after normalization was applied (black dots), with the largest amount of distortion at items where the two buyers differed the most in their offers. Regression coefficients were tested for significance using a two-sided, one-sample t-test against zero, where ** indicates $p < 0.01$, *** is $p < 0.001$, and n.s. indicates a non-significant test result. C. Average responses at the different levels of value difference (left minus right). Individual participants in grey, group average in black. Curves were spline-smoothed for visualization.

Next, we tested whether models based on the well-described neural mechanisms of divisive normalization could explain these effects. Normalization models predict that any inputs within a set are considered relative to each other, rather than on their absolute value. Such coding has been used to describe non-standard receptive fields in the visual cortex (Heeger, 1992) as well as, more recently, some surprising choice reversals in economic tasks (Bavard et al., 2018; Klein et al., 2017; Louie et al., 2013, 2011). Broadly speaking, normalization models decorrelate inputs so as to increase contrast. Hence, they should be able to capture the behavioural effects well. We employed two normalization models, one based on the classic description for neural firing rates (Louie et al., 2013) with three parameters, and another simplified version with just one parameter where options were boosted (or penalized) according to the relative difference between the two offers (see Methods). For example, a car that Pablo likes most, but María likes least would be maximally affected by normalization. We found that the simplified model performed best in a 6-fold cross-validation analysis (one fold per scanner run): The simpler model had a lower summed log-likelihood at the group-level (LL = -2038.3 versus LL = -2050.4) and clearly outperformed the larger model in a Bayesian model selection analysis (exceedance $p > 0.999$) , where the models are compared using a random effects approach (Stephan et al., 2009). Importantly, these models explained choices better than a constant model without any normalization where the relevant parameters were set to zero with only the slope parameter allowed to vary (LL = -2186.3). The distortion of the value space arising from the winning normalization model is plotted in Fig. 2b.

In our task, normalization was suboptimal because the two value dimensions were orthogonal. Thus, a simple strategy whereby the agent only learned the critical input variable to the normalization models, the signed difference in offers for a given car (e.g. Pablo minus María offer), performed relatively poorly (73.58% in the single markets), albeit better than chance. Importantly, however, this strategy yielded significantly lower performance than that of participants (73.58% versus 82.42%, two-sample t-test, $t_{50} = 4.92$, $p < 1 \times 10^{-5}$), indicating that participants were only partly driven by the relative difference.

The univariate signal does not reflect the irrelevant value dimension. Next, we turned our attention to the BOLD data acquired while participants made choices in the market phase. We began by asking whether we could reliably find expected areas in the medial prefrontal cortex encoding the value of the chosen option minus the unchosen option (i.e. the value difference) of the relevant dimension (e.g. Pablo's offers in a Pablo market). As anticipated, in GLM1 we were able to identify regions in the ventromedial prefrontal cortex (vmPFC) encoding this regressor positively as well as in the dorsomedial PFC exhibiting negative encoding, as has been described previously (Bartra et al., 2013). Next, we asked whether this encoding was influenced by the currently irrelevant value dimension (e.g. Maria's offers in a Pablo market), but failed to find any significant clusters encoding this regressor, even at liberal thresholds. We found the same to hold for the interaction of the relevant and irrelevant value differences. However, such absence of a univariate effect may be explained by a number of factors: First, the behavioural effect of the irrelevant dimension is small compared to the relevant dimension and

may thus only explain a small portion of the BOLD signal with the relevant value explaining the bulk of the variance. Second, and most importantly, it is not necessarily surprising that the univariate signal fails to capture an essentially two-dimensional value space. Thus, next we turned to representational similarity analysis (RSA), a technique which compares multivariate patterns of activity across conditions. This allowed us to ask whether the observed distances in the neural signal between conditions correlated to those predicted by a range of models. For example, we may ask whether the multivariate signal reflects the relevant or irrelevant value dimensions, see Fig. 4 for the representational dissimilarity matrices (RDMs) resulting from the models.

Medial prefrontal cortex represents the difference between offers when this is relevant. First, we employed RSA to investigate whether the signal in the medial PFC was sufficient to derive meaningful conclusions from multivariate analysis and to derive functionally defined regions of interest (ROI) for the subsequent analyses. In order to do so, we drew on the independent "refresher" dataset in which participants saw a single car on the screen and had to indicate who they wished to sell to, with the instruction that the value incurred would count towards their bonus. Since the task was to select the highest bidder for a given car, we asked whether the medial PFC would encode the absolute difference in the offers, essentially a measure of how easy a decision was to the participant (assuming sufficient learning). We found this to be the case for a large cluster centred around the dorsomedial PFC, with a peak in the dACC (8, 46, 28; x, y, z) and . Other significant clusters included the bilateral angular gyri and parietal lobes (right peak:

29, -63, 52, left peak: -34,-66, 35). Next, we visualized the resulting neural RDM from the dACC region (Fig. 3a) and applied multi-dimensional scaling (MDS) to it. MDS attempts to project distances from high-dimensional spaces (e.g a neural RDM from dozens of voxels) onto a smaller set of dimensions while keeping the resulting distances as similar as possible to the original high-dimensional distances. For this particular projection, we chose a three-dimensional projection (Fig. 3b). It was immediately apparent that the four options where both buyers offered the same amount (in green) clustered together and were separated from the other options. Second, the options for which Pablo pays more (in blue) can be arranged on a line indicating the amount he pays extra, and similarly, the options which María likes better than Pablo (in red) follow a similar arrangement, albeit on a different axis, Fig. 3b. It should be noted, however, that these MDS plots are used mainly for visualization and not for statistical inference.

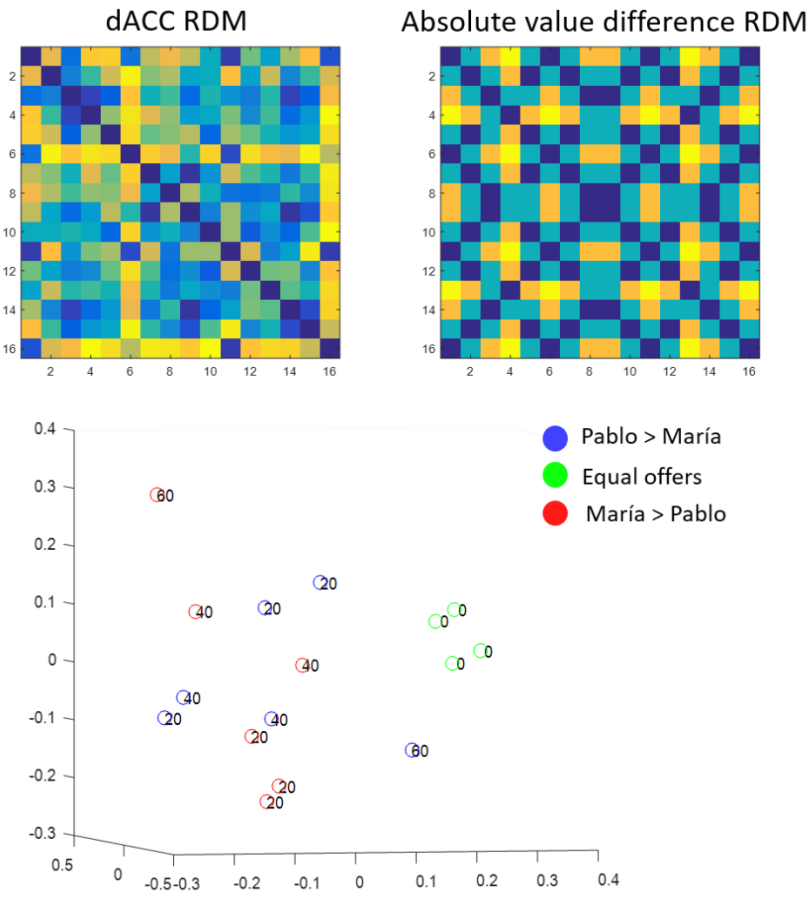


Figure 3. Neural RDM and MDS for dACC in the refresher task. A. The RDMs depict the distances between individual cars in the neural domain (left) and those predicted by the absolute value difference between the buyers' offers. B. The MDS projection leads to a value space in which all options with zero difference (in green) cluster together, while the other options can be arranged along two axes, one for those that Pablo pays more for (in blue) and those that María pays more for (in red).

What is the dimensionality of value encoding in the PFC? Equipped with this proof of concept, we next investigated four regions of interest (ROI), the dACC, vmPFC, and lateral frontopolar cortex (see Methods for how these ROI were defined). Each of these regions contributes uniquely to the decision-making process: while vmPFC is closely associated to the decision value and a "common currency" view of value (Levy and Glimcher, 2011; McNamee et al., 2013), dACC and the lateral frontopolar cortex have been associated with the value of unchosen options, alternative

courses of action, or currently unavailable background options (Boorman et al., 2009; Fouragnan et al., 2019; Kolling et al., 2016). Within each ROI we extracted the neural patterns for each individual car, split into two sets, one for all cars presented in a María market, and one for all cars presented in a Pablo market. This allowed us to ask whether the multivariate signal contained information about the irrelevant value. We compared the neural RDMs to four candidate model RDMs: One for the relevant value, another for the irrelevant value, and one each for the maximum and minimum value offered across buyers as value nuisance regressors. The results for the two RDMs of interest (relevant and irrelevant) across all ROIs are displayed in Fig. 4. First, we checked that a region that would not reasonably be expected to encode these variables, did not exhibit any correlations. We selected a region of primary visual cortex which correlated positively with trial onset in the univariate analysis and found that it encoded none of these variables ($ts_{25} < 1.06$, $ps > 0.29$).

The multivariate signals in the dACC and frontopolar cortex reflect opportunity costs. Next, we directed our attention to the dACC ROI, finding that it strongly encoded the relevant value, similar to the finding in the refresher market ($t_{25} = 3.51$, $p < 0.002$). Interestingly, however, the multivariate signal also carried information about the irrelevant value ($t_{25} = 2.50$, $p < 0.02$). Importantly, this does not reflect the counterfactual *unchosen* option as has previously been found (Boorman et al., 2013, 2009), but rather indicates that dACC retains some information about the value of the same option in a different *context*, as if it reflected the opportunity cost of selling this option now which would render it unavailable for selling in the following markets. Neither the maximum, nor the minimum value correlated

significantly with the dACC RDM ($ts_{25} < 0.41$, $ps > 0.68$). A similar pattern emerged when we looked at the bilateral frontopolar cortex, in a region similar to that reported in Boorman et al. (2009). Again, we found that it significantly encoded the relevant value ($t_{25} = 3.23$, $p < 0.004$), as well a slightly more weakly the irrelevant value ($t_{25} = 2.36$, $p < 0.03$), with no correlation with the other two variables ($ts_{25} < 1.29$, $ps > 0.21$).

The hippocampus supports reasoning about opportunity costs. In a recent study, Fouragnan et al. (2019) found that the hippocampus reflects the value of the currently unavailable choice in a binary decision-making task where options were sampled from a set of three options, one being unavailable. Analogously, we hypothesized that the hippocampus may selectively supply information about the value of the same item in the currently unavailable *context*. It may then supply this information to the dACC and frontopolar cortex. Indeed, the hippocampus exhibited the strongest encoding of the irrelevant value ($t_{25} = 3.68$, $p < 0.002$) out of all ROIs analysed, but did neither reflect the relevant value ($t_{25} < 1.63$, $ps > 0.11$), nor the other two value regressors ($ts_{25} > -0.56$, $ps > 0.57$).

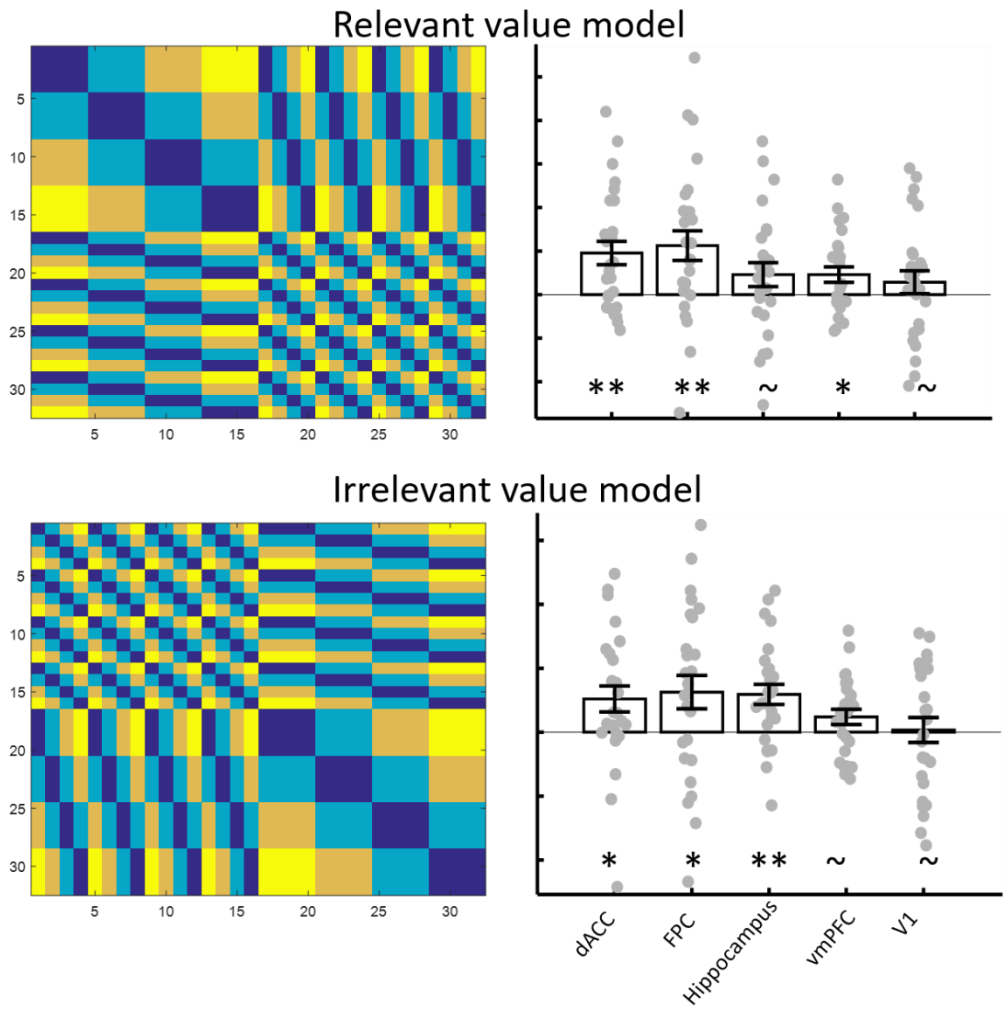


Figure 4. Model RDMs and regression results within each ROI. Each RDM is arranged as a 32x32 matrix, where rows and columns one to sixteen correspond to individual cars in the Pablo markets, and rows and columns 17 to 32 to cars in a María market. Cars were arranged according to Pablo value, hence the big squares in the upper left quadrant. Regression coefficients were tested for significance using a two-sided, one-sample t-test against zero, where * indicates $p < 0.5$, ** is $p < 0.01$, and ~ indicates a non-significant test result.

The vmPFC represents only the relevant value. Next, we asked whether the ventromedial prefrontal cortex, an area implicated in general-purpose value codes (Bartra et al., 2013; Lebreton et al., 2009; Levy and Glimcher, 2012) would nevertheless reflect the irrelevant value. However, this was not the case ($t_{25} < 1.97$, $ps > 0.06$), although the statistical test almost reached significance. Nevertheless,

vmPFC encoded the relevant value (t_{25} = 2.54, $p < 0.02$), but not the maximum or minimum value (ts_{25} < 1.11, $ps > 0.27$). This finding may corroborate the view that vmPFC encodes the decision value, i.e. the value which has already been processed to reflect the relevant aspects of a choice.

Discussion

Humans are able to flexibly value items according to arbitrary novel rules. However, the process underlying this revaluation is still poorly understood (Mante et al., 2013). Here, we presented a task in which participants flexibly reweighted the learned value of items in novel contexts. We found that they did so relatively accurately, but were negatively influenced by irrelevant item attributes, unlike previous work in similar domains (Flesch et al., 2018; Mante et al., 2013). Interestingly, the dACC and frontpolar cortex represented these currently irrelevant attributes as if they encoded the opportunity cost of parting with an item in a context where its value was inferior compared to a subsequent context.

Counterfactuals and opportunity costs

The effect we report contrasts with earlier accounts in that irrelevant attributes exerted a negative influence on behaviour. This effect was readily explained by normalization models (Carandini and Heeger, 2012; Heeger, 1992). However, here normalization occurred across attributes, rather than within a single domain as usually assumed, such as when agents compare items attribute by attribute (Gluth et al., 2018). Why might we observer such an effect? First, in our settings attributes (buyer's offer) were mutually exclusive, such that whenever

participants chose to sell a car to María, they perceived it as a waste if Pablo would have offered more for the same car. This is closely related to the concept of opportunity cost, however, in this case the cost is incurred because sales are made within an inferior context. We found this opportunity cost to be reflected in the dACC and frontopolar cortex. However, whilst these results are somewhat preliminary at this stage, they nevertheless fit remarkably well with and extend existing accounts of "counterfactual" reasoning: Previous studies have found that both the dACC (Fouragnan et al., 2019; Hayden et al., 2009), and the frontopolar cortex (Boorman et al., 2009; Buckley et al., 2009; Mansouri et al., 2017) encode the value of unchosen options. Importantly, however, these areas encoded only the options which was either currently unavailable for choice, or only the "second-best" option, as if they reflected the opportunity cost. In the case where options are currently unavailable this may seem suboptimal – a choice between two items should not be influenced by irrelevant items which are currently unavailable (Von Neumann and Morgenstern, 1944). However, when viewing the task as a whole, reasoning about the unavailable option enables the agent to make better choices the next time these become available. Through such processing, the agent may thus optimize global task performance rather than any local comparison.

The neural representation of opportunity cost

We found that three areas, the dACC, frontopolar cortex, and the hippocampus carried information about the currently irrelevant value dimension, as if they represented opportunity costs. However, unlike in previous work, there was

no apparent advantage to this representation, because these opportunity costs could not be used to inform decisions for a block of trials.

However, such counterfactuals may support decision-making in a more abstract way. First, these representations may intervene to stabilise the remembered value space which may otherwise suffer from interference through action selection, e.g. by imbuing the chosen option with *cached* value irrespective of market. Relatedly, such reasoning may hone participants' sensitivity for similar items in the subsequent context, thus aiding decision-making through mental simulation. Both processes may require the hippocampus, which was indeed found to selectively encode the irrelevant dimension, but not the relevant one. This pattern is remarkably similar to that observed by Fouragnan et al. (2019) in macaque monkeys.

However, more work needs to be done in the analysis of the neural data. One particular pertinent question is how different brain areas distort the value space to support the re-weighting according to the current context (Mante et al., 2013). In summary, in our preliminary analyses, we found the dACC, frontopolar cortex, and hippocampus to support counterfactual reasoning about opportunity costs, whereas the vmPFC was more readily explained by the relevant dimension alone. These results fit remarkably well with recent findings in the macaque monkey, but extend the potential role of the dACC and frontopolar cortex in the encoding of opportunity costs to higher-dimensional value spaces.

Methods

Participants

Twenty-six healthy volunteers with no history of psychiatric or neurological disorder participated in the study (12 male, mean age = 25.8, SD = 4.76). All participants gave written informed consent and the study received approval from the ethics board of the University of Granada. Participants were compensated for their time with €25 plus a performance-based bonus of up to €10. The exact sum paid was not saved due to a software error, but ranged between €8-10 in most cases.

Task

Participants performed that involved selling 16 different types of antique cars to buyers in different markets. The task consisted of four distinct phases. First, participants performed an "arena" task, in which they had to arrange all 16 cars on the screen according to how valuable they assumed them to be. Next, participants performed the "training phase" , in which they learned the value that two potential buyers (Pablo and María) would pay to acquire a car of this type. In order to do so, they saw each car 15 times in the centre of the screen, flanked by the names of the two buyers. Participants responded which of the two buyers they wished to sell to and then received full feedback about both buyers' offers. They were instructed to learn the exact values for each car, because they would need those in the subsequent phases. This phase lasted approximately 20 minutes and was performed outside of the scanner. Immediately after this learning phase,

participants performed another "arena" in which they had to arrange the cars according to the values they had just learned.

Whilst in the scanner, participants performed a shorter version of the learning (5 repetitions per car) without feedback, but were told they would receive bonus in accordance to the price their chosen buyer paid for the car. This phase lasted approximately 10 minutes (300 TRs). Lastly, the remainder of the scanning time consisted of six runs of 190 TRs each, in which participants found themselves in three different markets: One in which only Pablo was available to buy, another one in which only Maria was available, and an "unknown" market in which they did not know who would buy the car. Markets consisted of blocks of 15 trials, with the order of markets counterbalanced across runs. On each trial, participants saw two cars arranged horizontally along the screen. For each trial, cars were sampled uniformly at random without replacement. Participants had to choose which car they wished to sell. A central fixation dot indicated that their response had been registered, but otherwise no feedback was given. For bonus payment, ten trials were drawn at random from the scanner portion of the task, summed and then scaled so as to fall within the range of 0-10€.

Logistic regression analyses

We analysed participants' choices using a random effects model predicting either (GLM1) choosing the option on the left side, or (GLM2) choosing the correct option, i.e. the one with highest value according to the market. Three independent regressors were entered in each analysis: relevant value difference, irrelevant value difference, and their interaction. Value differences were entered as left option

minus right option (GLM1) or absolute differences (GLM2). GLM1 was also performed separately for each market using the value differences of each buyer to test how weighting of these factors differed across contexts. All regressors were z-scored and significance was assessed using a two-sided one-sample t-test against zero of the beta coefficients across subjects.

Normalization models

We tested two models of divisive normalization: Model one was based on the classic divisive normalization formula following (Louie et al., 2013), where each option is represented relative to the sum of all options:

$$nV_i = K \frac{V_i}{\sigma + w \times (V_i + V_j)}$$

Where K is a gain parameter, σ is the semi-saturation parameter, owing to the nature of the model as a model of neuronal firing rates, and a weighting parameter w. Options *i* and *j* correspond to Pablo's and Maria's value for each car.

The second model was a simplified version in which the options got boosted (or penalized) based on the relative difference between the values paid by the two buyers, as follows:

$$nV_i = V_i \times (1 + w \frac{V_i - V_j}{V_i + V_j})$$

Where the parameter w governs the boost (or penalty) conferred on each option based on the relative difference.

Parameters were estimated using maximum likelihood fitting and models were then compared using Bayesian model selection (Stephan et al., 2009).

FMRI Data Acquisition and pre-processing

MRI data were acquired on a 3T Siemens scanner. T1 weighted structural images were recorded directly prior to the task using an MPRAGE sequence: 1x1x1mm^3 voxel resolution, 176x256x256 grid, TR = 1900ms, TE = 2.52ms, TI = 900ms. Each fMRI image contained 32 axial echo-planar images (EPI) acquired in descending sequence with a voxel resolution of 3.5 mm isotropic, slice spacing of 4.2mm, TR = 2000ms, flip angle = 80, and TE of 30ms. 1440 EPI images were recorded per participant, resulting in a scanning time of around 48 minutes. Data were pre-processed using SPM12 (Wellcome Trust, London). As EPI acquisition used a descending sequence, images were corrected for slice time acquisition with the middle slice (at TR/2 = 1 s) as reference to minimize interpolation errors (Sladky et al., 2011). All RSA analyses, however, were carried out on the re-aligned images and thus did not involve slice time correction. Scans were realigned to the first scan within each session. The anatomical scan was co-registered to the mean functional image. Anatomical scans were normalized to the standard MNI152 template brain. The functional EPI images were then normalized and smoothed with a full width half maximum Gaussian kernel of 8mm for the mass-univariate analyses.

Univariate fMRI analyses

Data were analysed using SPM12 and custom scripts. Analyses were carried on the slice-time corrected and smoothed images. Stimulus onsets were incremented by 1s to account for slice time correction to the middle slice, which occurred at TR/2

= 1s after stimulus onset (Sladky et al., 2011). All GLMs used the standard hemodynamic response function and its time and dispersion derivatives, convolved with a stick function of zero seconds duration time-locked to event onset. The six motion correction vectors from realignment were included as nuisance regressors. Automatic orthogonalization was switched off. Data were visualized using the xjview toolbox for SPM (http://www.alivelearn.net/xjview).

We carried out two main GLMs. All GLMs modelled five distinct events: The onset of every trial during the refresher block, each screen indicating the next market (three per run), all trials containing the "single" markets, all trials containing the "mixed" market, as well as all missed response trials as a nuisance regressor (~ 1.3% of trials, range: 0%-18.3%). GLM1 included the following parametric regressors of interest: At single market trials, the chosen minus unchosen value of the relevant dimension (e.g. Pablo's value in a Pablo market), the irrelevant dimension, and their interaction. For completeness, we included for the refresher chosen and unchosen value, and reaction time. All events where the participant responded contained a nuisance regressor encoding response hand.

GLM2 included the same regressors as GLM1 with one variation: Instead of modelling chosen minus unchosen value, we entered the high minus low value differences and their interaction.

Representational Similarity Analysis

RSA analyses were carried out on the realigned images using custom scripts and the RSA toolbox (Nili et al., 2014). We fit two GLMs to derive the patterns of

activation: Each GLM contained one event for each car that was presented on the screen, i.e. on market trials there were always two events which were decorrelated across the experiment due to the unconstrained sampling of cars. GLM1 modelled only the refresher run, whereas GLM2 modelled all runs. For GLM2, car presentations were further split according to which market they occurred in, leading to a total of 48 unique events. These unique patterns were first pre-whitened using the residual covariance matrix voxels within the ROI/searchlight. Their dissimilarity was then assessed using 6-fold cross-validated correlation distance across runs of the market phase. For GLM1, no cross-validation was used as only a single run was modelled.

We assessed whether a region's RDM correlated with a (z-scored) model-predicted RDM using multiple regression into which all candidate models were entered, thus accounting for potential correlations between RDMs. Significance was then assessed using a one-sample t-test against zero of the resulting beta coefficients. For the searchlight analysis, the resulting beta maps were transformed back into MNI-space, smoothed and then tested at the group-level.

ROI selection.

ROI were selection based on orthogonal contrasts from the imaging data. The dACC and frontopolar regions were selected based on the RSA analysis of the refresher task; visual cortex and vmPFC were selected based on the clusters engaged by trial onset in the refresher task; finally, the hippocampus ROI was defined without functional contrasts using xjview's database.

Chapter Six: Where does value come from?

Abstract

The computational framework of reinforcement learning (RL) has allowed us to both understand biological brains and build successful artificial agents. However, in this article we highlight open challenges for RL as a model of animal behaviour in natural environments. We ask how the external reward function is designed for biological systems, and how we can account for the context sensitivity of valuation. We summarise both old and new theories proposing that animals track current and desired internal states and seek to minimise the distance to goal across multiple value dimensions. We suggest that this framework can readily account for canonical phenomena observed in the fields of psychology, behavioural ecology, and economics, and recent findings from brain imaging studies of value-guided decision-making.

The reward paradox

What is it that animals seek to achieve by their behaviour? For evolutionary biologists, there is a simple answer to this question – that animal behaviour has evolved to maximise reproductive fitness, i.e. the propensity to produce offspring that will carry forth our genes. However, for those studying the behaviour of humans and animals in the fields of ethology, economics, psychology and neuroscience, this answer is only partly satisfying. Critically, it stops short of specifying how animals can learn which behaviours will increase or decrease their

fitness. Instead, behavioural scientists propose that the environment furnishes reinforcement signals that indicate the likely costs or benefits of an action and assume that these signals can be directly sensed by the agent, allowing them to behave in ways that maximise utility over the short or long term. In machine learning research, this is sometimes called the "reward hypothesis" (Sutton and Barto, 2018).

Over the past half century, the fields psychology and artificial intelligence (AI) have jointly developed a normative theory, known as reinforcement learning (RL), that is founded on the reward hypothesis (Dayan and Daw, 2008; Sutton and Barto, 2018). RL conceives of each encounter with the world as involving three distinct stages: (i) an agent receives sensory observations from the environment; (ii) the agent takes actions that influence its future states; and (iii) the agent receives a scalar reward signal that is emitted by the environment, and that is processed by the agent via a dedicated input channel (Fig. 1a). This conceptualisation lends itself naturally to modelling the canonical laboratory paradigm for operant behaviour, whereby an animal receives a sensory stimulus (e.g. a visual signal or tone), takes an action (e.g. presses a button or licks a spout) and receives a positive or negative reinforcer (e.g. liquid, food, or money). The framework successfully accounts for a wide array of habit-based and goal-directed behaviours in humans and other animals (Dolan and Dayan, 2013), and prominently explains the neural signals that accompany the expectation and delivery of reinforcers (Schultz et al., 1997). RL has furnished the canonical computational framework for understanding value-guided decision-making in humans and other primates, and has been widely used to explain

how different brain regions participate in valuation and choice (Dayan and Daw, 2008). Moreover, RL models are currently offering some of the most exciting advances in AI research, producing artificial agents that can behave intelligently in complex, dynamic environments such as board and video games, thereby providing a proof of concept that intelligent behaviour can emerge *de novo* under the assumptions of the reward hypothesis (Mnih et al., 2015; Silver et al., 2017).

The RL framework offers an important set of computational tools for understanding animal learning and behaviour. However, as a theory of biological intelligence, RL focusses on how rewards influence behaviour but does not specify the provenance of the external reward function against which behaviour is optimised. In machine learning, the reward function for an RL model is hand-designed by the researcher. For example, when training an autonomous vehicle, she might specify that it needs to reach its destination as quickly as possible whilst obeying the traffic laws, by assigning benefits and costs for desired and undesired behaviours. However, in natural environments, no external entity exists that can directly quantify the consequences of each action, like the points that are awarded in a video game for completing levels or shooting monsters (Fig. 1b). Nor is it obvious that biological systems have a dedicated channel for receipt of external rewards that is distinct from the classical senses. Rather, rewards and punishments *are* sensory observations – the taste of an apple, the warmth of an embrace – and so stimulus value must be inferred by the agent, not conferred by the world. In other words, rewards must be intrinsic, not extrinsic (Box 1). We call this "the reward paradox".

Here, we consider the challenges of applying the reward hypothesis to the study of biological agents in the natural world, with a focus on the neural and computational mechanisms by which humans and other animals make value-guided choices. In doing so, we highlight some ways in which the RL framework might be modified, adapted or nuanced to account for the reward paradox. We appeal to a diversity of theories concerning motivated behaviour, and ask how they can explain the behaviour and brain activity observed during human decisions in cognitive, social and economic settings.

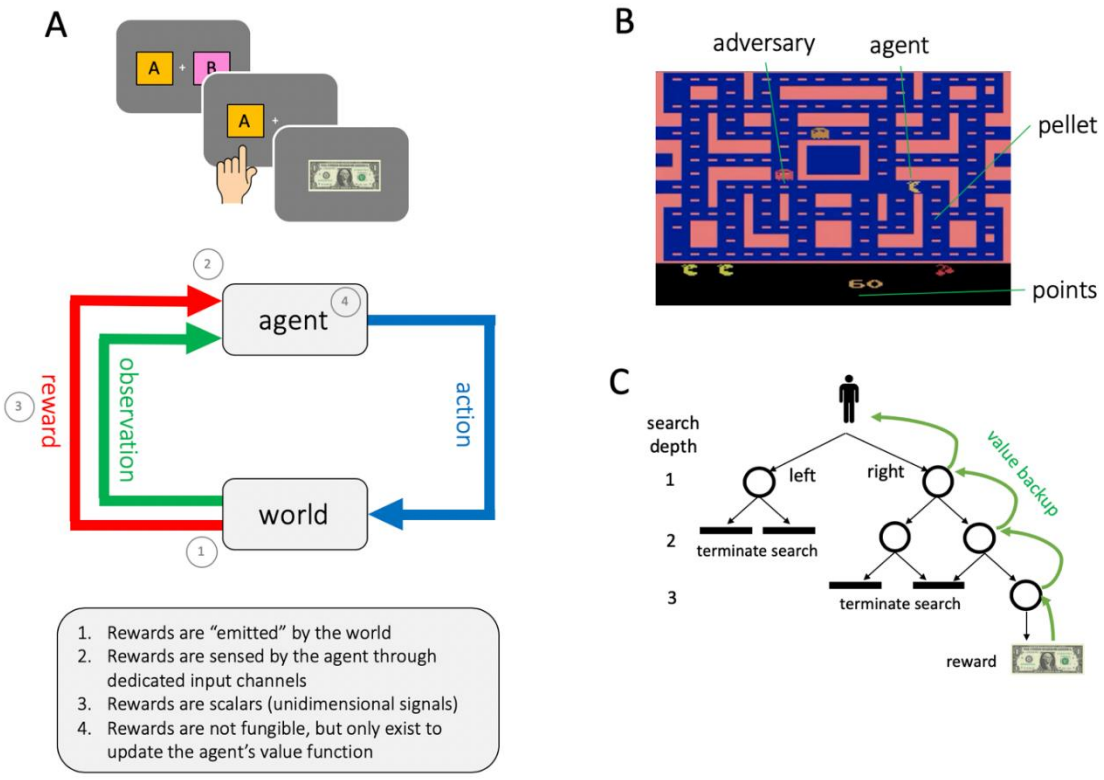


Figure 1. A. Schematic representation of the RL framework. Top: a typical "bandit" task in which participants choose between options (A and B) with uncertain reward probability, incurring financial outcomes. Below: in the RL framework, an agent receives observations (green), takes an action (blue) and receives feedback in the form of reward (red). The numbers highlight some assumptions of the basic RL framework that are less well suited to explaining motivated behaviour in biological

organisms (grey box below). B. The RL framework applied to video games (here, Ms. Pac-Man). The agent acts in order to maximise game score, which is a scalar quantity. Unlike humans playing the game, the agent receives the score differential directly (e.g. points for eating pellets) as reward signal via a dedicated input channel. C. Schematic of model-based RL, where the agent searches through a tree of possible future states and actions, sometimes stopping if no reward is expected from this branch of the search tree, and ultimately choosing the trajectory through the tree that leads to the highest return.

Context-dependent valuation and the representation of internal state

To circumvent the reward paradox, it is sometimes assumed that the reward function has been designed by evolution. In other words, reward computation may rely exclusively on fast, mandatory, phylogenetically prespecified mechanisms that automatically convert sensory observations into hedonic signals that act as a proxy for "external" rewards. Indeed, there is no doubt that dedicated mechanisms exist for those physiological drives that lie at the bottom of our "hierarchy of needs" (Maslow, 1943), such as hunger or fatigue. These needs are regulated by projections from the hypothalamus to neural systems that respond to rewards, including dopaminergic neurons in the ventral tegmental area (Fadel and Deutch, 2002). On this basis, proponents of RL models have dubbed dopamine neurons the "reward retina" as if they were responsible for directly sensing reinforcement from the environment (Schultz, 2015).

However, one challenge for this perspective is that the reward value of a stimulus depends on the agent's internal state as well as the context provided by the external environment. Intuitively, the value of a potentially rewarding stimulus (e.g. a hamburger) will depend on the agent's ongoing bodily needs (am I hungry?).

Under the RL framework, this sensitivity to internal context might be accounted for by assuming that bodily states form part of the environment, i.e. that the state of "I am hungry" is external to the agent in just the same way that the presence or absence of an apple is a property of the world, not just the mind. However, this argument becomes harder to defend when one considers the range of affective and cognitive factors that can potentially modulate value. For example, a hamburger might be less valuable to someone who is nervous before an exam, or to a committed vegetarian. It seems harder to argue that private mental phenomena such as moods or moral attitudes are properties of the external environment.

More generally, standard RL models assume that reward is a purely hedonic signal that is used solely to update the agent's value function. Reward is thus not *fungible* – that is, once acquired, it is not exchangeable for other assets, and cannot be used in a way that would alter future observations or facilitate future actions. For example, when RL agents are learning to play video games, the rewards they receive (i.e. game score differential) evaporate instantly without providing future opportunities for disbursement within the game. This is unlike natural settings, where rewards are typically accumulated because they lead to persistent changes in state. For example, bodies act as repositories for energetic resources that are harvested by the agent during trophic behaviours, and bank accounts are used to store for income acquired in exchange for work. The agent can thus choose to exchange one stored asset for another that is needed, the defining feature of economic behaviour in the marketplace. In the laboratory, rewards may be implicitly accumulated as experimental animals or human participants become more sated or

wealthy as a result of the liquid rewards or financial incentives that they receive for performing the task, but these changes in state are typically considered to be nuisance variables by the experimenter. For example, when a monkey working for liquid reward is no longer thirsty enough to work, the researcher will typically terminate the experiment. To fully account for biological behaviour, our models of value learning need to account for the internal state of the organism.

Psychologists and neuroscientists who study appetitive responses to primary reinforcers such as food and liquid have long appreciated that theories motivated behaviour need to include a representation of internal state. One prominent model, known as incentive salience theory, proposes that whilst an animal might "like" a stimulus on the basis of cached state values (e.g. sugar tastes good) as proposed in standard RL, the extent to which it "wants" a stimulus also depends on its physiological state (e.g. hunger), which adjusts valuation via a gain control mechanism. This ensures that stimuli that satisfy natural appetites will be consumed more readily (e.g. food when hungry and water when thirsty). Incentive salience can account for a wider range of empirical phenomena. For example, the rate of approach towards food depends on hunger levels, and adapts immediately according to internal levels of satiety irrespective of prior training (Mollenauer, 1971). Similarly, food items that have been administered during states of food deprivation are preferred over control items during a later test when the animal is sated, suggesting they have acquired greater value (Pompilio and Kacelnik, 2005). However, incentive salience does not describe exactly how internal state is represented, or specify how the animal chooses which appetite it wishes to satisfy. More widely, a rich literature has studied the consequences of satiety on food

choices, often invoking model-based RL to describe why animals take actions that elicit a reinforcer that has not been devalued over one that has (Wikenheiser and Schoenbaum, 2016). However, this literature has largely been silent about the mechanisms by which devaluation itself occurs, focussing instead on how animals learn the causal structure of the task.

In the remainder of this Chapter, we explore the relevance of these considerations for the study of human decision-making. Humans routinely make decisions about complex stimuli or abstract plans of action in cognitive, social and economic settings. Our argument is that a representation of internal state may be computationally relevant even for decisions between consumer products, social opportunities, or lifestyle alternatives. We adapt models that have been previously proposed to account for choices among basic needs to these more intricate decision settings, and describe how they might account for some intriguing findings concerning the neurobiology of human choice.

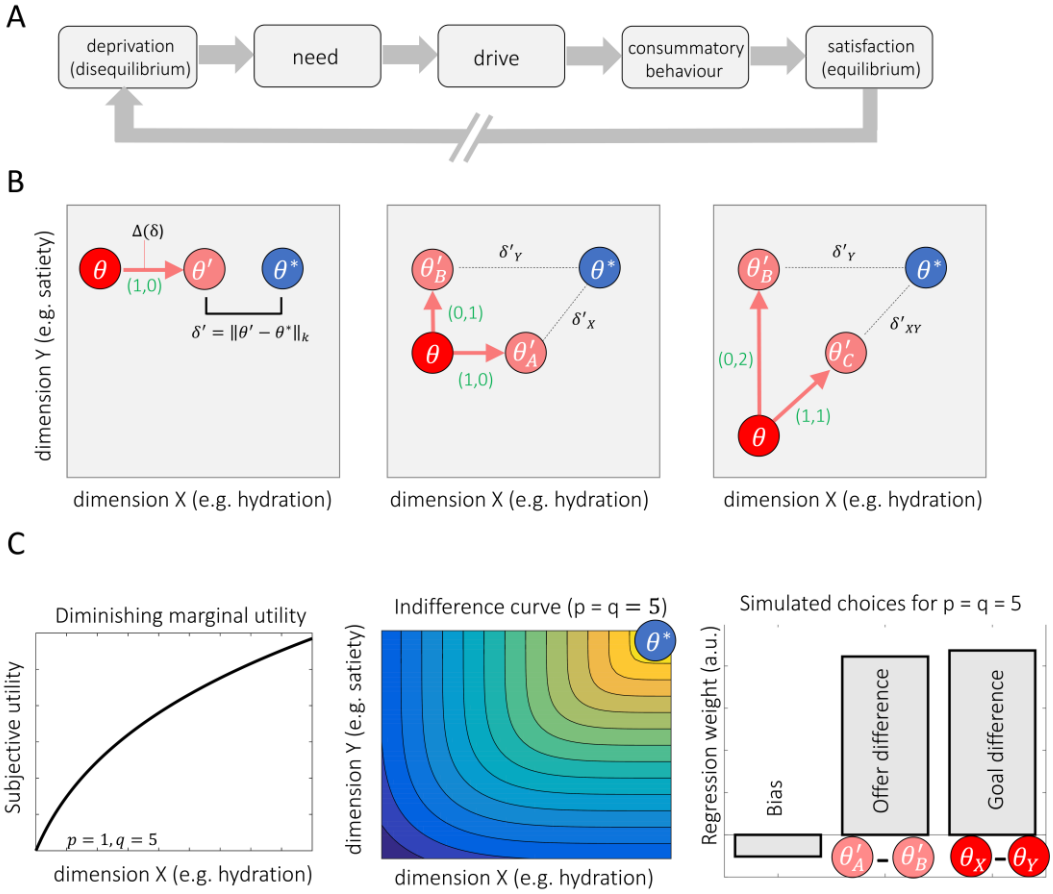


Figure 2. A. Schematic of Drive Reduction Theory. Behaviours are initiated to reduce drives, which in turn are provoked by needs that arise following homeostatic imbalance. B. Illustration of homeostatic reinforcement learning (HRL) , here in two example dimensions (labelled satiety and hydration). The agent's current state is denoted θ, candidate next states θ' and the desired state (setpoint) θ^*. Offers in the (x,y) dimension are shown in green. In the leftmost panel, the agent is sated but thirsty, as in a laboratory experiment where a monkey is working for liquid. An action reduces the distance δ between the current and desired state $\|\theta - \theta^*\|_k$ (where k indicates the Minkowski distance; we use this here for simplicity, but note that a more general, two-parameter version depicted in panel C is needed to account for non-linearities in valuation, see Box 2 for details). Notice that the reduction in distance is proportional to what would (in the RL framework) be construed as an external reward signal. In the middle panel, an agent who is more thirsty than hungry is offered an equivalent quantity of food (0,1) or water (1,0). Water reduces distance to goal (Euclidean, $k = 2$) more than food and is thus preferred. In the rightmost panel, the agent is both very hungry and very thirsty. An offer comprising a bundle of both food and water (1,1) is preferred over a large quantity of food (0,2). C. Predictions of HRL with two parameters, p and q, governing non-linearities. The model can reproduce the ubiquitous finding of

diminishing marginal utility (left panel). As p and q increase, distance estimates depend more strongly on just one dimension, indicating that maximum weight is given to the currently most needed asset (middle panel). Intermediate values p and q can reproduce one of the central findings of Chapter Four: that choices are both driven by offers and needs when assets need to be kept in equilibrium (see also Fig. 4).

Goals as cognitive setpoints

Our argument begins with the claim that human motivated behaviour is intrinsically goal-directed in nature. Humans will often work tirelessly to complete tasks whose only apparent reward is the satisfaction of task completion itself, for example when seeking to win a race, solve a puzzle, or climb a mountain (O'Reilly et al., 2014). Indeed, this purposive, self-consistent property of human behaviour is so powerful that humans will seek to complete tasks even when the outcome is known to be a net loss, a phenomenon known as the sunk cost fallacy (Knox and Inkster, 1968). Although the RL framework embraces goal-directed behaviours, it does so under the assumption that animals explicitly learn the transition probabilities between states and use mental tree search to identify the most rewarding trajectories, which then guide "model-based" decisions (see Fig. 1c). Thus, for RL, goals are not the objects of computation themselves but emerge as a by-product of policies which optimise expected future reward (Daw et al., 2005). Other elaborations of the RL framework explicitly propose additions to the cost function, such as a desire for novelty, agency or consistency in behaviour, which draw upon the assumption from cognitive psychology that humans do more than simply maximise reward (see Box 1).

Box 1. Intrinsic Motivation. Machine learning and AI researchers know that designing a reward function by hand is costly and does not scale to real world settings. To avoid this problem, agents have been built that generate their own intrinsic motivational signals (Oudeyer and Kaplan, 2007; Ryan and Deci, 2000; Schuck et al., 2016; Singh et al., 2010). For these agents, the cost function is populated by additional terms that encourage the agent to pursue adaptive behaviours even in the absence of reward. Many of these additional terms draw inspiration from psychological theory, which has long noted that humans and other animals often pursue activities for their own sake (e.g. a kitten furiously chasing a piece of string, or a commuter completing the crossword on her journey to work). For example, one key intrinsic motivation may be to explore the world in a structured fashion, satisfying curiosity and avoiding boredom (Gottlieb and Oudeyer, 2018). This allows agents to seek out experiences that may accelerate learning without directly enhancing reward, a phenomenon known as active learning. Other intrinsic motivation signals may encourage us to make actions with predictable consequences, or that exert maximal control over the environment or the behaviour of conspecifics, or those that minimise discrepant or dissonant inputs (see (Oudeyer and Kaplan, 2007) for a review). Augmenting the reward function with these intrinsic motivational signals often leads to faster learning and more stable patterns of behaviour in artificial agents (Guo et al., 2015). However, in most reports, intrinsic motivation signals supplement (rather than replace) rewards from the external environment, leaving the thorny question of where rewards come from only partially answered. Other approaches that avoid any external feedback by using purely unsupervised or self-supervised methods – where the agent is optimised to maximise information gain or reduce sensory surprise – hold undoubted promise but have not yet been observed to yield complex behaviours in dynamic environments that resemble the natural world (Friston, 2010).

In considering these phenomena, we turn to one of the earliest theories of motivated behaviour – one which assumes that animals seek to keep internal drives in equilibrium (Keramati and Gutkin, 2014; Pezzulo et al., 2015). This theory builds on early models of homeostatic function (Berridge, 2004) and found its first prominent expression in the 1940s with Hull's Drive Reduction Theory (Hull, 1943). These homeostatic models argue that motivated behaviours seek to restore imbalance among different physiological needs, such as warmth, satiety, and

hydration (Fig. 2a). A central idea is that neural circuits encoded desired states or "setpoints", and drives result from the disparity between current and desired state. Thus, rewards are multidimensional quantities signalling the drive reduction that occurs when an imbalance among the organism's multiple internal needs is redressed.

Here, we make a connection between homeostatic models of reward, which have emphasised balance among basic drives and interoceptive processes, and recent work on value-guided decision-making in humans and other primates (Kable and Glimcher, 2009; Rangel et al., 2008; Rushworth et al., 2011). We argue that motivated behaviour in cognitive, economic or social settings can similarly be understood as seeking to restore a balance among competing "setpoints" that correspond to higher-order goals. Unlike currently popular theories of value-guided choice, which are largely inspired by RL models (Dayan and Daw, 2008) or expected utility theory from economics (Glimcher, 2004), our view emphasises the multidimensional nature of value signals, the dependence of motivated behaviour on the internal state of the agent, and the desire to explicitly realise goal states (constraint satisfaction) rather than maximise reward. In what follows, we show how extant theories of motivation can explain some otherwise puzzling behavioural and neural findings in the literature on value-guided decision-making and economic choice.

Homeostatic reinforcement learning for cognitive behaviour

We suggest that during learning, humans form new setpoints pertaining to cognitive goals. For example, we might represent current and desired states on axes pertaining to financial stability, moral worth, or physical health as well as hunger, thirst or temperature. To sketch the computational underpinnings of this process, we appeal to an existing framework (Keramati and Gutkin, 2014) for understanding motivated behaviour for basic physiological processes, known as homeostatic reinforcement learning (HRL; Box 2). This theory proposes that current states and goals are encoded in a multidimensional "value map". Motivated behaviour can then be seen as an attempt to minimise the maximum distance to setpoints in this value space. Repurposing this framework for cognitive settings, agents commit to policies which focus on purposively driving the current state towards setpoints on a particular goal dimension, such as caching resources, building a shelter, obtaining a mate, or enhancing professional status. In doing so, their ultimate goal is to maintain equilibrium among all goal states, achieving what might be popularly characterised as a state of "wellbeing".

In HRL, reward emerges naturally as a consequence of the computations required for learning, rather than being furnished by the external environment. This allows the "reward paradox" highlighted above to be sidestepped. To illustrate, consider an animal whose internal state is currently described by parameters θ and who seeks to achieve a desired state θ^*. Let us assume that the agent has a biologically plausible functional form, such as a neural network. In order to calculate the gradients that will allow network weights to be optimised for future behaviour, it is

necessary to compute the loss term $\Delta(\|\theta^* - \theta\|_k)$, which indicates the change in distance between a current and desired state that is afforded by any action. Now consider a typical laboratory study in which a single reward dimension is relevant, for example because the experimenter reduces prior liquid intake and offers only water in exchange for behaviour. For a thirsty animal, this loss term is now mathematically identical to what one would construe as "reward" in the RL framework – i.e. the volume of liquid administered on a trial (Keramati and Gutkin, 2014). This equivalence will break down as θ tends towards θ^*, but by this point the researcher will typically have ended the experiment because the animal is fully hydrated. Thus, although in this experiment value is artificially constrained to be unidimensional, we note that from the agent's perspective, multiple value dimensions may continue to be relevant. For example, a human participant might be tempted to neglect the task in favour of checking their social media account, in order to enhance social capital. However, this individual would most likely be excluded from the analysis. Homeostatic reinforcement learning readily incorporates other accounts of reference-dependent evaluation by adjusting how the agent computes θ^*. For instance, it may do so by using a running average over past options and anchor its behaviour to this reference. When $k > 1$, simply assuming that other drives exist that cannot currently be satisfied immediately gives rise to the ubiquitous phenomenon of diminishing marginal utility. In other words, as the agent acquires the currently relevant asset (such as liquid reward) its attention will progressively tend towards other relevant assets, diminishing the contribution of liquid reward to its well-being (Fig. 2c).

Box 2: Homeostatic Reinforcement Learning

Let us assume that sensory observations, $\boldsymbol{X_t}$ are mapped onto a persistent internal state $\boldsymbol{\theta_t}$:

$$\theta_t = X_t \times U + \theta_{t-1} \times A_\theta$$

Eq.1

Where $\boldsymbol{U}$ is a matrix mapping the chosen option $\boldsymbol{X_t}$ onto $\boldsymbol{\theta_t}$. The internal state, $\boldsymbol{\theta_t}$, exhibits dynamics whereby it will change (decay) in the absence of any input, e.g. credit card debt will increase if it is not repaid, just as thirst will increase if not slaked. These dynamics are captured by the matrix $\boldsymbol{A_\theta}$, which governs the intrinsic changes over time.

Note that this implies that the correspondence between internal and external states must be learned, i.e. our preferences are not simply given to us but acquired with experience, and that we might be uncertain about whether we are hungry, tired or happy. This relates to the framework of "active inference", under which reward values are inferred rather than being directly conferred (Pezzulo et al., 2015). We conceive of this internal state space as a multi-dimensional map whose dimensions variously pertain to both physiological states (e.g. satiety) and cognitive goals (e.g. social status). Internal states are subject to momentum via recurrent connections so that states (or moods) may persist in the face of a change in external circumstance.

In HRL, the agent represents a set of desired (or aversive) points $\boldsymbol{\theta}^*$ on this space, allowing it to estimate the distance (and contributing features) of its current position to the goal state. In doing so, the agent needs to combine across the relevant goal dimensions and weight these according to their relative contribution to this distance by computing the Minkowski distance:

$$\delta_{t+1} = \|\theta_{t+1} - \theta^*\|_k = \sqrt[p]{\sum |\theta_{t+1} - \theta^*|^q}, for\ p = q$$

Eq.2

The parameter $\boldsymbol{k}$ allows the agent to prefer bundles that only consist of one single asset ($\boldsymbol{k < 1}$), be indifferent between all bundles ($\boldsymbol{k = 1}$) or prefer mixed bundles ($\boldsymbol{k > 1}$). The model is thus well placed to account for commonly observed indifference curves between two or more economic goods. If we further assume that there are two non-linearities, p and q, rather than just k, our model can further account for diminishing marginal utility, Fig. 2c. The implications of this computational step are described more fully in (Keramati and Gutkin, 2014).

Finally, where, then, is value (or: reward) in this framework? Value arises naturally when the agent compares its current distance to its desired state to its previous distance, which summarizes that the agent has progressed on at least one of its goals since the last time-point:

$$R_{t+1} = \delta_t - \delta_{t+1}$$

Eq.3

Value is thus a summary of whether the agent is approaching or retreating from its goals. This model encapsulates situations in which multi-attribute objects need to be integrated into a single reward signal, such that $\boldsymbol{R = X \times U}$ if $\boldsymbol{A_\theta = 0}$ and k = 1.

Implications of HRL for cognitive decisions

More generally, we argue that some of the most complex and abstract decisions that humans make might be better described by a process that optimises over states, rather than rewards. For example, consider a high school student choosing a career path. Under the (model-based) RL framework, she must consider an impossibly large number of potential futures, and select whichever is going to be most rewarding. This seems to imply the devotion of disproportionate levels of computational resources to the search problem. The approach advocated here implies that she first selects a goal state (e.g. become a lawyer) than then takes actions which minimise distance to that goal. For example, she seeks to go to law school; to maximise her chances, she first studies hard for her exams. This explanation seems to accord better with our common sense intuition of how the complex choices faced by modern humans are made. However, the computations involved may build upon more phylogenetically ancient mechanisms. For example, one of the most prominent theories of insect navigation proposes that in order to reach their home base, central place foragers such as honey bees and desert ants (Webb, 2019) initially encode an egocentric snapshot of their base and subsequently, on the return journey, use a similarity-matching process to gradually reach their goal, as if they are performing gradient descent over states.

Our view relates to an earlier proposal that animals are motivated to satisfy a "current concern" and understands disorders of mental health as disrupted goal selection, pursuit and appraisal (Klinger, 1975). A fleshed out computational theory

of this process will require the specification of how agents learn the relationship between individual states and the degree of goal realisation. One model proposes a distinction between “primary” and “learned” value, where the former resembles the traditional reward signal in the RL framework and the latter quantifies the eligibility of states for goal realisation (O’Reilly et al., 2014). A very appealing aspect of this framework is that it provides a natural way to understand the affective states that pervade our everyday mental landscape, including satisfaction (goal completion), frustration (goal obstruction), and disappointment (goal abandonment), which have largely eluded computational description thus far (O’Reilly et al., 2014). A more directly related machine learning approach directly computes the feasibility of attainment of goal states, showing how it can be used for compositional inference under the nonstationary reward functions characteristic of real world environments (Ringstrom and Schrater, 2019).

Many of the more complex behaviours exhibited by humans involve decisions about money. Money has the peculiar property of being fluidly exchangeable for other assets (except, as the Beatles remind us, for love). Our theory might seem to break down in financial settings, because it seems incongruous to propose a “setpoint” for wealth – anecdotally at least, the accumulation of wealth does not blunt the desire to acquire new wealth. However, there is evidence that even wealth accrual may be subject to some of the constraints identified here. For example, taxi drivers on shifts of variable length will tend to work until a fixed return has been achieved, returning home even when this “wealth setpoint” falls during a period of high yield – for example, even if customers are plentiful and fare rates

are high (Camerer et al., 1997). It is also observed that humans use mental accounting to allocate wealth to different potential value dimensions in advance (Thaler, 2018). Thus, although money can be deployed fluidly to facilitate movement in any direction in value space, humans mentally pre-allocate funds to the different directions of advancement in value space, treating an unexpected excess in one direction with profligacy rather than reallocating it facilitate progress along other dimensions.

Pure drive reduction theories fell from favour with the realisation that animals will work for rewarding stimuli when their drives are satisfied, i.e. purely for their hedonic value (Miller and Kessen, 1952; Olds and Milner, 1954), thus it seems likely that some behaviours are driven by "liking" alone, irrespective of any relation to the current setpoint (Zhang et al., 2009). Similarly, it is unlikely that all human cognitive decisions are based on minimising distance to a desired goal – the mechanisms proposed here presumably exist alongside other processes that allow behaviours that have been pleasurable or successful in the past to be repeated in the future. Thus, we might buy a pair of shoes on impulse, sneak an extra bite of dessert even when we are full, or linger at a party even after our social duties have been fulfilled. However, our argument is that the human tendency to optimise towards desired states has been overlooked in the face of enthusiasm about rewards as the primary drivers of behaviour.

Goals, states, and values in the OFC

In the final sections, we consider the implications of the perspective advanced here for the neurobiology of decision-making. For valuation to depend on the internal state of the agent, these states must be explicitly represented in brain signals. However, most neural studies of value-guided choice have used a paradigm in which the animal or human participant makes a succession of unrelated, independent choices between food items or monetary gambles. This paradigm is not well suited to measuring how variation in internal state (how sated am I right now? How much wealth have I accumulated?) might be coded in neural circuits. However, recently researchers have begun to employ more complex paradigms that may shed some light on this issue.

At the neural level, the integrity of the orbitofrontal cortex (OFC) is critical for goal-driven reward behaviour (Burke et al., 2008). Interestingly, a recent theory proposes that the OFC constitutes a "map" of both observable and latent states within a task (Schuck et al., 2016; Wilson et al., 2014), similar to the framework proposed here (c.f. Eq.1 in Box 2). Moreover, OFC may code endogenous (internal) as well as exogenous (external) value, as firing rates (Padoa-Schioppa, 2013) and BOLD signals (Abitbol et al., 2015) can be choice-predictive even in the baseline period prior to stimulus onset. An emerging theory suggests that OFC encodes affective states that pertain to ongoing positive or negative value, i.e. moods or "momentum" in internal value states (Eldar et al., 2016; Vinckier et al., 2018). Together, these findings imply that OFC is a likely candidate for representing ongoing estimates of position on the value map.

One recent study directly tested whether BOLD signals in OFC tracked the internal accumulation of financial resources across an episode in a way consistent with the ongoing internal state representation proposed in Chapter Three. Participants performed an economic task that involved gambling for financial incentives over successive, independent trials. For a rational agent performing this task, there is no imperative to track the accumulated level of payoff incurred by choices – to maximise return, one simply needs to select, on each trial, the gamble with the highest expected value. However, when humans performed this gambling task over multiple discrete episodes, each signalled by a prominent contextual cue that was structurally irrelevant to the task, brain activity in the medial OFC tracked the level of accumulated resources (or "wealth") over the context (Fig. 3). This "value accumulation" signal occurred even though the accumulated rewards were never overtly signalled to participants, nor were they instructed or incentivised to calculate their ongoing wealth.

Another brain imaging study revealed that a nearby region of medial prefrontal cortex tracks internal states during wealth accumulation (Tsetsos et al., 2014). Participants were asked to choose among four alternatives (bandits) that either paid or charged an uncertain sum of money. Critically, each selected bandit was fungible – it continued to incur financial costs or benefits across a short block of trials, as if the agent had purchased that asset and was continually profiting (or otherwise) from its possession ("rule in" condition). In a different condition, the bandits began in the agent's possession and decisions had to be made about which (if any) bandits to "rule out". This allowed us to assess how neural encoding of value depended on

whether an asset was being bought or sold. Critically, BOLD signals in the ventromedial prefrontal cortex (vmPFC) encoded the cumulative expected value of assets across the block (Fig. 3), and only coded for momentary rewards when decisions were made with future consequences for reward (e.g. when a bandit was selected in the "rule in" condition, or not selected in the "rule out" condition). Together, these studies suggest that the reward system tracks level of accumulated resources over time, for example so that future decision policies can adapt according to the current internal state, in a way that is compatible with the theory advanced here.

Budget rules: aspiration and avoidance points

The natural world is structured in such a way that some states are critical for survival or have substantial impact on long-run future outcomes. For example, the student introduced above might work hard to pass her exams in the knowledge that it will open up interesting career opportunities. These states are often attained when accumulated resources reach, or fall below, a critical threshold. Behavioural ecologists have argued that animals' risky foraging behaviour adapts to satisfy a "budget rule" that seeks to maintain energetic resources at aspirational levels that safely offset future scarcity. For example, birds make risky foraging choices at dusk in order to accrue sufficient energy to survive a cold night (Caraco, 1981). This view is neatly accommodated within the framework proposed here, in that the aspiration level reflects the setpoint against which current resource levels are compared, and the driver of behaviour is the disparity between current state and goal.

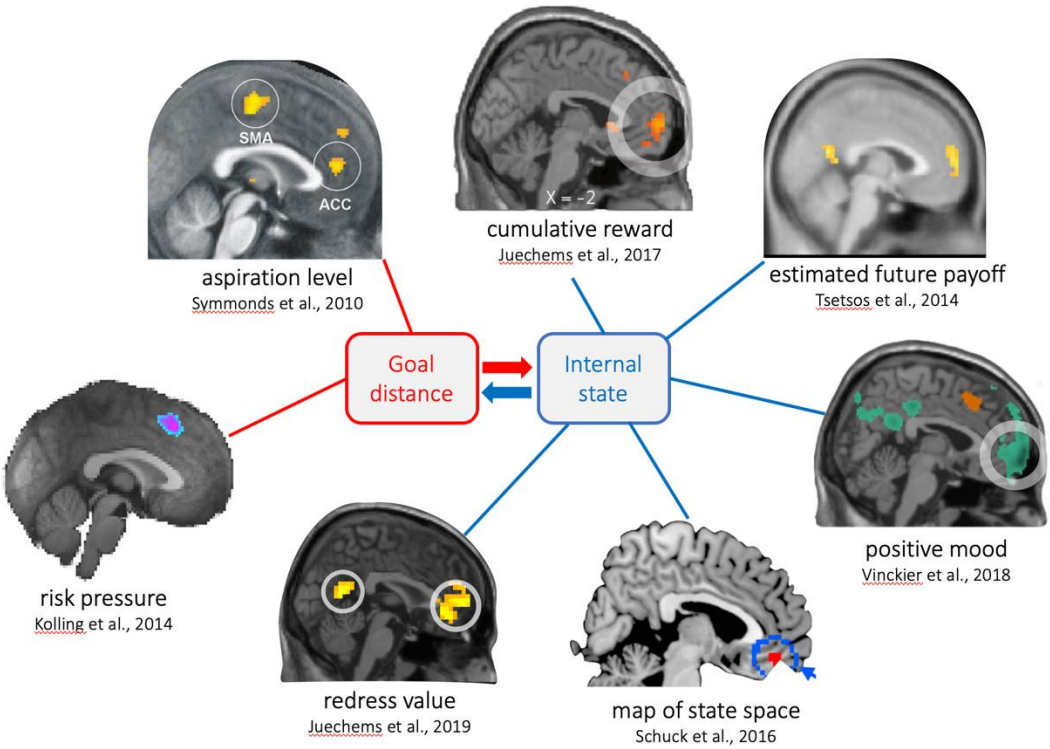


Figure 3. Cortical areas on the medial surface correlating with diverse aspects of internal state and goal distance. Reprinted with permission from refs (Juechems et al., 2019, 2017; Kolling et al., 2014; Schuck et al., 2016; Symmonds et al., 2010; Tsetsos et al., 2014).

Recent work with brain imaging (Kolling et al., 2014; Symmonds et al., 2010) suggests that this goal distance signal might be computed in a different medial prefrontal cortex region, the dorsal anterior cingulate cortex (dACC). Symmonds et al. (2010) asked participants to make a series of gambles which only paid out if a critical threshold or aspiration point was reached. BOLD signals in the vmPFC scaled with the expected future return over a series of gambles, similar to (Tsetsos et al., 2014), whereas dACC coded for the aspiration point itself (Fig. 3). Using a very similar design, Kolling et al. (2014) asked human participants to accept or reject monetary gambles over a short block, with cumulative earnings increased by a multiplier if they reached a fixed aspiration level. BOLD signals in the dACC coded

for number of steps to goal, as well as the discrepancy between accumulated return and the aspiration level, scaled by the distance to the end of the block. This signal, which the authors call "risk pressure", seems to be signalling the agent's current position with respect to a goal, rather than merely the arrival of a punctate reward. In another study involving planning in a virtual subway environment, dACC BOLD signals coded for the distance to goal in number of context switches (Balaguer et al., 2016). Indeed, the dACC is one brain region where single-cell responses gradually build up in association with a series of actions that precede reward delivery (Shidara and Richmond, 2002), and where lesions provoke failures of task persistence (Kennerley et al., 2006), as if this region participated more generally computing the disparity signal $\theta^* - \theta$ in the service of goal-directed behaviour.

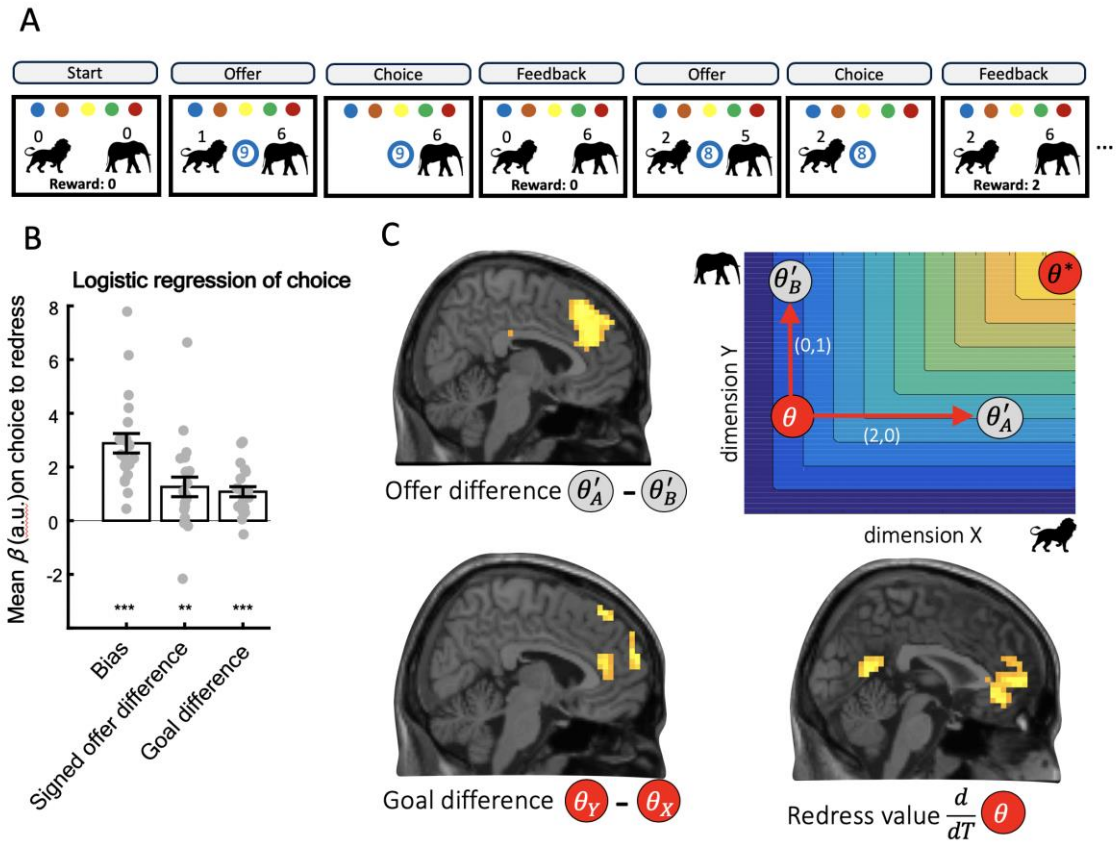


Figure 4. Neural signals for computing goal equilibrium. A. Experimental task used in Chapter Four. Participants managed virtual zoos in which they needed to accumulate as many lions and elephants as possible. Starting out from an empty zoo, participants were offered varying amounts of lions and elephants to acquire. Importantly, reward was calculated as the lower tally of animals currently owned within the zoo (e.g. when participants owned 2 lions and 6 elephants, they earned 2 units of reward). This enforced a rule by which the two tallies should be held approximately in equilibrium across time. Participants completed five zoos per block with their respective length indicated as a countdown in the centre of the screen. B. Logistic regression predicting a "redress" choice – i.e. whether participants chose to increase the lower tally of animals currently in the zoo (as given by the feedback screen). Participants placed approximately equal weight on the offer and goal (tally) differences, as in the simulation in Fig. 2c. C. Indifference map over the two assets (lions and elephants) and three key neural correlates. Regions in the dACC and vmPFC encoded the offer difference, goal difference, and the redress (the decrease in distance to goal between trials).

The multidimensional nature of value

In the standard RL framework, each state-action pair is associated with a scalar value, often called a Q-value (Sutton and Barto, 2018). In other words, most RL models assume that value is a unidimensional quantity (although "multi-objective" RL models relax this assumption (Roijers et al., 2013)). This seems incongruous when considering the challenges that animals face in natural environments, which involve jointly satisfying many different constraints (for example, maintaining levels of satiety, hydration, warmth and social capital). Rewards take many different forms – from food items to social and sexual signals to secondary reinforcers such as money. A popular view in the hybrid field of neuroeconomics is that neural signals allow potentially incommensurable rewards to be encoded in a single neural "common currency" (Kable and Glimcher, 2009). For example, single cells in OFC code for the "offer value" of different liquid rewards irrespective of their taste (Padoa-Schioppa and Assad, 2006). The existence of a common neural code for

different assets allows economists to understand rational decision-making in consumer settings, which might involve choosing between a meal in a restaurant or tickets to the theatre. However, HRL suggests a different perspective, whereby assets are explicitly encoded along different dimensions (needs) in the value map, and decisions are made according to whichever need is currently greatest.

Without further assumptions, HRL predicts that animals will treat different assets as if they are imperfect substitutions for one another. Multidimensional prospects (or "bundles") should be preferred if they contain a mixture of different assets, because this allows multiple drives to be satisfied simultaneously (Fig. 2b). For example, it is better to be neither too hungry nor too thirsty than to be starving or dehydrated. Several longstanding empirical results support this view. First of all, in early conditioning experiments it was observed that a cue that is associated either simultaneously or successively with multiple assets – such as food and water - has greater value than an equivalent cue associated with only a single asset (or drive) (Wire and Barrientos, 1958). Finally, a preference for compromise decisions is also apparent from studies of so-called "decoy" effects in multi-attribute, multi-alternative choices. Faced with a choice between two iso-preferred items that have complementary costs and benefits (e.g. an expensive but high quality consumer product A, or a cheaper but lower quality product B, the introduction of a "decoy" that is higher in price/quality than A will bias preferences towards A, whereas an item that is lower in price/quality than B will bias choices towards B, as if participants had an intrinsic preference for the compromise solution (Simonson, 2002). This preference is predicted by a need to keep goals in equilibrium, by

assuming that agents aspire to high quality and low price simultaneously, rendering the agent's indifference curve over options convex to the origin (Fig. 2b).

In natural environments we are faced with choices between offers of different quantity and quality, in variable states of need. In such scenarios, our proposal predicts that both goal difference (the relative level of two internal resources A and B) and offer difference (the relative quality of offer of A and B) will predict choices, Fig. 2c. This prediction was tested in a recent study in which participants made decisions that directly trade off multiple assets. In Chapter Four, we asked participants to perform a task that involved managing a virtual zoo containing lions and elephants. On each trial they populated the zoo by choosing among offers of variable numbers of each animal, and the financial success of their zoo was proportional to the minimum number of either lions or elephants present on each trial (which was not directly observed), Fig. 4a. Each choice was classified according to whether it "redressed" the imbalance in cumulative lions and elephants and decisions to "redress" were predicted by both goal difference and offer difference, Fig. 4b. Critically, BOLD signals in the vmPFC signalled the level of redress incurred by a choice – the extent to which a choice brought the numbers of animals into equilibrium, but independent of the total animal numbers and thus of any overall financial return. Interestingly, the dACC coded for the level of imbalance among assets, i.e. the pressure to redress this imbalance, reminiscent of the "risk pressure" signal described above (Kolling et al., 2014), Fig. 4c.

Where do goals come from?

Our focus here has been on "where value comes from". Drawing upon homeostatic theories of motivation, we argue that the value of actions taken in cognitive, social and economic settings depends at least in part on the extent to which they drive an agent towards – or away from – a setpoint. However, this begs another, potentially more thorny question: how are the setpoints set in the first place? In those use cases to which HRL has been applied thus far, pertaining to physiological needs, this is relatively uncontroversial, as it is safe to assume that setpoints for satiation and hydration are innate. What, then, of our desire to climb a mountain or become a lawyer?

In the RL framework, goals are chosen because of their potential reward value. But if reward is a reflection of the distance to a goal, then in our hands this argument becomes circular. One possibility is that goal-setting involves a hierarchical means-end reasoning process, by which proximal needs beget simple goals, the achievement of which requires more complex goals, and so on in a hierarchy of abstraction. Another factor that should not be overlooked is the role of social influence in goal setting. We often seek to achieve goals because they were suggested by our friends and associates, or by commercial advertising. However, we have to admit that we do not know how goals are computed. We think it is likely that the process by which we set ourselves goals is at least partly arbitrary, and otherwise intricately linked with the most intricate and ephemeral aspects of human cognition that pertain to our identity and sense of self.

Concluding Remarks

Our contribution here is twofold. Firstly, we highlight some challenges for RL as a computational theory of motivated behaviour in biological agents. The RL framework as developed within machine learning assumes that rewards originate in the environment, that they are sensed by dedicated input channels, that they are not fungible, and (typically) unidimensional. Secondly, we suggest that an alternative framework (Hull, 1943; Keramati and Gutkin, 2014; Klinger, 1975; O'Reilly et al., 2014) that treats reward as the by-product of computing distance to largely self-defined (*intrinsic*) goals, may have rich explanatory power for understanding human decisions. We show how this framework readily accounts for commonly observed effects in human and animal behaviour, such as reward-maximization, reference-dependence, diminishing marginal utility, and convex indifference between goods. Empirical tests of the theory's wider predictions will require tasks (or field observations) that involve accumulation of multiple assets. Such tasks are rare in the literature concerning value-guided choices, and where multiple value dimensions inadvertently come into play, they are often even discarded as nuisance variables.

General Discussion

In this thesis, I described three neuroimaging experiments and one review targeting the question of how value is constructed in the medial prefrontal cortex. Specifically, I asked how humans monitor their internal goals by keeping track of covert resource accumulation (Chapter Three), how they arbitrate between two competing goals (Chapter Four), what this implies for the underlying neural representation of value (Chapter Five), and how we may harness homeostatic models of behaviour to explain some of these phenomena (Hull, 1943; Keramati and Gutkin, 2014), Chapter Six. I built on existing models of economic decision-making (Kahneman and Tversky, 1979; Von Neumann and Morgenstern, 1944), models of value learning (Sutton and Barto, 2018), and concepts from the study of motivated behaviour (Berridge, 2004; Keramati and Gutkin, 2014). Chapter Six already discussed the bulk of this thesis, so I will use the remaining paragraphs as an opportunity to outline a subjective selection of key questions for future research relating to this thesis:

First, how does devaluation happen online? There must exist a neural mechanism by which our diverse drives and goals interact so as to route actions towards the most promising goal. Indeed, some signals relating to primary drives are broadcast throughout cortex (Allen et al., 2019) and may thus alter activity calculating value for other, less relevant drives. However, for more complex, cognitive goals, we lack clear understanding of how this might happen. Does it rely on specific sub-cortical structures that diverse goal representations all project to? Does it happen within cortex based on, e.g. mutual inhibition (Hunt and Hayden, 2017)?

Second, how do we select at the level of goals? Humans often engage particular goals and follow them for an extend period of time, unobstructed by potentially interfering needs (O'Reilly et al., 2014). However, this is often mixed with phases of indecision in which appropriate goals are selected and weighed against each other irrespective of the exact course of action to be implemented. Thus, arguably, this process happens at a different level of abstraction and its results impose constraints on the lower levels (where goal pursuit is optimized). Thus the agent needs to continually compute both the feasibility of goals, and the order (or priority) with which goals are pursued (Ringstrom and Schrater, 2019).

Lastly, how can we combine insights from devaluation (or aberrant valuation) and goal selection to inform research into psychiatric disorders, particularly mood disorders and addiction? If, for instance, depressed individuals are relatively unimpaired in their processing of reward prediction errors, then what computational deficit better describes their behaviour (Rutledge et al., 2017)? Could a deficit in goal selection better capture some phenomena related to mood disorders?

In summary, this thesis represents a modest contribution to our understanding of flexible valuation. It highlights the importance of using tasks with several assets to gain insight into flexible valuation where utility often depends on incommensurable attributes. Finally, it emphasizes the need for theories of goal selection, a process which is arguably less well understood at present than goal pursuit (O'Reilly et al., 2014).

References

Abitbol, R., Lebreton, M., Hollard, G., Richmond, B.J., Bouret, S., Pessiglione, M., 2015. Neural Mechanisms Underlying Contextual Dependency of Subjective Values: Converging Evidence from Monkeys and Humans. J. Neurosci. 35, 2308–2320. doi:10.1523/JNEUROSCI.1878-14.2015

Abrahams, M. V, Dill, L.M., 1989. A determination of the energetic equivalence of the risk of predation. Ecology 70, 999–1007.

Albert, H., Arnold, D., Maier-Rigaud, F., 2012. Model Platonism: Neoclassical economic thought in critical light. J. Institutional Econ. 8, 295–323. doi:10.1017/S1744137412000021

Allais, M., 1953. Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. Econometrica 21, 503–546.

Allen, W.E., Chen, M.Z., Pichamoorthy, N., Tien, R.H., Pachitariu, M., Luo, L., Deisseroth, K., 2019. Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. Science (80-.). 3932, 0–10. doi:10.1126/science.aav3932

Aristotle, Crisp, R., 2000. Nicomachean Ethics. Cambridge University Press, Cambridge, UK.

Arrow, K., 1971. The theory of risk aversion, in: Essays in the Theory of Risk-Bearing. pp. 90–120.

Badre, D., 2008. Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn. Sci. 12, 193–200. doi:10.1016/j.tics.2008.02.004

Balaguer, J., Spiers, H., Hassabis, D., Summerfield, C., 2016. Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. Neuron 90, 893–903. doi:10.1016/j.neuron.2016.03.037

Balleine, B.W., Dickinson, A., 1998. Goal-Directed Instrumental Action: Contingency and Incentive Learning and Their Cortical Substrates. Neuropharmacology 37, 407–419. doi:10.1016/S0028-3908(98)00033-1

Barlow, H.B., 1961. Possible principles underlying the transformation of sensory messages, in: Rosenblith, W.A. (Ed.), Sensory Communication. MIT Press, Cambridge, pp. 217–234. doi:10.1080/15459620490885644

Barreto, A., Dabney, W., Munos, R., Hunt, J.J., Schaul, T., van Hasselt, H., Silver, D., 2016. Successor Features for Transfer in Reinforcement Learning. Adv. Neural Inf. Process. Syst.

Bartra, O., McGuire, J.T., Kable, J.W., 2013. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76, 412–427. doi:10.1016/j.neuroimage.2013.02.063

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., Palminteri, S., 2018. Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. Nat. Commun. 9. doi:10.1038/s41467-018-06781-2

Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A., Stacenhfeld, K.L., Kurth-Nelson, Z., 2018. What is a cognitive map? Organising knowledge for flexible behaviour. BioRxiv. doi:10.1101/365593

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221. doi:10.1038/nn1954

Bernacchia, A., Seo, H., Lee, D., Wang, X.J., 2011. A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. 14, 366–372. doi:10.1038/nn.2752

Bernoulli, D., 1954. Exposition of a New Theory on the Measurement of Risk. Econometrica 22, 23–36.

Berridge, K.C., 2004. Motivation concepts in behavioral neuroscience. Physiol. Behav. 81, 179–209. doi:10.1016/j.physbeh.2004.02.004

Boorman, E.D., Behrens, T.E., Rushworth, M.F., 2011. Counterfactual choice and learning in a Neural Network centered on human lateral frontopolar cortex. PLoS Biol. 9. doi:10.1371/journal.pbio.1001093

Boorman, E.D., Behrens, T.E.J., Woolrich, M.W., Rushworth, M.F.S., 2009. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron 62, 733–743. doi:10.1016/j.neuron.2009.05.014

Boorman, E.D., Rushworth, M.F., Behrens, T.E.J., 2013. Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. J. Neurosci. 33, 2242–53. doi:10.1523/JNEUROSCI.3022-12.2013

Bordley, R., LiCalzi, M., 2000. Decision analysis using targets instead of utility functions. Decis. Econ. Financ. 23, 53–74.

Botvinick, M., Ritter, S., Wang, J.X., Kurth-Nelson, Z., Blundell, C., Hassabis, D., 2019. Reinforcement Learning, Fast and Slow. Trends Cogn. Sci. 23, 408–422. doi:10.1016/j.tics.2019.02.006

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D., 2001. Conflict Monitoring and Cognitive Control. Psychol. Rev. 108, 624–652.

Botvinick, M.M., Niv, Y., Barto, A.C., 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113, 262–280. doi:10.1016/j.cognition.2008.08.011

Bouret, S., Richmond, B.J., 2015. Sensitivity of Locus Ceruleus Neurons to Reward Value for Goal-Directed Actions. J. Neurosci. 35, 4005–4014. doi:10.1523/JNEUROSCI.4553-14.2015

Bouret, S., Richmond, B.J., 2010. Ventromedial and Orbital Prefrontal Neurons Differentially Encode Internally and Externally Driven Motivational Values in Monkeys 30, 8591–8601. doi:10.1523/JNEUROSCI.0049-10.2010

Bowman, E.H., 1982. Risk seeking by troubled firms. Sloan Manage. Rev. 23, 33–42.

Bradfield, L.A., Dezfouli, A., Van Holstein, M., Chieng, B., Balleine, B.W., 2015. Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. Neuron 88, 1268–1280. doi:10.1016/j.neuron.2015.10.044

Bruckner, D.W., 2009. In defense of adaptive preferences. Philos. Stud. 142, 307–324. doi:10.1007/s11098-007-9188-7

Buckley, M.J., Mansouri, F.A., Hoda, H., Mahboubi, M., Browning, P.G.F., Kwok, S.C., Philips, A., Tanaka, K., 2009. Dissociable Components of Rule-Guided Behavior Depend on Distinct Medial and Prefrontal Regions. Science (80-.). 325, 52–58. doi:10.1126/science.1172377

Burke, K.A., Franz, T.M., Miller, D.N., Schoenbaum, G., 2008. The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. Nature 454, 340–344. doi:10.1038/nature06993.The

Camerer, C., Babcock, L., Loewenstein, G., Thaler, R., 1997. Labor Supply of New York City Cabdrivers: One Day at a Time. Q. J. Econ. 112, 407–441. doi:10.1162/003355397555244

Cannon, W.B., 1929. Organization for Physiological Homeostasis. Physiol. Rev. 9, 399–431. doi:10.1152/physrev.1929.9.3.399

Caraco, T., 1981. Energy budgets, risk and foraging preferences in dark-eyed juncos (Junco hyemalis). Behav. Ecol. Sociobiol. 8, 213–217. doi:10.1007/BF00299833

Carandini, M., Heeger, D.J., 2012. Normalization as a canonical neural computation. Nat. Rev. Neurosci. 13, 51–62. doi:nrn3136 [pii]10.1038/nrn3136

Carpenter, R.H.S., 2004. Homeostasis: a plea for a unified approach. AJP Adv. Physiol. Educ. 28, 180–187. doi:10.1152/advan.00012.2004

Charnov, E.L., 1976. Optimal foraging theory: the marginal value theorem. Theor. Popul. Biol. 9, 129–136.

Chau, B.K.H., Kolling, N., Hunt, L.T., Walton, M.E., Rushworth, M.F.S., 2014. A neural mechanism underlying failure of optimal choice with multiple alternatives. Nat. Neurosci. 17, 463–70. doi:10.1038/nn.3649

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., De Gardelle, V., Herce Castañón, S., Summerfield, C., 2014. Adaptive gain control during human perceptual choice. Neuron 81, 1429–1441. doi:10.1016/j.neuron.2014.01.020

Chib, V.S., Rangel, A., Shimojo, S., O'Doherty, J.P., 2009. Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. J. Neurosci. 29, 12315–12320. doi:10.1523/JNEUROSCI.2575-09.2009

Chumbley, J.R., Friston, K.J., 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage 44, 62–70. doi:10.1016/j.neuroimage.2008.05.021

Cone, J.J., Fortin, S.M., McHenry, J.A., Stuber, G.D., McCutcheon, J.E., Roitman, M.F., 2016. Physiological state gates acquisition and expression of mesolimbic reward prediction signals. Proc. Natl. Acad. Sci. 113, 1943–1948. doi:10.1073/pnas.1519643113

Cox, X.K.M., Kable, J.W., 2014. BOLD Subjective Value Signals Exhibit Robust Range Adaptation. J. Neurosci. 34, 16533–16543. doi:10.1523/JNEUROSCI.3927-14.2014

Davis, T., Love, B.C., Preston, A.R., 2012. Striatal and Hippocampal Entropy and Recognition Signals in Category Learning: Simultaneous Processes Revealed by Model-Based fMRI. J. Exp. Psychol. Learn. Mem. Cogn. 38, 821–839. doi:10.1038/mp.2011.182.doi

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215. doi:10.1016/j.neuron.2011.02.027

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8, 1704–11. doi:10.1038/nn1560

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

Dayan, P., 2009. Goal-directed control and its antipodes. Neural Networks 22, 213–219. doi:10.1016/j.neunet.2009.03.004

Dayan, P., Daw, N.D., 2008. Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. 8, 429–453. doi:10.3758/CABN.8.4.429

De Martino, B., Fleming, S.M., Garrett, N., Dolan, R.J., 2013. Confidence in value-based choice. Nat. Neurosci. 16, 105–10. doi:10.1038/nn.3279

De Martino, B., Kumaran, D., Holt, B., Dolan, R.J., 2009. The neurobiology of reference-dependent value computation. J. Neurosci. 29, 3833–3842. doi:10.1523/JNEUROSCI.4832-08.2009

De Martino, B., Kumaran, D., Seymour, B., Dolan, R.J., 2006. Frames, biases, and rational decision making in the human brain. Science (80-.). 313, 684–687.

Dener, E., Kacelnik, A., 2016. Pea Plants Show Risk Sensitivity. Curr. Biol. 1–5. doi:10.1016/j.cub.2016.05.008

Doebeli, M., Ispolatov, Y., Simon, B., 2017. Towards a mechanistic foundation of evolutionary theory. Elife 6, 1–17. doi:10.7554/eLife.23804

Doeller, C.F., Barry, C., Burgess, N., 2010. Evidence for grid cells in a human memory network. Nature 463, 657–61. doi:10.1038/nature08704

Dolan, R.J., Dayan, P., 2013. Goals and Habits in the Brain. Neuron 80, 312–325. doi:10.1016/j.neuron.2013.09.007

Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., Daw, N.D., 2015. Model-based choices involve prospective neural activity. Nat. Neurosci. 18, 767–772. doi:10.1038/nn.3981

Doll, B.B., Simon, D.A., Daw, N.D., 2012. The ubiquity of model-based reinforcement learning. Curr. Opin. Neurobiol. 22, 1075–1081. doi:10.1016/j.conb.2012.08.003

Donoso, M., Collins, A.G.E., Koechlin, E., 2014. Human cognition. Foundations of human reasoning in the prefrontal cortex. Science (80-.). 344, 1481–6. doi:10.1126/science.1252254

Ebitz, R.B., Hayden, B.Y., 2016. Dorsal anterior cingulate: a Rorschach test for cognitive neuroscience. Nat. Neurosci. 19, 1278–1279. doi:10.1038/nn.4387

Edwards, W., 1962. Subjective probabilities inferred from decisions. Psychol. Rev. 69, 109–135. doi:10.1037/h0038674

Eklund, A., Knutsson, H., Nichols, T.E., 2018. Cluster Failure Revisited: Impact of First Level Design and Data Quality on Cluster False Positive Rates. bioRxiv. doi:10.1101/296798

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. 113, 7900–7905. doi:10.1073/pnas.1602413113

Eldar, E., Rutledge, R.B., Dolan, R.J., Niv, Y., 2016. Mood as Representation of Momentum. Trends Cogn. Sci. 20, 15–24. doi:10.1016/j.tics.2015.07.010

Ellsberg, D., 1961. Risk, Ambiguity, and the Savage Axioms. Q. J. Econ. 75, 643–669.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H.J., Ornelas, S., Koay, S.A., Thiberge, S., Daw, N.D., Tank, D.W., Witten, I.B., 2018. Specialized and spatially organized coding of sensory, motor, and cognitive variables in midbrain dopamine neurons. bioRxiv 456194. doi:10.1101/456194

Eshel, N., Tian, J., Bukwich, M., Uchida, N., 2016. Dopamine neurons share common response function for reward prediction error 19. doi:10.1038/nn.4239

Etzioni, A., 1986. the Case for a Multiple-Utility Conception. Econ. Philos. 159–183.

Fadel, J., Deutch, A.Y., 2002. Anatomical substrates of orexin-dopamine interactions: Lateral hypothalamic projections to the ventral tegmental area. Neuroscience 111, 379–387.

doi:10.1016/S0306-4522(02)00017-9

Fellows, L.K., 2006. Deciding how to decide: ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. Brain 129, 944–952. doi:10.1093/brain/awl017

Fellows, L.K., Farah, M.J., 2007. The role of ventromedial prefrontal cortex in decision making: Judgment under uncertainty or judgment per se? Cereb. Cortex 17, 2669–2674. doi:10.1093/cercor/bhl176

Flesch, T., Balaguer, J., Dekker, R., Nili, H., Summerfield, C., 2018. Comparing continual task learning in minds and machines. Proc. Natl. Acad. Sci. 115, 10313–10322. doi:10.1073/pnas.1800755115

Fouragnan, E.F., Chau, B.K.H., Folloni, D., Kolling, N., Verhagen, L., Klein-Flügge, M., Tankelevitch, L., Papageorgiou, G.K., Aubry, J.F., Sallet, J., Rushworth, M.F.S., 2019. The macaque anterior cingulate cortex translates counterfactual choice value into actual behavioral change. Nat. Neurosci. 22. doi:10.1038/s41593-019-0375-6

Friedman, M., Savage, L.J., 1948. The Utility Analysis of Choices Involving Risk. J. Polit. Econ. 56, 279–304.

Friston, K., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127–138. doi:10.1038/nrn2787

Frydman, C., Barberis, N., Camerer, C., Bossaerts, P., Rangel, A., 2014. Using neural data to test a theory of investor behavior: An application to realization utility. J. Finance 69, 907–946. doi:10.1111/jofi.12126

Gardner, M.P.. H., Conroy, J.C., Styer, C. V., Huynh, T., Whitaker, L.R., Schoenbaum, G., 2018. Medial orbitofrontal inactivation does not affect economic choice. Elife 7, 1–22.

Gehring, W.J., Willoughby, A.R., 2002. The medial frontal cortex and the rapid processing of monetary gains and losses. Science (80-.). 295, 2279–2282. doi:10.1126/science.1066893

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15, 870–878. doi:10.1006/nimg.2001.1037

Gershman, S.J., 2019. What does the free energy principle tell us about the brain ? 1–11.

Gershman, S.J., 2018. The successor representation: its computational logic and neural substrates. bioRxiv 38, 1–20. doi:10.1523/JNEUROSCI.0151-18.2018

Gershman, S.J., Malmaud, J., Tenenbaum, J.B., 2017. Structured Representations of Utility in Combinatorial Domains. Decision 4, 1–43.

Gläscher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., Paul, L.K., Tranel, D., 2012. Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. Proc. Natl. Acad. Sci. 109, 14681–14686. doi:10.1073/pnas.1206608109

Glimcher, P.W., 2004. Decision, Uncertainty and the Brain: The Science of Neuroeconomics. MIT Press.

Gluth, S., Spektor, M.S., Rieskamp, J., 2018. Value-based attentional capture affects multi-alternative decision making. Elife 7, 1–36. doi:10.7554/eLife.39659

Gonzalez, R., Wu, G., 1999. On the Shape of the Probability Weighting Function. Cogn. Psychol. 38, 129–166.

Gottfried, J.A., O'Doherty, J., Dolan, R.J., 2003. Value in Human Amygdala and Orbitofrontal Cortex. Science (80-.). 301, 1104–1107. doi:10.1126/science.1087919

Gottlieb, J., Oudeyer, P.Y., 2018. Towards a neuroscience of active sampling and curiosity. Nat. Rev. Neurosci. 19, 758–770. doi:10.1038/s41583-018-0078-0

Guitart-Masip, M., Duzel, E., Dolan, R., Dayan, P., 2014. Action versus valence in decision making. Trends Cogn. Sci. 18, 194–202. doi:10.1016/j.tics.2014.01.003

Guo, X., Singh, S., Lewis, R., Lee, H., 2015. Deep Learning for Reward Design to Improve Monte Carlo Tree Search in ATARI Games. arXiv.

Hall-McMaster, S., Luyckx, F., 2019. Revisiting foraging approaches in neuroscience. Cogn. Affect. Behav. Neurosci. 19, 225–230.

Hare, T.A., Camerer, C.F., Rangel, A., 2009. Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. Science (80-.). 324, 646–648.

Hare, T.A., Malmaud, J., Rangel, A., 2011. Focusing Attention on the Health Aspects of Foods Changes Value Signals in vmPFC and Improves Dietary Choice. J. Neurosci. 31, 11077–11087. doi:10.1523/JNEUROSCI.6383-10.2011

Harlow, H.F., 1949. The formation of learning sets. Psychol. Rev. 56, 51–65.

Haslett, D.W., 1990. What is Utility? Econ. Philos. 6, 65–94.

Hayden, B.Y., Pearson, J.M., Platt, M.L., 2011. Neuronal basis of sequential foraging decisions in a patchy environment. Nat. Neurosci. 14, 933–939. doi:10.1038/nn.2856

Hayden, B.Y., Pearson, J.M., Platt, M.L., 2009. Fictive Reward Signals in the Anterior Cingulate Cortex. Science (80-.). 324, 948–950. doi:10.1126/science.1168488

Hayden, B.Y., Platt, M.L., 2009. Gambling for Gatorade: Risk-sensitive decision making for fluid rewards in humans. Anim. Cogn. 12, 201–207. doi:10.1007/s10071-008-0186-8

Heeger, D.J., 2017. Theory of cortical function. Proc. Natl. Acad. Sci. 114, 1773–1782. doi:10.1073/pnas.1619788114

Heeger, D.J., 1992. Normalization of cell responses in cat striate cortex. Vis. Neurosci. 9, 181–197. doi:10.1017/S0952523800009640

Hirshleifer, J., 1985. The Expanding Domain of Economics. Am. Econ. Rev. 75, 53–68.

Holroyd, C.B., Yeung, N., 2012. Motivation of extended behaviors by anterior cingulate cortex. Trends Cogn. Sci. 16, 122–128. doi:10.1016/j.tics.2011.12.008

Hong, H., Kubik, J.D., Stein, J.C., 2004. Social Interaction and Stock-Market Participation. J. Finance 59, 137–163. doi:10.1111/j.1540-6261.2004.00629.x

Houston, A.I., Fawcett, T.W., Mallpress, D.E.W., McNamara, J.M., 2014. Clarifying the relationship between prospect theory and risk-sensitive foraging theory. Evol. Hum. Behav. 35, 502–507. doi:10.1016/j.evolhumbehav.2014.06.010

Houston, A.I., McNamara, J.M., Steer, M.D., 2007. Do we expect natural selection to produce rational behaviour? Philos. Trans. R. Soc. Lond. B. Biol. Sci. 362, 1531–1543. doi:10.1098/rstb.2007.2051

Howard, J.D., Kahnt, T., 2017. Identity-specific reward representations in orbitofrontal cortex are modulated by selective devaluation. J. Neurosci. 37, 3473–16. doi:10.1523/jneurosci.3473-16.2017

Howard, L.R., Javadi, A.H., Yu, Y., Mill, R.D., Morrison, L.C., Knight, R., Loftus, M.M.,

Staskute, L., Spiers, H.J., 2014. The Hippocampus and Entorhinal Cortex Encode the Path and Euclidean Distances to Goals during Navigation. Curr. Biol. 24, 1331–1340. doi:10.1016/j.cub.2014.05.001

Hull, C.L., 1943. Principles of behavior: an introduction to behavior theory. Appleton-Century-Crofts, New York.

Hunt, L.T., Hayden, B.Y., 2017. A distributed, hierarchical and recurrent framework for reward-based choice. Nat. Rev. Neurosci. 18, 172–182. doi:10.1038/nrn.2017.7

Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F.S., Behrens, T.E.J., 2012. Mechanisms underlying cortical activity during value-guided choice. Nat. Neurosci. 15, 470–476. doi:10.1038/nn.3017

Hurly, T.A., 1992. Energetic reserves of marsh tits (Parus palustris): food and fat storage in response to variable food supply. Behav. Ecol. 3, 181–188.

Izquierdo, A., Suda, R.K., Murray, E.A., 2004. Bilateral Orbital Prefrontal Cortex Lesions in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward Contingency. J. Neurosci. 24, 7540–7548. doi:10.1523/JNEUROSCI.1921-04.2004

Jevons, W.S., 1866. Brief Account of a General Mathematical Theory of Political Economy. J. R. Stat. Soc. XXIX, 282–287.

Juechems, K., Balaguer, J., Herce Castañón, S., Ruz, M., O'Reilly, J.X., Summerfield, C., 2019. A Network for Computing Value Equilibrium in the Human Medial Prefrontal Cortex. Neuron 101, 1–11. doi:10.1016/J.NEURON.2018.12.029

Juechems, K., Balaguer, J., Ruz, M., Summerfield, C., 2017. Ventromedial Prefrontal Cortex Encodes a Latent Estimate of Cumulative Reward. Neuron 93, 705-714.e4. doi:10.1016/j.neuron.2016.12.038

Kable, J.W., Glimcher, P.W., 2009. The Neurobiology of Decision: Consensus and Controversy. Neuron 63, 733–745. doi:10.1016/j.neuron.2009.09.003

Kacelnik, A., Bateson, M., 1997. Risk-sensitivity: crossroads for theories of decision-making. Trends Cogn. Sci. 1, 304–309.

Kacelnik, A., El Mouden, C., 2013. Triumphs and trials of the risk paradigm. Anim. Behav. 86, 1117–1129. doi:10.1016/j.anbehav.2013.09.034

Kahneman, D., Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263–292. doi:10.2307/1914185

Kennerley, S.W., Walton, M.E., Behrens, T.E.J., Buckley, M.J., Rushworth, M.F.S., 2006. Optimal decision making and the anterior cingulate cortex. Nat. Neurosci. 9, 940–947. doi:10.1038/nn1724

Keramati, M., Ahmed, S.H., Gutkin, B.S., 2017. Misdeed of the need: towards computational accounts of transition to addiction. Curr. Opin. Neurobiol. 46, 142–153. doi:10.1016/j.conb.2017.08.014

Keramati, M., Gutkin, B., 2014. Homeostatic reinforcement learning for integrating reward collection and physiological stability. Elife 3, 1–26. doi:10.7554/eLife.04811

Keramati, M., Smittenaar, P., Dolan, R.J., Dayan, P., 2016. Adaptive integration of habits into depth-limited planning defines a habitual-goal – directed spectrum. Proc. Natl. Acad. Sci. 113, 12868–12873. doi:10.1073/pnas.1609094113

Kim, H., Shimojo, S., O'Doherty, J.P., 2011. Overlapping Responses for the Expectation of Juice and Money Rewards in Human Ventromedial Prefrontal Cortex. Cereb. Cortex 21,

769–776. doi:10.1093/cercor/bhq145

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1982. Optimization by simulated annealing. Science (80-.). 220, 671–680.

Klein, T.A., Ullsperger, M., Jocham, G., 2017. Learning relative values in the striatum induces violations of normative decision making. Nat. Commun. 8, 1–12. doi:10.1038/ncomms16033

Kliger, D., Tsur, I., 2011. Prospect Theory and Risk-Seeking Behavior by Troubled Firms. J. Behav. Financ. 12, 29–40. doi:10.1080/15427560.2011.555028

Klinger, E., 1975. Consequences of Commitment to and Disengagement from Incentives. Psychol. Rev. 82, 1–25. doi:10.1037/h0021465

Knox, R.E., Inkster, J.A., 1968. Postdecision dissonance at post time. J. Pers. Soc. Psychol. 8, 319–323. doi:10.1037/h0025528

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., Glover, G., 2005. Distributed neural representation of expected value. J. Neurosci. 25, 4806–4812. doi:10.1523/JNEUROSCI.0642-05.2005

Koechlin, E., 2016. Prefrontal executive function and adaptive behavior in complex environments. Curr. Opin. Neurobiol. 37, 1–6. doi:10.1016/j.conb.2015.11.004

Koechlin, E., Hyafil, A., 2007. Anterior Prefrontal Function and the Limits of Human Decision-Making. Science (80-.). 318, 594–598.

Koechlin, E., Ody, C., Kouneiher, F., 2003. The architecture of cognitive control in the human prefrontal cortex. Science (80-.). 302, 1181–1185. doi:10.1126/science.1088545

Koechlin, E., Summerfield, C., 2007. An information theoretical approach to prefrontal executive function. Trends Cogn. Sci. 11, 229–235. doi:10.1016/j.tics.2007.04.005

Kolling, N., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., 2012. Neural Mechanisms of Foraging. Science (80-.). 336, 95–98. doi:10.1126/science.1216930

Kolling, N., Scholl, J., Chekroud, A., Trier, H.A., Rushworth, M.F.S., 2018. Prospection, Perseverance, and Insight in Sequential Behavior. Neuron 99, 1069-1082.e7. doi:10.1016/j.neuron.2018.08.018

Kolling, N., Wittmann, M., Rushworth, M.F.S., 2014. Multiple neural mechanisms of decision making and their competition under changing risk pressure. Neuron 81, 1190–1202. doi:10.1016/j.neuron.2014.01.033

Kolling, N., Wittmann, M.K., Behrens, T.E.J., Boorman, E.D., Mars, R.B., Rushworth, M.F.S., 2016. Value, search, persistence and model updating in anterior cingulate cortex. Nat. Neurosci. 19, 1280–1285. doi:10.1038/nn.4382

Korn, C.W., Bach, D.R., 2015. Maintaining Homeostasis by Decision-Making. PLoS Comput. Biol. 11, 1–19. doi:10.1371/journal.pcbi.1004301

Kőszegi, B., Rabin, M., 2007. Reference-Dependent Risk Attitudes. Am. Econ. Rev. 97, 1047–73.

Kőszegi, B., Rabin, M., 2006. A Model of Reference-Dependent Preferences. Q. J. Econ. 121, 1133–1165.

Krakauer, J.W., Ghazanfar, A.A., Gomez-marin, A., MacIver, M.A., Poeppel, D., 2017. Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron 93, 480–490. doi:10.1016/j.neuron.2016.12.041

Krubitzer, L., 1995. The organization of neocortex in mammals: are species differences really so different? Trends Neurosci. 18, 408–417. doi:10.1016/0166-2236(95)93938-T

Laughlin, S., 1981. A simple coding procedure enhances a neuron's information capacity. Zeitschrift fur Naturforsch. C 36, 910–912. doi:10.1515/znc-1981-9-1040

Lebreton, M., Abitbol, R., Daunizeau, J., Pessiglione, M., 2015. Automatic integration of confidence in the brain valuation signal. Nat. Neurosci. 18, 1159–1167. doi:10.1038/nn.4064

Lebreton, M., Jorge, S., Michel, V., Thirion, B., Pessiglione, M., 2009. An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. Neuron 64, 431–439. doi:10.1016/j.neuron.2009.09.040

Levy, D.J., Glimcher, P.W., 2012. The root of all value: a neural common currency for choice. Curr. Opin. Neurobiol. 22, 1027–1038. doi:10.1016/j.conb.2012.06.001

Levy, D.J., Glimcher, P.W., 2011. Comparing Apples and Oranges: Using Reward-Specific and Reward-General Subjective Value Representation in the Brain. J. Neurosci. 31, 14693–14707. doi:10.1523/JNEUROSCI.2218-11.2011

Li, V., Michael, E., Balaguer, J., Herce Castañón, S., Summerfield, C., 2018. Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making. Proc. Natl. Acad. Sci. 201805224. doi:10.1073/pnas.1805224115

Lim, S.-L., O'Doherty, J.P., Rangel, a., 2011. The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. J. Neurosci. 31, 13214–13223. doi:10.1523/JNEUROSCI.1246-11.2011

Lin, A., Adolphs, R., Rangel, A., 2012. Social and monetary reward learning engage overlapping neural substrates. Soc. Cogn. Affect. Neurosci. 7, 274–281. doi:10.1093/scan/nsr006

Liu, Z., Zhou, J., Li, Y., Hu, F., Lu, Y., Ma, M., Feng, Q., Zhang, J., Wang, D., Zeng, J., Bao, J., Kim, J., Chen, Z., El Mestikawy, S., Luo, M., 2014. Dorsal Raphe Neurons Signal Reward through 5-HT and Glutamate. Neuron 81, 1360–1374. doi:10.1016/j.neuron.2014.02.010

Lopes, L., Oden, G., 1999. The Role of Aspiration Level in Risky Choice: A Comparison of Cumulative Prospect Theory and SP/A Theory. J. Math. Psychol. 43, 286–313. doi:10.1006/jmps.1999.1259

Lopes, L.L., 1994. Psychology and Economics: Perspectives on Risk, Cooperation and the Marketplace. Annu. Rev. Psychol. 45, 197–227.

Lopez-Persem, A., Domenech, P., Pessiglione, M., 2016. How prior preferences determine decision-making frames and biases in the human brain. Elife 5, 1–20. doi:10.7554/eLife.20317

Losecaat Vermeer, A.B., Boksem, M.A.S., Sanfey, A.G., 2014. Neural mechanisms underlying context-dependent shifts in risk preferences. Neuroimage 103, 355–363. doi:10.1016/j.neuroimage.2014.09.054

Louie, K., Glimcher, P.W., 2012. Efficient coding and the neural representation of value. Ann. N. Y. Acad. Sci. 1251, 13–32. doi:10.1111/j.1749-6632.2012.06496.x

Louie, K., Grattan, L.E., Glimcher, P.W., 2011. Reward Value-Based Gain Control: Divisive Normalization in Parietal Cortex. J. Neurosci. 31, 10627–10639. doi:10.1523/JNEUROSCI.1237-11.2011

Louie, K., Khaw, M.W., Glimcher, P.W., 2013. Normalization is a general neural mechanism

for context-dependent decision making. Proc. Natl. Acad. Sci. U. S. A. 110, 6139–44. doi:10.1073/pnas.1217854110

Luyckx, F., Nili, H., Spitzer, B., Summerfield, C., 2018. Neural structure mapping in human probabilistic reward learning 1–16. doi:10.1101/366757

Machens, C.K., Romo, R., Brody, C.D., 2005. Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. Science (80-.). 307, 1121–1124. doi:10.1126/science.1104171

Mallpress, D.E.W., Fawcett, T.W., Houston, A.I., Mcnamara, J.M., 2015. Risk Attitudes in a Changing Environment: An Evolutionary Model of the Fourfold Pattern of Risk Preferences 122, 364–375.

Mansouri, F.A., Koechlin, E., Rosa, M.G.P.P., Buckley, M.J., 2017. Managing competing goals - A key role for the frontopolar cortex. Nat. Rev. Neurosci. 18, 645–657. doi:10.1038/nrn.2017.111

Mante, V., Sussillo, D., Shenoy, K. V., Newsome, W.T., 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84. doi:10.1038/nature12742

Markowitz, H., 1952. The Utility of Wealth. J. Polit. Econ. 60, 151–158.

Mas-Colell, A., Whinston, M.D., Green, J.R., 1995. Microeconomic Theory. Oxford University Press, New York.

Maslow, A.H., 1943. A Theory of Human Motivation. Psychol. Rev. 370–396.

Mattar, M.G., Daw, N.D., 2018. Prioritized memory access explains planning and hippocampal replay. Nat. Neurosci. 21, 1609–1617. doi:10.1038/s41593-018-0232-z

McDermott, R., Fowler, J.H., Smirnov, O., 2008. On the Evolutionary Origin of Prospect Theory Preferences. J. Polit. 70. doi:10.1017/S0022381608080341

McNamara, J.M., Houston, A.I., 1985. Optimal Foraging and Learning. J. Theor. Biol. 117, 231–249.

McNamee, D., Rangel, A., O'Doherty, J.P., 2013. Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. Nat. Neurosci. 16, 479–85. doi:10.1038/nn.3337

Meder, D., Kolling, N., Verhagen, L., Wittmann, M.K., Scholl, J., Madsen, K.H., Hulme, O.J., Behrens, T.E., Rushworth, M.F., 2017. Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. Nat. Commun. 1–39. doi:10.1101/195842

Miller, K.J., Botvinick, M.M., Brody, C.D., 2017. Dorsal hippocampus contributes to model-based planning. Nat. Neurosci. 20, 1269–1276. doi:10.1038/nn.4613

Miller, K.J., Shenhav, A., Ludvig, E.A., 2016. Habits without Values. BioRxiv 1–27.

Miller, N.E., Kessen, M.L., 1952. Behavioral Effects of Food Via Stomach Fistula Compared with Those of Food Via Mouth. J. Comp. Physiol. Psychol. 45, 555–564.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533. doi:10.1038/nature14236

Moffaert, K. Van, Nowé, A., Auer, P., Hutter, M., Orseau, L., 2014. Multi-Objective

Reinforcement Learning using Sets of Pareto Dominating Policies. J. Mach. Learn. Res. 15, 3483–3512.

Mollenauer, S.O., 1971. Shifts in deprivation level: Different effects depending on amount of preshift training. Learn. Motiv. 2, 58–66. doi:10.1016/0023-9690(71)90048-8

Momennejad, I., Howard, M.W., 2018. Predicting the Future with Multi-scale Successor Representations. bioRxiv 1–27. doi:10.1101/449470

Momennejad, I., Otto, A.R., Daw, N.D., Norman, K.A., 2018. Offline replay supports planning in human reinforcement learning. Elife 7, 1–25. doi:10.7554/elife.32548

Momennejad, I., Russek, E.M., Cheong, J.H., Botvinick, M.M., Daw, N.D., Gershman, S.J., 2017. The successor representation in human reinforcement learning. Nat. Hum. Behav. 1, 680–692. doi:10.1038/s41562-017-0180-8

Morales, M., Margolis, E.B., 2017. Ventral tegmental area: Cellular heterogeneity, connectivity and behaviour. Nat. Rev. Neurosci. 18, 73–85. doi:10.1038/nrn.2016.165

Mullainathan, S., Shafir, E., 2014. Scarcity: The True Cost of Not Having Enough. Penguin.

Murray, J.D., Bernacchia, A., Freedman, D.J., Romo, R., Wallis, J.D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., Wang, X.J., 2014. A hierarchy of intrinsic timescales across primate cortex. Nat. Neurosci. 17, 1661–1663. doi:10.1038/nn.3862

Neubert, F.-X., Mars, R.B., Sallet, J., Rushworth, M.F.S., 2015. Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. Proc. Natl. Acad. Sci. 112, E2695–E2704. doi:10.1073/pnas.1410767112

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A Toolbox for Representational Similarity Analysis. PLoS Comput. Biol. 10. doi:10.1371/journal.pcbi.1003553

Nogueiras, R., Tschöp, M.H., Zigman, J.M., 2008. Central nervous system regulation of energy metabolism: Ghrelin versus leptin. Ann. N. Y. Acad. Sci. 1126, 14–19. doi:10.1196/annals.1433.054

Noonan, M.P., Walton, M.E., Behrens, T.E.J., Sallet, J., Buckley, M.J., Rushworth, M.F.S., 2010. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc. Natl. Acad. Sci. 107, 20547–20552. doi:10.1073/pnas.1012246107

O'Neill, M., Schultz, W., 2010. Coding of Reward Risk by Orbitofrontal Neurons Is Mostly Distinct from Coding of Reward Value. Neuron 68, 789–800. doi:10.1016/j.neuron.2010.09.031

O'Reilly, J.X., Schüffelgen, 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. Proc. Natl. Acad. Sci. 110, E3660–E3669. doi:10.1073/pnas.1305373110

O'Reilly, R.C., Hazy, T.E., Mollick, J., Mackie, P., Herd, S., 2014. Goal-Driven Cognition in the Brain: A Computational Framework. arXiv. doi:10.1097/CCM.0000000000001615

Olds, J., Milner, P., 1954. Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain. J. Comp. Physiol. Psychol. 47, 419–427. doi:10.1037/h0058775

Öngür, D., Ferry, A.T., Price, J.L., 2003. Architectonic subdivision of the human orbital and medial prefrontal cortex. J. Comp. Neurol. 460, 425–449. doi:10.1002/cne.10609

Oudeyer, P., Kaplan, F., 2007. What is intrinsic motivation? A typology of computational approaches. Front. Neurorobot. 1, 1–14. doi:10.3389/neuro.12.006.2007

Padoa-Schioppa, C., 2013. Neuronal origins of choice variability in economic decisions. Neuron 80, 1322–1336. doi:10.1016/j.neuron.2013.09.013

Padoa-Schioppa, C., 2009. Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. J. Neurosci. 29, 14004–14014. doi:10.1523/JNEUROSCI.3751-09.2009

Padoa-Schioppa, C., Assad, J.A., 2008. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat. Neurosci. 11, 95–102. doi:10.1038/nn2020

Padoa-Schioppa, C., Assad, J.A., 2006. Neurons in the orbitofrontal cortex encode economic value. Nature 441, 223–226. doi:10.1038/nature04676

Padoa-Schioppa, C., Conen, K.E., 2017. Orbitofrontal Cortex: A Neural Circuit for Economic Decisions. Neuron 96, 736–754. doi:10.1016/j.neuron.2017.09.031

Parducci, A., 1965. Category judgment: A range-frequency model. Psychol. Rev. 72, 407–418.

Paulsen, D.J., Platt, M.L., Huettel, S.A., Brannon, E.M., 2011. Decision-making under risk in children, adolescents, and young adults. Front. Psychol. 2, 1–6. doi:10.3389/fpsyg.2011.00072

Pelletier, G., Fellows, L.K., 2018. A critical role for human ventromedial frontal lobe in value comparison based on multi-attribute configuration. bioRxiv. doi:10.1101/483719

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., Frith, C.D., 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442, 1042–5. doi:10.1038/nature05051

Pezzulo, G., Rigoli, F., Friston, K., 2015. Active Inference, homeostatic regulation and adaptive behavioural control. Prog. Neurobiol. 134, 17–35. doi:10.1016/j.pneurobio.2015.09.001

Philiastides, M.G., Biele, G., Heekeren, H.R., 2010. A mechanistic account of value computation in the human brain. Proc. Natl. Acad. Sci. U. S. A. 107, 9430–9435. doi:10.1073/pnas.1001732107

Plassmann, H., O'Doherty, J., Rangel, A., 2007. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J. Neurosci. 27, 9984–9988. doi:10.1523/JNEUROSCI.2131-07.2007

Platt, M.L., Huettel, S.A., 2008. Risky business: the neuroeconomics of decision making under uncertainty. Nat. Neurosci. 11, 398–403. doi:10.1038/nn2062

Polanía, R., Woodford, M., Ruff, C., 2019. Efficient coding of subjective value. Nat. Neurosci. 22, 134–142. doi:10.1101/358317

Polanía, R., Woodford, M., Ruff, C., 2018. Efficient coding of subjective value. Nat. Neurosci. 22, 134–142. doi:10.1101/358317

Pompilio, L., Kacelnik, A., 2005. State-dependent learning and suboptimal choice: When starlings prefer long over short delays to food. Anim. Behav. 70, 571–578. doi:10.1016/j.anbehav.2004.12.009

Pratt, J.W., 1964. Risk Aversion in the Small and in the Large. Econometrica 32, 122–136. doi:10.2307/1913738

Prelec, D., 1998. The Probability Weighting Function. Econometrica 66, 497–527.

Prelec, D., Loewenstein, G., 1991. Decision Making Over Time and Under Uncertainty: A Common Approach. Manage. Sci. 37, 770–786. doi:10.1287/mnsc.37.7.770

Puig, M.V., Rose, J., Schimdt, R., Freund, N., 2014. Dopamine modulation of learning and memory in the prefrontal cortex: insights from studies in primates, rodents, and birds. Front. Neural Circuits 8, 1–15. doi:10.3389/fncir.2014.00093

Ramsay, F.P., 1926. Truth and Probability, in: Braithwaite, R.B. (Ed.), The Foundations of Mathematics and Other Logical Essays. Kegan, Paul, Trench, Trubner & Co, London, pp. 156–198.

Rangel, A., Camerer, C., Montague, P.R., 2008. A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. 9, 545–556. doi:10.1038/nrn2357

Rangel, A., Clithero, J.A., 2012. Value normalization in decision making: Theory and evidence. Curr. Opin. Neurobiol. 22, 970–981. doi:10.1016/j.conb.2012.07.011

Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87. doi:10.1038/4580

Redish, A.D., 2016. Vicarious trial and error. Nat. Rev. Neurosci. 17, 147–159. doi:10.1038/nrn.2015.30

Ribas-Fernandes, J.J.F., Solway, A., Diuk, C., McGuire, J.T., Barto, A.G., Niv, Y., Botvinick, M.M., 2011. A Neural Signature of Hierarchical Reinforcement Learning. Neuron 71, 370–379. doi:10.1016/j.neuron.2011.05.042

Rich, E.L., Wallis, J.D., 2016. Decoding subjective decisions from orbitofrontal cortex. Nat. Neurosci. 19, 1–10. doi:10.1038/nn.4320

Rigoli, F., Friston, K.J., Martinelli, C., Selaković, M., Shergill, S.S., Dolan, R.J., 2016. A Bayesian model of context-sensitive value attribution. Elife 5, 1–25. doi:10.7554/eLife.16127

Rigoli, F., Rutledge, R.B., Dayan, P., Dolan, R.J., 2015. The influence of contextual reward statistics on risk preference. Neuroimage 11. doi:10.1016/j.neuroimage.2015.12.016

Ringstrom, T.J., Schrater, P.R., 2019. Constraint Satisfaction Propagation: Non-stationary Policy Synthesis for Temporal Logic Planning. arXiv.

Roijers, D.M., Vamplew, P., Dazeley, R., 2013. A Survey of Multi-Objective Sequential Decision-Making 48, 67–113. doi:10.1613/JAIR.3987

Roy, M., Shohamy, D., Wager, T.D., 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn. Sci. 16, 147–156. doi:10.1016/j.tics.2012.01.005

Rudorf, S., Preuschoff, K., Weber, B., 2012. Neural correlates of anticipation risk reflect risk preferences. J. Neurosci. 32, 16683–92. doi:10.1523/JNEUROSCI.4235-11.2012

Rushworth, M.F.S., Behrens, T.E.J., 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. Nat. Neurosci. 11, 389–397. doi:10.1038/nn2066

Rushworth, M.F.S., Kolling, N., Mars, R.B., Sallet, J., Mars, R.B., 2012. Valuation and decision-making in frontal cortex: one or many serial or parallel systems? Curr. Opin. Neurobiol. 22, 1–10. doi:10.1016/j.conb.2012.04.011

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., Behrens, T.E., 2011. Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron 70, 1054–1069. doi:10.1016/j.neuron.2011.05.014

Russek, E.M., Momennejad, I., Botvinick, M.M., Gershman, S.J., Daw, N.D., 2017. Predictive representations can link model-based reinforcement learning to model-free mechanisms, PLoS Computational Biology. doi:10.1371/journal.pcbi.1005768

Rustichini, A., Conen, K.E., Cai, X., Padoa-Schioppa, C., 2017. Optimal coding and neuronal adaptation in economic decisions. Nat. Commun. 8. doi:10.1038/s41467-017-01373-y

Rutledge, R.B., Dean, M., Caplin, A., Glimcher, P.W., 2010. Testing the reward prediction error hypothesis with an axiomatic model. J Neurosci 30, 13525–13536. doi:10.1523/JNEUROSCI.1747-10.2010

Rutledge, R.B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., Lam, J., Skandali, N., Siegel, J.Z., Ousdal, O.T., Prabhu, G., Dayan, P., Fonagy, P., Dolan, R.J., 2017. Association of neural and emotional impacts of reward prediction errors with major depression. JAMA Psychiatry 74, 790–797. doi:10.1001/jamapsychiatry.2017.1713

Ryan, R.M., Deci, E.L., 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemp. Educ. Psychol. 25, 54–67. doi:10.1006/ceps.1999.1020

Sadacca, B.F., Wied, H.M., Lopatina, N., Saini, G.K., Nemirovsky, D., Schoenbaum, G., 2018. Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task 1–17.

Sallet, J., Mars, R.B., Noonan, M.P., Neubert, F.-X., Jbabdi, S., O'Reilly, J.X., Filippini, N., Thomas, A.G., Rushworth, M.F., 2013. The Organization of Dorsal Frontal Cortex in Humans and Macaques. J. Neurosci. 33, 12255–12274. doi:10.1523/JNEUROSCI.5108-12.2013

Savage, L.J., 1971. Elicitation of personal probabilities and expectations. J. Am. Stat. Assoc. 66, 783–801. doi:10.1080/01621459.1971.10482346

Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., Botvinick, M.M., 2013. Neural representations of events arise from temporal community structure. Nat. Neurosci. 16, 486–92. doi:10.1038/nn.3331

Schaul, T., Horgan, D., Gregor, K., Silver, D., 2015. Universal Value Function Approximators. Int. Conf. Mach. Learn. 1–9. doi:3045118.3045258

Schoenbaum, G., Takahashi, Y., Liu, T.L., Mcdannald, M.A., 2011. Does the orbitofrontal cortex signal value? Ann. N. Y. Acad. Sci. 1239, 87–99. doi:10.1111/j.1749-6632.2011.06210.x

Schuck, N.W., Cai, M., Wilson, R.C., Niv, Y., 2016. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron 91, 1402–1412. doi:10.1016/j.neuron.2016.08.019

Schuck, N.W., Niv, Y., 2019. Sequential replay of non-spatial task states in the human hippocampus. Science (80-.). 364, 1–9. doi:10.1126/science.aaw5181

Schultz, W., 2015. Neuronal Reward and Decision Signals: From Theories to Data. Physiol. Rev. 95, 853–951. doi:10.1152/physrev.00023.2014

Schultz, W., Apicella, P., Ljungberg, T., 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J. Neurosci. 13, 900–913. doi:8441015

Schultz, W., Dayan, P., Montague, P.R., 1997. A Neural Substrate of Prediction and Reward. Science (80-.). 275, 1593–1599. doi:10.1126/science.275.5306.1593

Schultz, W., O'Neill, M., Tobler, P.N., Kobayashi, S., 2011. Neuronal signals for reward risk in frontal cortex. Ann. N. Y. Acad. Sci. 1239, 109–117. doi:10.1111/j.1749-6632.2011.06256.x

Shadlen, M.N., Shohamy, D., 2016. Perspective Decision Making and Sequential Sampling from Memory. Neuron 90, 927–939. doi:10.1016/j.neuron.2016.04.036

Sharpe, M.J., Batchelor, H.M., Mueller, L.E., Yun, C., Maes, E.J.P., Niv, Y., Schoenbaum, G., 2019a. Dopamine transients delivered in learning contexts do not act as model-free prediction errors. bioRxiv 574541. doi:10.1101/574541

Sharpe, M.J., Stalnaker, T., Schuck, N.W., Killcross, S., Schoenbaum, G., Niv, Y., 2019b. An Integrated Model of Action Selection: Distinct Modes of Cortical Control of Striatal Decision Making. Annu. Rev. Psychol. 70, 53–76. doi:10.1146/annurev-psych-010418-102824

Sharpe, M.J., Wikenheiser, A.M., Niv, Y., Schoenbaum, G., 2015. The State of the Orbitofrontal Cortex. Neuron 88, 1075–1077. doi:10.1016/j.neuron.2015.12.004

Shenhav, A., Botvinick, M., Cohen, J., 2013. The expected value of control: An integrative theory of anterior cingulate cortex function. Neuron 79, 217–240. doi:10.1016/j.neuron.2013.07.007

Shenhav, A., Cohen, J.D., Botvinick, M.M., 2016. Dorsal anterior cingulate cortex and the value of control. Nat. Neurosci. 19, 1286–1291. doi:10.1038/nn.4382

Shenhav, A., Straccia, M.A., Cohen, J.D., Botvinick, M.M., 2014. Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. Nat. Neurosci. 17, 1249–1254. doi:10.1038/nn.3771

Shidara, M., Richmond, B.J., 2002. Anterior Cingulate: Single Neuronal Signals Related to Degree of Reward Expectancy. Science (80-.). 296, 1709–1711. doi:10.1126/science.1069504

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. Nature 550, 354–359. doi:10.1038/nature24270

Simonson, I., 2002. Choice Based on Reasons: The Case of Attraction and Compromise Effects. J. Consum. Res. 16, 158. doi:10.1086/209205

Singh, S., Lewis, R.L., Barto, A.G., Sorg, J., 2010. Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. IEEE Trans. Auton. Ment. Dev. 2, 70–82. doi:10.1109/tamd.2010.2051031

Sladky, R., Friston, K.J., Tröstl, J., Cunnington, R., Moser, E., Windischberger, C., 2011. Slice-timing effects and their correction in functional MRI. Neuroimage 58, 588–594. doi:10.1016/j.neuroimage.2011.06.078

Smith, A., 1786. An inquiry into the nature and causes of the wealth of nations, 4th ed. London.

Soltani, A., de Martino, B., Camerer, C., 2012. A range-normalization model of context-dependent choice: A new model and evidence. PLoS Comput. Biol. 8. doi:10.1371/journal.pcbi.1002607

Spitzer, B., Waschke, L., Summerfield, C., 2017. Selective overweighting of larger magnitudes during noisy numerical comparison. Nat. Hum. Behav. 1.

Srivastava, N., Schrater, P., 2015. Learning What to Want: Context-Sensitive Preference Learning. PLoS One 10, 1–21. doi:10.1371/journal.pone.0141129

Stachenfeld, K.L., Botvinick, M.M., Gershman, S.J., 2014. Design Principles of the Hippocampal Cognitive Map. Adv. Neural Inf. Process. Syst. 1–9. doi:10.1016/j.ydbio.2008.02.022

Steiner, J., Stewart, C., 2016. Perceiving Prospects Properly. Am. Econ. Assoc. 106, 1601–1631.

Steinmetz, N.A., Zatka-Haas, P., Carandini, M., Harris, K.D., 2018. Distributed correlates of visually-guided behavior across the mouse brain. bioRxiv 474437. doi:10.1101/474437

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–1017. doi:10.1016/j.neuroimage.2009.03.025

Stephens, D.W., 2008. Decision ecology: foraging and the ecology of animal decision making. Cogn. Affect. Behav. Neurosci. 8, 475–484. doi:10.3758/CABN.8.4.475

Stephens, D.W., 1981. The Logic of Risk-Sensitive Foraging Preferences. Anim. Behav. 29, 628–629.

Stevens, S.S., 1960. The Psychophysics of Sensory Function. Am. Sci. 48, 226–253.

Stewart, N., Chater, N., Brown, G.D.A., 2006. Decision by sampling. Cogn. Psychol. 53, 1–26. doi:10.1016/j.cogpsych.2005.10.003

Stott, J.J., Redish, a. D., 2015. Representations of Value in the Brain: An Embarrassment of Riches? PLOS Biol. 13, e1002174. doi:10.1371/journal.pbio.1002174

Strait, C.E., Blanchard, T.C., Hayden, B.Y., 2014. Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. Neuron 82, 1357–1366. doi:10.1016/j.neuron.2014.04.032

Summerfield, C., Tsetsos, K., 2015. Do humans make good decisions? Trends Cogn. Sci. 19, 27–34. doi:10.1016/j.tics.2014.11.005.Do

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning, 2nd ed. MIT Press, Cambridge, Mass.

Sutton, R.S., Modayil, J., Delp, M., Degris, T., Pilarski, P.M., White, A., Precup, D., 2011. Horde: A Scalable Real-time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction. AAMAS 2, 761–768.

Syed, E.C.J., Grima, L.L., Magill, P.J., Bogacz, R., Brown, P., Walton, M.E., 2016. Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat. Neurosci. 19, 34–36. doi:10.1038/nn.4187

Symmonds, M., Bossaerts, P., Dolan, R.J., 2010. A behavioral and neural evaluation of prospective decision-making under risk. J. Neurosci. 30, 14380–14389. doi:10.1523/JNEUROSCI.1459-10.2010

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: statistics, structure, and abstraction. Supporting Material. Science (80-.). 331, 1279–85. doi:10.1126/science.1192788

Tesauro, G., 1994. TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. Neural Comput. 6, 215–219. doi:10.1162/neco.1994.6.2.215

Thaler, R., 2018. From Cashews to Nudges: The Evolution of Behavioral Economics. Am. Econ. Rev. 108, 1265–1287. doi:10.1257/aer.108.6.1265

Thaler, R.H., Johnson, E.J., 1990. Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. Manage. Sci. 36, 643–660. doi:10.1287/mnsc.36.6.643

Thorndike, E.L., 1898. Animal Intelligence. Psychol. Rev. 2, 1–107. doi:10.4324/9781351321044-2

Tinbergen, N., 1963. On aims and methods of Ethology. Z. Tierpsychol. 20, 410–433. doi:https://0-doi-org.pugwash.lib.warwick.ac.uk/10.1163/157075605774840941

Tobler, P.N., Fiorillo, C.D., Schultz, W., 2005. Adaptive coding of reward value by dopamine neurons. Science (80-.). 307, 1642–5. doi:10.1126/science.1105370

Tolman, E.C., 1948. Cognitive maps in rats and men. Psychol. Rev. 55, 189–208.

Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R. a, 2007. The neural basis of loss aversion in decision-making under risk. Science (80-.). 315, 515–518. doi:10.1126/science.1134239

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G.D., Bonci, A., de Lecea, L., Deisseroth, K., 2009. Phasic Firing in Dopaminergic Neurons Is Sufficient for Behavioral Conditioning. Science (80-.). 324, 1080–1084. doi:10.1126/science.1168878

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., Summerfield, C., 2016. Economic irrationality is optimal during noisy decision making. Proc. Natl. Acad. Sci. 113, 3102–3107. doi:10.1073/pnas.1519157113

Tsetsos, K., Usher, M., Chater, N., 2010. Preference reversal in multiattribute choice. Psychol. Rev. 117, 1275–1293. doi:10.1037/a0020580

Tsetsos, K., Wyart, V., Shorkey, S.P., Summerfield, C., 2014. Neural mechanisms of economic commitment in the human medial prefrontal cortex. Elife 3, e03701. doi:10.7554/eLife.03701

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. J. Risk Uncertain. 5, 297–323. doi:10.1007/BF00122574

Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science (80-.). 211, 453–458. doi:10.1126/science.7455683

Valentin, V. V., Dickinson, A., O'Doherty, J.P., 2007. Determining the Neural Substrates of Goal-Directed Learning in the Human Brain. J. Neurosci. 27, 4019–4026. doi:10.1523/JNEUROSCI.0564-07.2007

Vamplew, P., Issabekov, R., Dazeley, R., Foale, C., Berry, A., Moore, T., Creighton, D., 2017. Steering approaches to Pareto-optimal multiobjective reinforcement learning. Neurocomputing 263, 26–38. doi:10.1016/j.neucom.2016.08.152

Vinckier, F., Rigoux, L., Oudiette, D., Pessiglione, M., 2018. Neuro-computational account of how mood fluctuations arise and affect decision making. Nat. Commun. 9. doi:10.1038/s41467-018-03774-z

Vlaev, I., Chater, N., Stewart, N., Brown, G.D.A., 2011. Does the brain calculate value? Trends Cogn. Sci. 15, 546–554. doi:10.1016/j.tics.2011.09.008

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press. doi:10.1177/1468795X06065810

Wajnberg, E., Fauvergue, X., Pons, O., 2000. Patch leaving decision rules and the Marginal Value Theorem: An experimental analysis and a simulation model. Behav. Ecol. 11, 577–586. doi:10.1093/beheco/11.6.577

Wallis, J.D., 2011. Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat. Neurosci. 15, 13–19. doi:10.1038/nn.2956

Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., Botvinick, M., 2018. Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. 21, 860–868. doi:10.1038/s41593-018-0147-8

Webb, B., 2019. The internal maps of insects. J. Exp. Biol. 222, 1–13. doi:10.1242/jeb.188094

Webb, R., Glimcher, P.W., Louie, K., 2014. Rationalizing Context-Dependent Preferences : Divisive Normalization and Neurobiological Constraints on Choice. SSRN.

Wei, X.-X., Stocker, A.A., 2017. Lawful relation between perceptual bias and discriminability. Proc. Natl. Acad. Sci. 114, 201619153. doi:10.1073/pnas.1619153114

Wikenheiser, A.M., Schoenbaum, G., 2016. Over the river, through the woods: Cognitive maps in the hippocampus and orbitofrontal cortex. Nat. Rev. Neurosci. 17, 513–523. doi:10.1038/nrn.2016.56

Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., Niv, Y., 2014. Orbitofrontal cortex as a cognitive map of task space. Neuron 81, 267–279. doi:10.1016/j.neuron.2013.11.005

Wire, E.L., Barrientos, G., 1958. Secondary Reinforcement and Multiple Drive Reduction. J. Comp. Physiol. Psychol. 51, 640–643.

Wolf, M., van Doorn, G.S., Leimar, O., Weissing, F.J., 2007. Life-history trade-offs favour the evolution of animal personalities. Nature 447, 581–584. doi:10.1038/nature05835

Wong, K.-F., Wang, X.-J., 2006. A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. J. Neurosci. 26, 1314–1328. doi:10.1523/JNEUROSCI.3733-05.2006

Woodford, M., 2012. Prospect Theory as Efficient Perceptual Distortion. Am. Econ. Rev. 102, 41–46. doi:http://www.aeaweb.org/aer/

Xue, G., Lu, Z., Levin, I.P., Bechara, A., 2010. The impact of prior risk experiences on subsequent risky decision-making: The role of the insula. Neuroimage 50, 709–716. doi:10.1016/j.neuroimage.2009.12.097

Yamada, H., Louie, K., Tymula, A., Glimcher, P.W., 2018. Free choice shapes normalized value signals in medial orbitofrontal cortex. Nat. Commun. 9, 1–11. doi:10.1038/s41467-017-02614-w

Yamada, H., Tymula, A., Louie, K., Glimcher, P.W., 2013. Thirst-dependent risk preferences in monkeys identify a primitive form of wealth. Proc. Natl. Acad. Sci. 110, 15788–15793.

Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. Psychol. Rev. 111, 931–959. doi:10.1037/0033-295X.111.4.931

Zhang, J., Berridge, K.C., Tindell, A.J., Smith, K.S., Aldridge, J.W., 2009. A neural computational model of incentive salience. PLoS Comput. Biol. 5, 9–14. doi:10.1371/journal.pcbi.1000437

Appendix A. Supplementary materials for Chapter Three

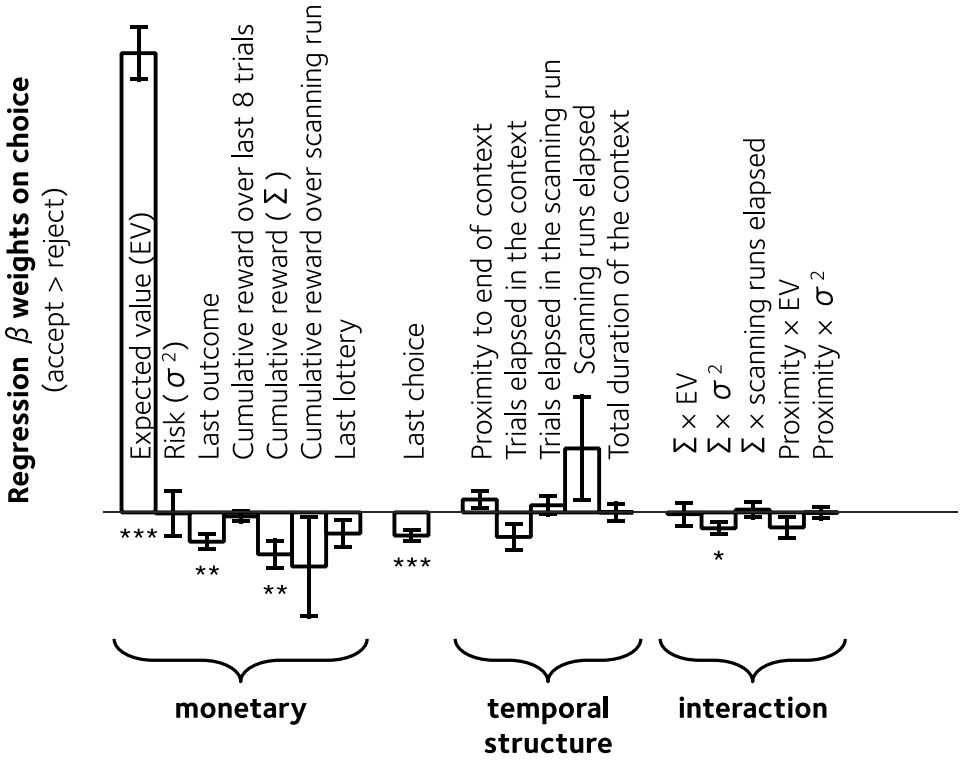


Figure S1. Related to Figure 1b. Logistic regression model 2 with controls. All regressors entered in the same competitive analysis. Bar height is mean, and error bars are standard error of the mean (SEM), where *** is $p < 0.001$, ** is $p < 0.01$, and * is $p < 0.05$ for a two-tailed t-test against zero with 19 degrees of freedom.

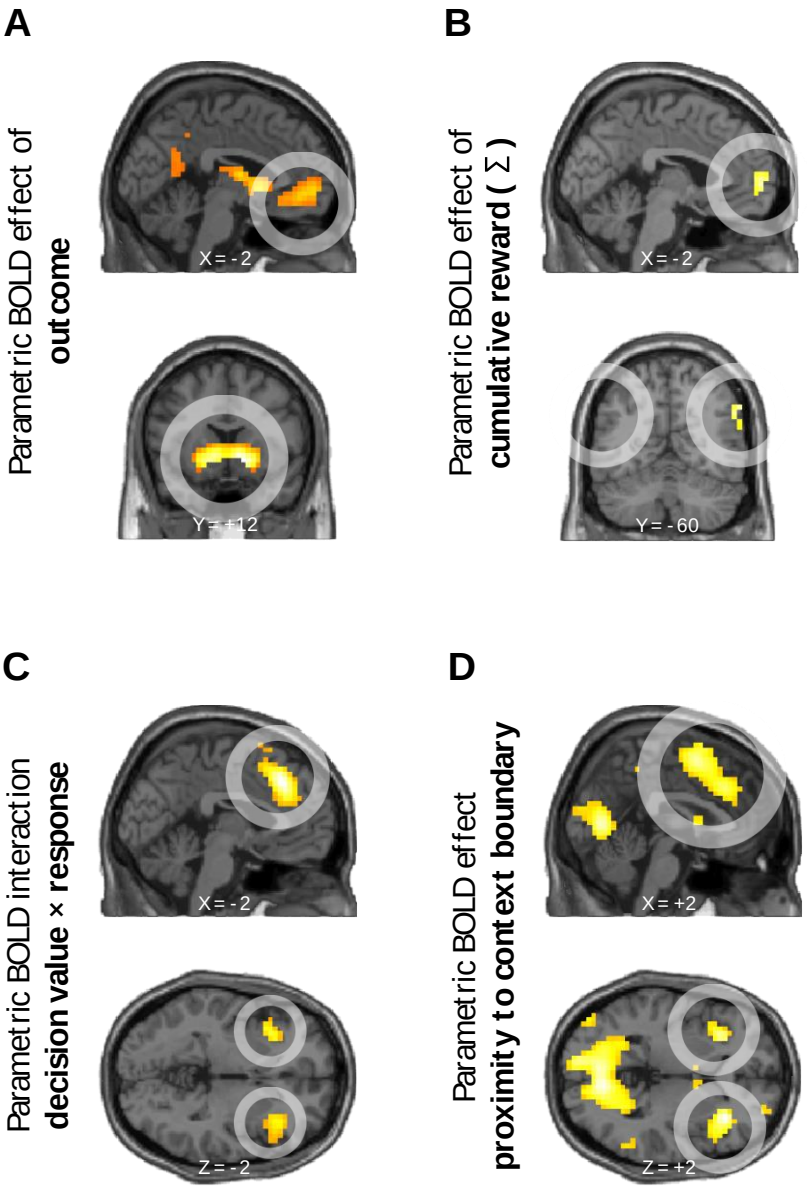


Figure S2. Related to Figures 2-5. SnPM re-analysis of all main BOLD effects. We reanalyzed all main contrasts using SnPM13 and display the results on the same slices as in the corresponding main figures thresholded for significant clusters using the non-parametric procedure described in the **methods**. **(a-b)** corresponds to **Fig. 2a-b**, **(c)** to **Fig. 4**, and **(d)** to **Fig. 5**.

Activation tables

All tables only include clusters with at least 10 voxels.

Table S1. Related to Figure 2a. Positive BOLD correlates of trial outcome

Approximate region	cluster p (FDR-corrected)	voxels in cluster	|T-Value|	Z-Value	peak p, unc.	X	Y	Z
Striatum	0.00	752	9.34	5.66	$< 1 \times 10^{-5}$	14	12	-6
			8.07	5.26	$< 1 \times 10^{-5}$	-14	12	-6
			5.57	4.24	0.00001	-26	-20	2
Right occipital	0.46	10	5.22	4.06	0.00002	30	-100	-2
Left cerebellum	0.01	76	5.07	3.99	0.00003	-22	-44	-22
			4.96	3.92	0.00004	-22	-28	-34
			4.65	3.75	0.00009	-22	-32	-22
Posterior cingulate gyrus (PCC)	0.00	166	5.05	3.97	0.00004	-10	-48	38
			4.85	3.87	0.00006	18	-48	38
			4.24	3.51	0.00022	-2	-60	26
Left occipital	0.00	132	5.02	3.96	0.00004	-14	-104	2
			4.57	3.71	0.00011	-42	-80	30
			4.28	3.54	0.00020	-26	-92	26
Left dorsolateral prefrontal cortex (dlPFC)	0.12	29	4.55	3.69	0.00011	-22	32	50
Left temporal	0.12	29	4.49	3.66	0.00013	-62	-4	-10
			4.34	3.57	0.00018	-54	-8	-14
			4.01	3.37	0.00037	-58	0	-26
Right cerebellum	0.37	13	3.94	3.33	0.00044	6	-68	-38
			3.93	3.32	0.00045	14	-60	-30

Right occipital	0.22	20	3.91	3.31	0.00047	34	-88	18
			3.81	3.25	0.00059	26	-92	22
			3.70	3.17	0.00075	38	-80	22

Table S2. Related to Figure 2b. Positive BOLD effects of cumulative reward.

Approximate region	cluster p (FDR-corrected)	voxels in cluster	|T-Value|	Z-Value	peak p, unc.	X	Y	Z
Right angular gyrus	0.06	31	8.65	5.45	$< 1 \times 10^{-5}$	58	-56	34
			4.16	3.47	0.00027	62	-60	18
Right Occipital lobe	0.00	170	6.37	4.60	$< 1 \times 10^{-5}$	38	-92	14
			5.92	4.41	0.00001	46	-76	-6
Left temporal lobe	0.04	35	5.30	4.10	0.00002	-66	-28	-6
			4.83	3.86	0.00006	-58	-28	-10
Left Occipital and Cuneus	0.03	42	5.16	4.03	0.00003	-26	-100	18
Ventromedial prefrontal cortex (PFC)	0.00	88	4.98	3.94	0.00004	-6	64	2
			4.44	3.63	0.00014	-14	52	-14
			4.44	3.63	0.00014	-14	56	-2
Left angular gyrus	0.02	48	4.81	3.84	0.00006	-54	-60	30
			4.77	3.82	0.00007	-50	-64	38
			4.06	3.40	0.00033	-62	-48	14
Left dorsolateral PFC	0.21	15	4.74	3.80	0.00007	-30	20	50
			4.04	3.39	0.00035	-26	16	42
left inferior frontal gyrus	0.16	18	4.59	3.72	0.00010	-34	48	-10
			4.30	3.55	0.00019	-38	40	-6
Left superior parietal lobule	0.13	21	4.55	3.69	0.00011	-22	-80	46
ventral striatum	0.01	59	4.54	3.69	0.00011	-14	24	-2
			4.18	3.48	0.00025	10	20	2
			4.05	3.40	0.00034	-10	4	-2
inferior occipital lobe	0.25	13	4.42	3.62	0.00015	-46	-84	-2
Left temporal lobe	0.29	11	4.38	3.59	0.00016	-58	-4	-26
			3.72	3.18	0.00073	-54	-16	-18

Table S3. Related to Figure 4a. BOLD effects of the DV by choice interaction

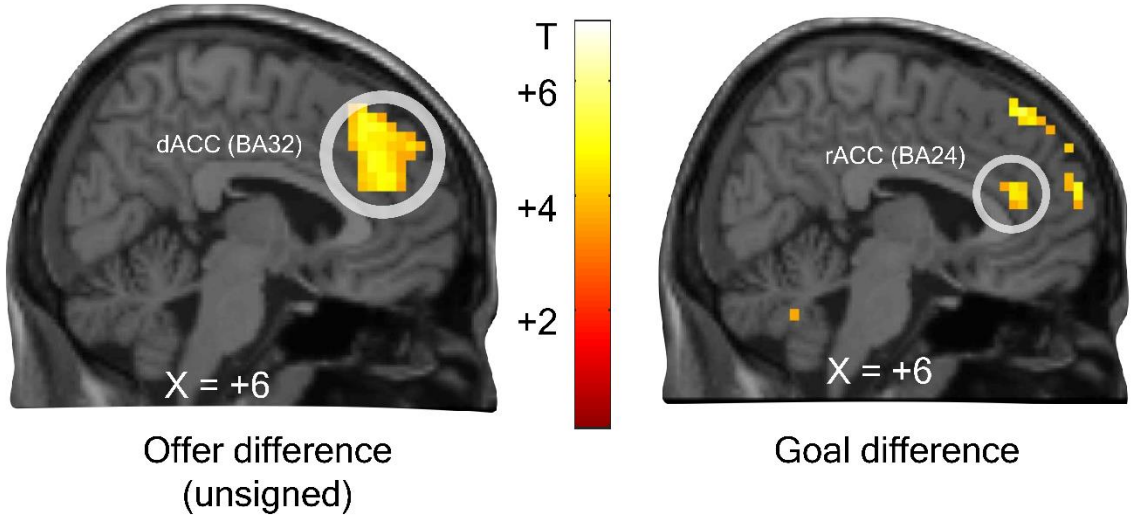
Activation	Approximate region	cluster p (FDR-corrected)	voxels in cluster	|T-Value|	Z-Value	voxel p, unc.	X	Y	Z
Positive									
	Paracentral lobule	0.27	25	4.80	3.84	0.00006	2	-32	58
				4.09	3.42	0.00031	14	-44	58
	Left precentral gyrus	0.46	14	4.22	3.50	0.00023	-54	-12	38
Negative									
	Bilateral superior frontal gyrus (sFG), medial frontal gyrus (mFG)	0.00	281	8.37	5.36	$< 1 \times 10^{-5}$	-2	32	42
				5.82	4.36	0.00001	-6	32	26
				4.88	3.88	0.00005	18	12	70
	Right inferior frontal gyrus (IFG), Insula	0.00	106	6.66	4.73	$< 1 \times 10^{-5}$	34	20	-6
	Left IFG, Insula	0.00	79	6.15	4.51	$< 1 \times 10^{-5}$	-34	20	-2
	Right angular gyrus	0.00	58	4.97	3.93	0.00004	54	-52	42
	Left sFG	0.24	10	4.94	3.91	0.00005	-14	16	70
				3.64	3.13	0.00086	-14	28	62

Table S4. Related to Figure 5. BOLD effects of trials to boundary (proximity)

Activation	Approximate region	cluster p (FDR-corrected)	voxels in cluster	|T-Value|	Z-Value	peak p, unc.	X	Y	Z
Positive									
	Occipital, prefrontal cortex (PFC), Insula, Precuneus,	0.00	7151	8.68	5.46	$< 1 \times 10^{-5}$	10	-72	2
	supplementary motor area			8.41	5.37	$< 1 \times 10^{-5}$	6	8	62
				8.34	5.35	$< 1 \times 10^{-5}$	46	24	14
	Striatum	0.00	145	6.07	4.47	$< 1 \times 10^{-5}$	6	12	10
				4.92	3.90	0.00005	-14	-4	-2
				3.99	3.36	0.00039	10	12	-6
	left dorsolateral PFC	0.00	195	5.63	4.27	0.00001	-42	40	30
				5.11	4.00	0.00003	-26	52	22
	left temporal lobe	0.30	12	4.90	3.89	0.00005	-46	-12	-6
	Left cerebellum	0.02	57	4.87	3.87	0.00005	-18	-24	-34
				4.71	3.79	0.00008	-46	-52	-34
				4.60	3.73	0.00010	-38	-44	-50
	Right cerebellum	0.29	14	4.30	3.55	0.00019	34	-40	-34
				3.63	3.13	0.00089	34	-40	-46
	Brainstem, Thalamus	0.28	16	4.15	3.46	0.00027	2	-24	2
Negative									
	Right occipital lobe	0.02	42	9.69	5.76	$< 1 \times 10^{-5}$	26	-96	-6
	Left occipital lobe	0.04	24	6.64	4.72	$< 1 \times 10^{-5}$	-26	-96	-10

Appendix B. Supplementary materials for Chapter Four

Figure S1. Related to Figure 2. BOLD correlates of offer difference and goal difference from GLM1 controlling for reaction time.



Offer difference (unsigned)

Goal difference

Figure S2. Related to Figure 2. BOLD correlates of receipt value, goal receipt value, and redress value as chosen minus foregone codes. Redress value was orthogonalized with respect to the other two value codes. GLMS1.

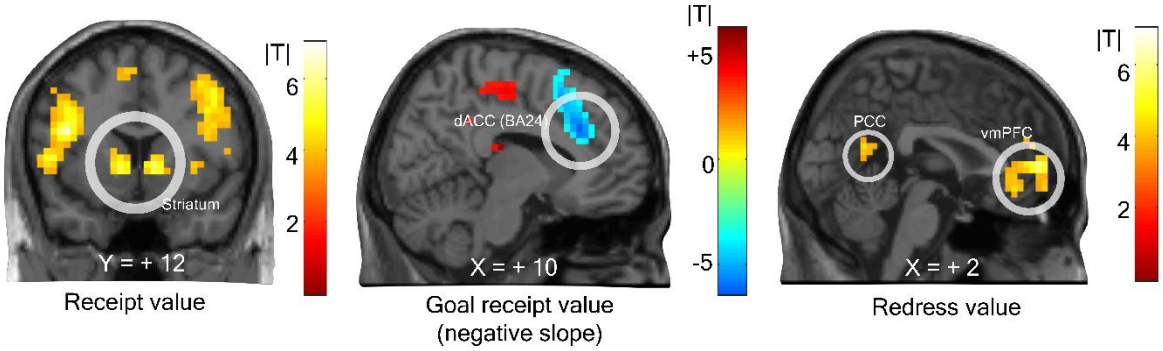


Figure S3. Related to Figure 1. Comparison of DP model and behavior. (A) Logistic regression of choice comparing participants, optimal model (with noise), and best fitting DP model. Bars are participants, error bars are SEM. (B) Choice to redress versus optimal Q-value from DP model. Grey lines are fit behaviour of individual participants based on a regression of optimal Q-value difference of their choices, black line is average across participant. Note that each participant would have encountered only a subset of the full Q-value range.

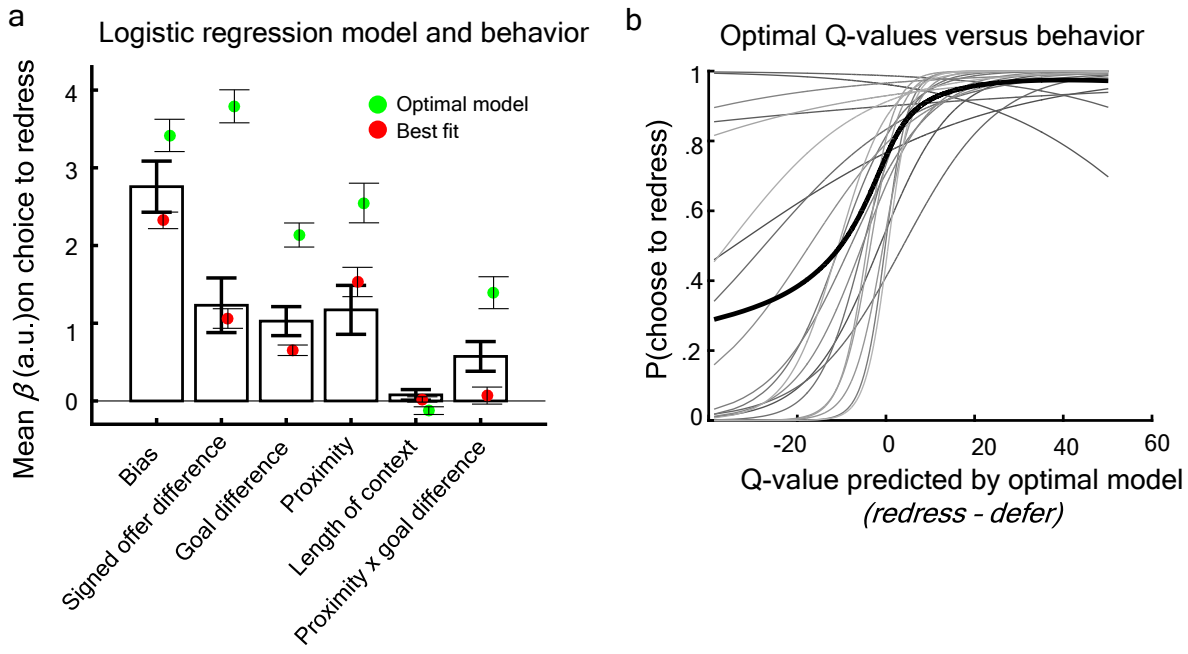
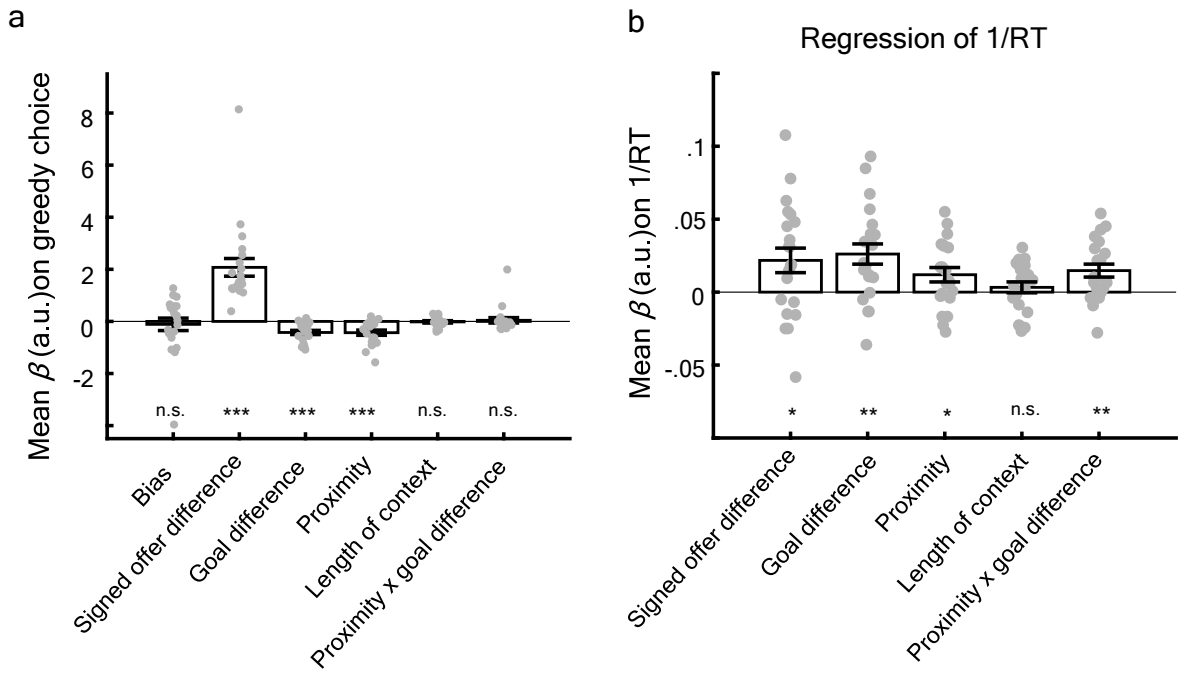


Figure S4. Related to Figure 1 and Methods. Regressions of greedy choice and reaction time. (A) Logistic regression of choice predicting choice of the higher offer (greedy). The same regressors as in the main text were used. (B) Regression of (inverse) reaction time using regressors from Fig 1b. Note that a positive beta corresponds to *faster* responses for high magnitudes of the regressor. Bar height corresponds to mean beta across participants, error bars are SEM. *** is p < 0.001, ** is p < 0.01, * is p < 0.05 for a two-tailed t-test against zero.



Tables

Table S1. Related to Figure 2. BOLD effects of unsigned offer difference, GLM1. Abbreviations: dACC – dorsal anterior cingulate cortex; PFC – prefrontal cortex; dlPFC – dorsolateral PFC; PCC - posterior cingulate cortex, dmPFC – dorsomedial PFC; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster-size	T	Z	p(uncorrected)	X	Y	Z
Dorsal anterior cingulate (dACC), bilateral dorsolateral prefrontal cortex (dlPFC)	0.0000	702	7.02	4.93	0.00000042	-14	28	50
			6.55	4.74	0.00000109	10	48	38
			6.19	4.58	0.00000237	10	32	50
Left angular gyrus	0.0014	93	6.88	4.87	0.00000055	-46	-56	30
			5.82	4.40	0.00000543	-46	-64	50
			5.58	4.28	0.00000923	-50	-60	42
Right anterior insula	0.0107	51	6.17	4.57	0.00000248	46	20	2
			5.07	4.02	0.00002921	34	20	-6
Right angular gyrus	0.0008	112	5.68	4.33	0.00000735	50	-56	30
			5.66	4.32	0.00000769	50	-52	38
			3.95	3.36	0.00039428	42	-72	46
Left anterior insula	0.0085	58	4.95	3.95	0.00003860	-34	20	-2
			4.86	3.91	0.00004695	-42	20	-6
			4.66	3.79	0.00007580	-38	16	10
Posterior cingulate (PCC)	0.1555	15	4.94	3.95	0.00003956	2	-20	26
			4.44	3.66	0.00012666	2	-28	34
Left ventrolateral PFC	0.2036	11	4.22	3.53	0.00021061	-38	44	14
			4.03	3.41	0.00033077	-34	56	10
Negative								
Left parahippocampal	0.5633	19	5.44	4.21	0.00001275	-30	-36	-18
			3.85	3.29	0.00049957	-34	-20	-18

Table S2. Related to Figure 2. BOLD effects of goal difference, GLM1; Abbreviations: ACC – anterior cingulate cortex; PFC – prefrontal cortex; dmPFC – dorsomedial PFC; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
Left ventral occipital lobe / fusiform gyrus	0.0011	132	6.91	4.88	0.00000052	-34	-48	-18
			5.87	4.42	0.00000484	-26	-44	-14
			5.44	4.21	0.00001257	-30	-56	-10
Dorsomedial PFC (dmPFC)	0.3426	20	6.07	4.52	0.00000310	10	40	54
Right rostral ACC (Brodmann area 24; BA)	0.0062	88	5.37	4.18	0.00001464	10	40	14
			5.14	4.05	0.00002509	6	64	18
			4.13	3.47	0.00025944	14	60	22
Left mid-occipital	0.1789	33	5.03	3.99	0.00003239	-30	-80	22
			4.09	3.45	0.00028366	-30	-92	14
			3.98	3.38	0.00036486	-26	-72	30
Right mid-occipital	0.2678	25	4.78	3.86	0.00005648	30	-68	30
			4.63	3.77	0.00008086	26	-68	38
Right cerebellum	0.5696	13	4.30	3.57	0.00017560	22	-68	-18
Right fusiform gyrus	0.5696	12	4.25	3.55	0.00019383	30	-48	-14
			3.93	3.35	0.00040976	30	-56	-10
			3.70	3.19	0.00070599	22	-52	-14

Table S3. Related to Figure 2. BOLD effect of receipt value, GLM2. Abbreviations: dACC – dorsal anterior cingulate cortex; PFC – prefrontal cortex; dlPFC – dorsolateral PFC; SMA - supplementary motor area; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
Left dlPFC	0.0857	52	6.53	4.72	0.00000116	-46	16	22
Bilateral striatum, thalamus	0.0000	320	5.76	4.37	0.00000612	-6	12	2
			5.74	4.36	0.00000646	10	12	2
			5.74	4.36	0.00000647	-10	-20	10
Left inferior parietal	0.2465	19	5.57	4.28	0.00000951	-38	-52	50
			5.17	4.07	0.00002310	-34	-52	42
Left mid-occipital	0.0884	42	5.31	4.15	0.00001685	-26	-68	30
Left pre-central gyrus	0.0884	44	5.25	4.11	0.00001953	-34	-4	62
			5.00	3.98	0.00003449	-46	0	50
			3.77	3.23	0.00060873	-38	0	34
Right insula	0.1842	28	5.00	3.98	0.00003470	34	28	-2
			4.00	3.39	0.00035040	34	16	-2
Right inferior parietal	0.1842	26	4.84	3.89	0.00004918	34	-48	38
			4.78	3.86	0.00005666	38	-52	46
Right pre-central gyrus	0.1842	27	4.78	3.86	0.00005739	38	4	42
Left insula	0.1926	24	4.68	3.80	0.00007224	-30	24	-2
Supplementary motor area (SMA)	0.4123	11	4.52	3.71	0.00010440	-10	16	54
			3.83	3.28	0.00052428	-6	8	58
Left cerebellum	0.2465	19	4.39	3.63	0.00014025	-6	-60	-2
Right dACC (BA24)	0.2473	18	4.30	3.58	0.00017480	10	40	6

			4.22	3.53	0.00020889	10	36	14
			4.09	3.45	0.00028527	10	28	22
Right dlPFC	0.3956	12	4.20	3.52	0.00021870	42	20	22
			4.07	3.43	0.00029979	34	24	26
Inferior occipital lobe	0.3084	15	4.17	3.50	0.00023368	-38	-56	-6

Table S4. Related to Figure 2. BOLD effect of goal receipt value, GLM2; Abbreviations: dACC – dorsal anterior cingulate cortex; BA – Brodmann area; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
Negative								
dACC (BA24)	0.0388	43	5.34	4.16	0.00001570	14	28	26
			5.06	4.01	0.00003015	14	40	14
Right insula	0.0548	30	5.30	4.14	0.00001726	30	20	-10
Left insula	0.1869	13	4.64	3.78	0.00007866	-34	24	-6

Table S5. Related to Figure 2. BOLD effect of redress value, GLM2; Abbreviations: PCC - posterior cingulate cortex; PFC – prefrontal cortex; vmPFC – ventromedial PFC; dmPFC – dorsomedial PFC; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
PCC	0.0040	105	6.53	4.73	0.00000114	-6	-60	14
			3.67	3.17	0.00076418	18	-56	26
vmPFC	0.0006	162	5.50	4.25	0.00001092	6	52	2
			4.78	3.86	0.00005769	2	44	-10
			4.54	3.72	0.00009877	-2	52	18
Left mid-cingulate cortex	0.7017	13	5.46	4.22	0.00001195	-10	-36	38
Left fusiform gyrus, hippocampus	0.1471	36	4.72	3.83	0.00006512	-34	-24	-14
			4.44	3.66	0.00012439	-30	-44	-6
			4.41	3.64	0.00013436	-26	-28	-18
Anterior dmPFC	0.7017	10	4.50	3.70	0.00010941	-6	60	34
Negative								
Left superior frontal gyrus	0.3219	10	5.27	4.12	0.00001854	-26	4	66
Right inferior parietal lobule	0.3219	11	4.20	3.51	0.00022188	38	-40	42
Left inferior parietal lobule	0.3097	22	4.17	3.50	0.00023418	-50	-48	54
			4.03	3.41	0.00032863	-46	-40	50
			3.83	3.27	0.00052858	-42	-56	58

The following tables are from GLMS1, which included the same regressors as GLM2, but as chosen minus foregone instead of chosen value. Redress value was orthogonalized with respect to receipt value and goal receipt value.

Table S6. Related to Figure 2. BOLD effect of receipt value (chosen-foregone), GLMS1; Abbreviations: PFC – prefrontal cortex; dlPFC – dorsolateral PFC; SMA – supplementary motor area; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
Left dlPFC	0.0000	392	7.32	5.05	0.00000022	-46	16	22
			5.64	4.31	0.00000803	-54	16	2
			5.36	4.17	0.00001514	-42	0	46
Left precentral gyrus	0.0001	164	6.78	4.83	0.00000068	-30	-64	42
			6.11	4.54	0.00000283	-38	-52	46
Right inferior and superior frontal gyri	0.0000	290	6.07	4.52	0.00000307	38	32	22
			6.02	4.50	0.00000346	42	8	42
			5.43	4.21	0.00001296	46	24	30
Bilateral striatum, thalamus	0.0000	238	5.84	4.41	0.00000520	10	8	2
			5.68	4.33	0.00000730	18	-4	2
			5.65	4.32	0.00000790	-10	8	2
Left inferior occipital lobe	0.1835	21	5.65	4.32	0.00000785	-38	-56	-6
Right angular gyrus	0.0001	166	5.36	4.17	0.00001521	34	-48	38
			5.20	4.09	0.00002189	34	-48	50
			4.84	3.89	0.00005016	-6	-68	54
Right anterior insula	0.0549	42	5.17	4.07	0.00002316	34	28	-2
			4.04	3.41	0.00032328	34	12	2

Right superior temporal lobe	0.0077	77	4.73	3.83	0.00006449	50	-48	14
			4.29	3.57	0.00017657	46	-52	-10
			3.91	3.33	0.00043258	46	-44	-2
Left cerebellum	0.1835	20	4.65	3.78	0.00007724	-6	-60	-2
			3.85	3.29	0.00049767	-14	-64	2
Left cerebellum	0.3214	13	4.50	3.70	0.00010942	-38	-56	-26
			4.01	3.39	0.00034745	-30	-60	-26
SMA	0.1604	24	4.47	3.68	0.00011666	-6	8	58
			4.10	3.45	0.00027606	-10	16	54
			3.64	3.15	0.00081270	-2	24	50
Right hippocampus	0.0727	36	4.39	3.63	0.00014088	18	-36	2
			4.21	3.52	0.00021632	14	-40	-10
			4.07	3.43	0.00029983	10	-48	-2

Table S7. Related to Figure 2. BOLD effect of goal receipt value (chosen-foregone), GLMS1; Abbreviations: SMA – supplementary motor area; dACC – dorsal anterior cingulate cortex; PFC – prefrontal cortex; dlPFC – dorsolateral PFC; PCC - posterior cingulate cortex, vmPFC – ventromedial PFC; BA – Brodmann area; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
Bilateral inferior occipital, hippocampi, posterior insula, SMA	0.0000	1182	6.19	4.58	0.00000238	-38	-36	-6
			6.03	4.50	0.00000343	22	-12	30
			5.84	4.41	0.00000514	-14	-48	22
Left superior temporal lobe	0.1260	40	5.88	4.43	0.00000469	-66	-16	6
			5.27	4.12	0.00001855	-62	-12	-6
			4.25	3.54	0.00019783	-62	-4	-18
vmPFC	0.1019	49	5.55	4.27	0.00000995	-2	56	-6
			4.72	3.82	0.00006619	-2	36	-10
			4.05	3.42	0.00031233	2	24	-10
Right medial temporal lobe	0.4587	20	5.45	4.22	0.00001230	58	-8	-14
Ventricle	0.6942	10	4.38	3.63	0.00014445	-2	-4	22
White matter	0.5402	14	4.30	3.58	0.00017388	22	40	-2
			4.08	3.44	0.00029504	22	32	-6
Mid-occipital lobe	0.5322	16	4.14	3.48	0.00025243	-22	-88	2
			4.05	3.42	0.00031095	-18	-80	18
			3.85	3.29	0.00049478	-26	-76	10
Negative								
dACC (BA24 and 32), SMA	0.0000	214	6.36	4.65	0.00000165	-6	16	50
			6.32	4.63	0.00000182	6	16	50
			6.07	4.52	0.00000309	10	28	30
Right anterior insula	0.0032	80	5.45	4.22	0.00001224	42	20	6
			4.82	3.88	0.00005167	34	24	2
Left anterior insula	0.0032	74	5.40	4.19	0.00001387	-42	20	2

	0.0784	23	5.19	4.08	0.00002235	-38	56	18

Table S8. Related to Figure 2. BOLD effect of redress value (chosen-foregone), GLMS1. Abbreviations: PFC – prefrontal cortex; vmPFC – ventromedial PFC; PCC - posterior cingulate cortex; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

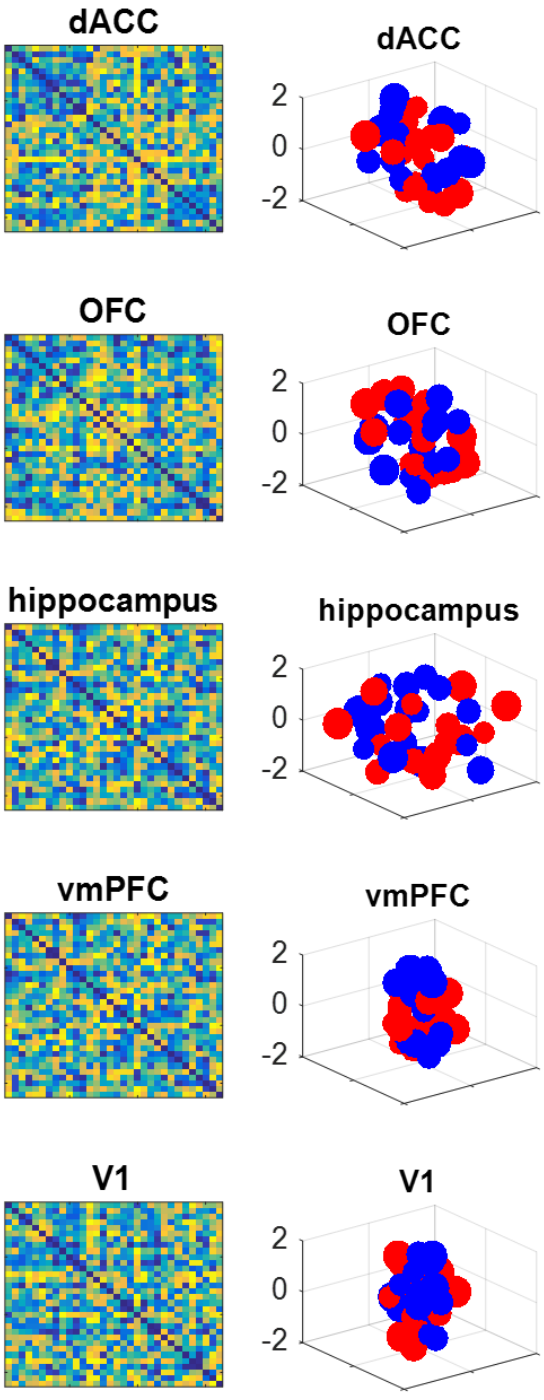
Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
vmPFC	0.0019	151	6.04	4.51	0.00000329	2	56	6
			4.90	3.93	0.00004289	-14	48	2
			4.43	3.65	0.00012960	14	40	-2
PCC	0.1366	41	5.15	4.06	0.00002457	-6	-60	14
			3.78	3.24	0.00059327	6	-52	14

Table S9. Related to Figure 4. BOLD effect of proximity from GLM1; Abbreviations: iPL – interior parietal lobule; dACC – dorsal anterior cingulate cortex; PFC – prefrontal cortex; dlPFC – dorsolateral PFC; dmPFC – dorsomedial PFC; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	cluster size	T	Z	p(uncorrected)	X	Y	Z
Left inferior parietal lobule (iPL), angular gyrus	0.0002	214	7.83	5.24	0.00000008	-54	-48	38
			4.70	3.82	0.00006799	-50	-64	30
			3.95	3.36	0.00039512	-38	-44	58
Right iPL, angular gyrus	0.0000	300	7.53	5.13	0.00000015	58	-40	38
			5.87	4.43	0.00000482	42	-44	34
			5.64	4.31	0.00000809	54	-40	54
Left dlPFC	0.0006	176	5.64	4.31	0.00000807	-34	20	46
			5.22	4.10	0.00002054	-26	12	58
			5.17	4.07	0.00002329	-30	28	38
Right dlPFC	0.0001	263	5.31	4.15	0.00001677	30	4	62
			5.22	4.10	0.00002096	46	12	14
			4.76	3.85	0.00005938	38	4	46
Precuneus	0.0017	143	4.89	3.92	0.00004380	-10	-56	58
			4.34	3.60	0.00015833	14	-60	58
			4.26	3.55	0.00018936	-6	-56	42
dmPFC	0.6747	14	4.82	3.88	0.00005256	-10	52	34
Ventrolateral PFC	0.2852	31	4.70	3.81	0.00006852	34	64	10
Mid cingulate gyrus	0.2852	29	4.32	3.59	0.00016688	10	-16	38
			4.03	3.40	0.00033135	2	-20	42
			4.00	3.39	0.00035049	-10	-8	38
Middle temporal gyrus	0.6431	16	4.14	3.48	0.00025476	54	-48	10
			4.01	3.39	0.00034537	54	-52	2
			3.89	3.32	0.00045135	46	-44	2

Appendix C. Supplementary materials for Chapter 5

Supplementary Figure 1. Neural RDMs for the five ROIs and MDS plots. The first 16 rows and columns represent each individual car in a Pablo market, and rows and columns 17 to 32 represent the same cars in a María market. MDS plots into three dimensions were z-transformed to allow comparison across ROIs. Red dots are from the María markets, blue dots from a Pablo market.



Supplementary Table 1. Results from searchlight analysis in refresher task for the absolute value difference RDMs. Abbreviations: dACC – dorsal anterior cingulate cortex; dlPFC – dorsolateral prefrontal cortex; SMA – supplementary motor area; FDR – cluster-wise false discovery rate; T – t-value test statistic; Z – standardized T statistic; X,Y,Z – coordinates according to the Montreal Neurological Institute template (in millimeters)

Region	FDR p	Cluster size	T	Z	p(uncorrected)	X	Y	Z
dACC, dlPFC, SMA	2.01E-12	3560	6.85	5.09	1.73E-07	-37.5	49	14
			6.71	5.02	2.45E-07	-23.5	17.5	52.5
			6.37	4.86	5.68E-07	8	45.5	28
Bilateral angular gyri, precuneus, inferior parietal lobe	6.95E-07	1413	5.77	4.56	2.53E-06	29	-63	52.5
			5.63	4.48	3.63E-06	53.5	-56	31.5
			5.13	4.20	1.33E-05	-34	-66.5	35
Right anterior temporal lobe	0.46	62	4.55	3.85	5.84E-05	36	17.5	-21
			4.11	3.56	0.000183	39.5	10.5	-17.5
			3.91	3.42	0.000312	39.5	24.5	-10.5
Right precentral gyrus	0.8729	17	4.08	3.54	0.000199	25.5	-24.5	59.5
Left medial temporal lobe	0.7654	29	4.02	3.50	0.000229	-55	-7	-21
Left posterior medial temporal lobe	0.8821	11	3.87	3.39	0.000345	-58.5	-35	-7