Neurobiology of social and individual choice

Nicholas David Wright

Wellcome Trust Centre for Neuroimaging
Institute of Neurology
University College London

Dissertation submitted for the degree of
Doctor of Philosophy
of
University College, London

September 2011

Declaration

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

23rd September, 2011

Abstract

In the course of our everyday lives, we are constantly faced with situations in which we must choose. Do we invest in the bank or the stock-market? Is a new wage deal so unfair that we should resort to a strike? These situations are elegantly described mathematically by Rational Choice Theory (RCT), which dominates the quantitative social sciences such as economics. However, unfortunately RCT often fails to predict how humans actually behave. Here I investigate choice using paradigms derived from the RCT framework, but aim to better predict actual choices by using a biological level of explanation. First, I examine simple choices that involve no social interaction, asking how choices are influenced by risk in potential outcomes, and by whether outcomes reflect potential gains or losses. The data reveal independent impacts of risk and loss on choice, findings not predicted by extant economic theories. Instead, I then harness functional Magnetic Resonance Imaging (fMRI) to suggest a biological mechanism by which risk and loss bias approach behaviour, and test this hypotheses in further behavioural experiments. Secondly, I examine social choices. Specifically, I examine biological systems that enable social behaviour to respond flexibly to environmental contingencies. I investigate the neural basis of the human fairness motivation using fMRI, and show how it flexibly adapts to external social context. Next, I show how this fairness motivation adapts to changes in an individual’s internal physiological state. Finally, I show how cooperation is modulated by the androgen hormone testosterone. Overall, in light of these non-social and social findings, I propose that a biologically-based account of choice can explain choices that are not predicted by existing theory.

Acknowledgements

I have been very lucky and I owe thanks to many people. When I first approached Ray Dolan with my ideas, he kindly listened and then supported my application to the Wellcome Trust. As my supervisor, Ray has given me freedom, provided guidance and demanded excellence. I owe Ray a very great deal. I can also highly recommend his taste in wine. Anyone who knows my second supervisor, Peter Dayan, will know his remarkable intellectual rigour. A parody of English history called *“1066 and all that”*, divided everything up into Good Things and Bad Things – and I think Peter would certainly be classed as a Good Thing. Equally, the FIL would not exist without the brilliance of Karl Friston, who is clearly another Good Thing.

However, those to whom I owe my greatest debt are my parents. They have always been kind, and I wish my mother had lived to see me complete this work. I also thank Marsha, whom I love and who always makes everything a lot more fun!

My PhD has also been immeasurably more fun because of the great friends I’ve made. Who doesn’t want to work in a lab with three hot blondes? Tamara was my first friend at the FIL, and instigated many memorable (if boozy) evenings. Tali and Rosalyn are the most elegant basic neuroscientists in the UK. And for symmetry I would mention three good friends from the bonobo club. Steve is simply a great guy. Mkael will be pleased I am advertising Peter Bossaerts’ belief that he is the most talented mathematician in British Neurology. Marc is my favourite ever Continental.

With the specific work in this thesis, I am particularly grateful for analytical help from Mkael Symmonds (Chapters 4 and 6), Steve Fleming (Chapter 6), and Bahador Bahrami (Chapter 8). Geraint Rees and Chris Frith collaborated on Chapter 8. I thank the MSc students who helped with data collection: Karen Hodgson (Chapters 4 and 7); Bonni Crawford (Chapter 5); and Emily Johnson and Gina Di Malta (Chapter 8).

There are, of course, too many others. Cathy Price’s attitude to science was inspirational to me. All the FIL support staff. Everyone, really.

Table of contents

List of Figures

List of Tables

Abbreviations

ACC	Anterior Cingulate Cortex
BIC	Bayesian Information Criterion.
DLPFC	Dorsolateral prefrontal cortex
DMPFC	Dorsomedial prefrontal cortex
E-A	Egocentric-Allocentric
EUT	Expected Utility Theory
fMRI	functional Magnetic Resonance Imaging
FWE	Family-wise error
IFG	Inferior Frontal Gyrus
MCC	Middle Cingulate Cortex
MFG	Middle Frontal Gyrus
MTG	Middle Temporal Gyrus
OFC	Orbitofrontal cortex
PT	Prospect Theory
RCT	Rational Choice Theory
RT	Reaction time
rTMS	Repetitive transcranial magnetic stimulation
SFG	Superior Frontal Gyrus
SMA	Supplementary Motor Are
STG	Superior Temporal Gyrus
STS	Superior Temporal Sulcus
SVC	Small volume corrected
TOM	Theory of Mind
UG	Ultimatum Game

Publications from this PhD thesis

All data chapters in this thesis have been submitted for publication. At the time of writing, the data contained in Chapter 6, which uses fMRI to examine the contextual manipulation of fairness, has been published in the *Journal of Neuroscience* (Wright et al., 2011).

Chapter 1. Introduction

The choices we make sculpt our lives. Do I continue working in my secure job, or do I take a risk and start my own business? Do I invest my savings in the stock market, in housing, in Government bonds, or do I take a trip and play the casinos in Las Vegas?[1] If I am part of a farming community where communal irrigation systems must be maintained: do I cooperate with my neighbours, or do I let them do the work and free-ride? Understanding how humans make such individual and social choices is important – not only because this determines how we now live in our society, but also because it constrains the potential structural arrangements in society.

What do we know about how humans make choices? One starting point is the philosopher Plato (2005) who painted a picture of the soul as a chariot, comprising a charioteer directing competing motivations: *“First the charioteer of the human soul drives a pair, and secondly one of the horses is noble and of noble breed, but the other quite the opposite in breed and character. Therefore in our case the driving is necessarily difficult and troublesome”.* Such richness and complexity was also evident two millennia later in the writings of Adam Smith, generally acknowledged as the father of modern economics. Smith discussed the more passionate side of human nature in his *Theory of Moral Sentiments* (1759), which begins *“How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it.”* Later in Smith’s book *The Wealth of Nations* (1776), which can be seen as the origin of neo-classical economics, he describes how even when the individual “*intends only his own gain, and he is in this, … led by an invisible hand to promote an end which was no part of his intention.”* Unfortunately, in the latter half of the twentieth century the quantitative social sciences, such as economics, lost the richness of earlier

[1] At the time of writing, some might consider Las Vegas to be the safest option.

descriptions and came to be dominated by consideration of only one of these components of human nature: rational self-interest.

Since the mid-twentieth century, economics in particular has been dominated by this one mode of analysis, centred on the Rational Choice Theory (RCT) (von Neumann and Morgenstern, 1944). RCT models individual choices through Expected Utility Theory, and social choices through Game Theory. RCT does in fact successfully explain many aspects of human behaviour and provides simple and mathematically rigorous descriptions of situations in which individuals must choose. However, as discussed in Chapter 2, RCT is limited as a descriptive model, and fails to predict many aspects of human choice.

To improve the predictive power of such models, over the past three decades a subfield of economics has begun to put the passions back into models of behaviour. This field of behavioural economics aims to *“increase the explanatory power of economics by providing it with more realistic psychological foundations”* (Camerer and Loewenstein, 2004). However, *“it is important to emphasize that the behavioural economics approach extends rational choice and equilibrium models; it does not advocate abandoning these models entirely*” (Ho et al., 2006). This psychologically-informed economics has been applied to individual choices, for example attempting to replace Expected Utility Theory with Prospect Theory (Kahneman and Tversky, 1979), and has been applied to social choices, for example replacing Game Theory with Behavioural Game Theory (Camerer, 2003). However, important behavioural regularities are still not predicted by these models, including behaviours described in this thesis.

More recently, biologically-based and neuroscientific approaches to choice have been combined with those from economics (Glimcher, 2003; Glimcher and Rustichini, 2004; Camerer et al., 2005). The term neuroeconomics has been coined in this context. Common to different perspectives on neuroeconomics, and to the work in this thesis, is that the fact that the main object of interest is the study of value-based decision-making. Such value-based decision-making occurs whenever an animal makes a choice from several alternatives on the basis of the subjective value it places upon them. In a sense, this inter-disciplinary approach permits the reintroduction of earlier richness and complexity into models of human behaviour, but within a mathematically specifiable and empirically grounded framework. Further, these inter-disciplinary approaches provide better descriptive models of choice.

In this thesis I aim to understand how, from a biological and psychological perspective, people make choices. Specifically, I investigate two paradigmatic influences on individual choice, namely risk and loss (Chapters 4 and 5); and two paradigmatic influences on social choice, namely fairness and cooperation (Chapters 6, 7 and 8). I use concepts from the quantitative social sciences, and in particular economics, to provide a tractable framework in which to examine these choices. To understand how these choices are determined biologically, I use behavioural model comparison in conjunction with functional magnetic resonance imaging (fMRI), and employ causal manipulations of hormones and physiological state (thirst).

1.1 Thesis outline

First, I will describe the theoretical and empirical background to the thesis (Chapter 2). Specifically, I will describe Rational Choice Theory and biologically-based approaches to individual and social choice. In Chapter 3, I describe the functional Magnetic Resonance Imaging methods that I employ in this thesis.

The empirical work presented in this thesis is contained in Chapters 4 to 8. In Chapters 4 and 5, I examine the effects of risk and valence on individual choices. In Chapter 4, I describe a new paradigm from which neural data is used to infer that risk and loss influence choice by biasing individuals away from approaching (choosing) stimuli incorporating these variables. This mechanistic hypothesis makes specific predictions concerning reaction time biases, which I then test in Chapter 5. Chapters 6 to 8 examine social choice, and specifically the biological systems that enable social behaviour to respond flexibly to environmental contingencies. In Chapter 6, I investigate the neural basis of a human fairness motivation using fMRI, and show how it flexibly adapts to external social context. In Chapter 7, I show how this fairness motivation adapts to changes in an individual’s internal physiological state. Finally, in Chapter 8, I show how cooperation is modulated by the androgen hormone testosterone.

Chapter 2. Literature review

2.1 Overview

My review of the literature contains four sections. First, I will describe Rational Choice Theory (RCT), which provides the dominant theoretical treatments of social and non-social choice in the quantitative social sciences. Second. I will address biologically-based theories that seek to move beyond RCT in the description of choice behaviour. Third, I will review how the treatments of RCT and biologically-based theories of choice relate to two specific influences on individual choice examined in this thesis – the risk and valence of potential outcomes. Finally, I will ask how these approaches relate to social choices, and specifically how they relate to the concepts of fairness and cooperation that are examined in this thesis.

2.2 Rational Choice Theory

2.2.1 Rational Choice Theory: setup and axioms

Rational Choice Theory can be simply described as follows: an agent chooses the best action according to that agent's preferences, from amongst all the actions available to the agent. No qualitative restriction is placed upon the decision-maker's preferences – the agent's "rationality" lies in the consistency of choices, not in the chooser's tastes. The model has two basic components: a set of actions available to the agent (denoted by the set *X* below); and a specification of the agent's preferences (Kreps, 1990).

2.2.1.1 The objects of choice: the set of available actions

We are interested in the behaviour of an individual, who is faced with the problem of choosing from among a set of objects (we can also say that this is a set of actions). Let *X* represent some set of objects. The agent knows the set of available objects. Within this set, *X*, we can describe consumption bundles, for example denoted by *x* or *y*. In individual, non-social choice an example of a consumption bundle with three commodities could be $x=(x_1, x_2, x_3)$, representing x_1 cans of beer, x_2 bottles of wine, and x_3 shots of whisky. In social choices an example could be a particular division of the bill at a restaurant.

2.2.1.2 The basic preference relation

As to preferences, we assume that if the agent is presented with any pair of actions, she knows which she prefers or if she is indifferent between them. For example when asking the agent to compare two alternatives, *x* and *y*, if the consumer prefers *x* to *y*, then we can write $x \succ y$, or state that *x* is strictly preferred to *y*.

2.2.1.3 Assumptions

As described above, no qualitative restriction is placed upon the agent's preferences – her "rationality" lies in the consistency of choices. To ensure this consistency the following assumptions must hold.

> *Assumption 1. Preferences are asymmetric: there is no pair x and y from X such that* $x \succ y$ *and* $y \succ x$.

This first assumption means that individuals cannot prefer *x* to *y* and also prefer *y* to *x*. However, there are problems with this assumption. One problem relates to the time of choice. For example, an agent might prefer more beer one day but more wine another. Indeed, how such changes in preference during social choices might be induced by social context, internal physiological state and hormones is subject of Chapters 6, 7 and 8 respectively in this thesis. Another problem relates to the framing of the choice as gains or losses, which is discussed later in this Chapter

A second assumption is that if an agent makes the judgment $x \succ y$, she is able to place any other option z somewhere on the ordinal scale set by these two.

Assumption 2. Preferences are negatively transitive. If $x \succ y$*, then for any third element* z*, either* $x \succ z$*, or* $z \succ y$*, or both.*

Finally, three further properties are necessary for strict preference:

Irreflexivity: For no x *is* $x \succ x$*.*

Transitivity: $x \succ y$ *and* $y \succ z$*, then* $x \succ z$*.*

Acyclicity: If, for a given integer n*,* $x_1 \succ x_2$*,* $x_2 \succ x_3$*, … ,* $x_{n-1} \succ x_n$*, then* $x_n \neq x_1$*.*

Note that we can also define further such preference relations where individuals are indifferent between alternatives, or where they weakly prefer alternatives.

2.2.2 Utility and utility functions

How do we describe an agent's preferences? We could specify preference relations for each pair of actions, but this is impractical in all but the simplest circumstances. Instead we can use a numerical scale, known as a utility function. This also has the advantage that we can turn a choice problem into a numerical maximisation problem. The utility function, *U*, represents an agent's preferences if, for any actions *x* in *X* and *y* in *X*

$$x \succ y \quad \text{if and only if} \quad U(x) \succ U(y) \qquad \text{Eq. 2.1}$$

That is, *U* measures all the objects of choice on a numerical scale, and a higher measure on the scale means the agent likes that object more. To permit a numerical representation, it is necessary that the assumptions above are met, and also that either the set *X* is small or that preferences are well behaved (Kreps, 1990).

The units in a utility scale have no particular meaning. An agent's preferences, in the sense used here, convey only ordinal information. Utility functions therefore also only convey ordinal information. Note, however that recent work has sought to correlate utility measures with neural activity, which is discussed below.

2.2.3 Relating choice and preference: revealed preference

Under this model, it is easy to see how choices derive from preferences, with the agent happy to choose anything that isn't bettered by something else that is available. However, unless we provide our consumer with a questionnaire or otherwise directly enquire as to her preferences, the only signs of her preferences are the actual choices she makes. We can thus take choice behaviour as the primitive and infer the preferences – and a method to achieve this is known as revealed preference. Samuleson's (1938) original observation that preference could be inferred from choice, relies on the assumption that when choosing between two options *A* and *B*, a subject will choose *A* more frequently than *B* if (and only if) *A* is more desirable.

$$\text{desirability}(A) > \text{desirability}(B) \rightarrow \text{p(choose } A) > \text{p(choose } B) \qquad \text{Eq. 2.2}$$

One strategy here is to present different amounts of *A* and *B* over multiple trials, and find the point at which she is indifferent between *A* and *B*, from which one can obtain an ordinal ranking of the desirability of each (Corrado et al., 2008).

2.2.4 Rational Choice Theory: emergence of behavioural economics

Before turning to biologically-based models of choice in the next section, it is important to mention that over the past three decades economists have sought to incorporate psychological insights into models derived from RCT, to improve the predictive power of such models. The resultant field is termed behavioural economics. As described by those in the field, "*It is important to emphasize that the behavioural economics approach extends rational choice and equilibrium models; it does not advocate abandoning these models entirely*" (Ho et al., 2006). We return to behavioural economic models in later sections, specifically Prospect Theory and behavioural game theory.

2.3 Biologically-based approaches to value-based choice

Biologically-based and neuroscientific approaches to choice have a long theoretical and empirical tradition, for example relating to the vast literature on associative learning and choice (Thorndike, 1911; Mackintosh, 1983). More recently, these biological approaches have been combined with economic approaches based in RCT to form the emerging field of neuroeconomics (Glimcher, 2003; Glimcher and Rustichini, 2004; Camerer et al., 2005). The term neuroeconomics has meant different things to different workers in the field: including the application to neurobiology of mathematical structures derived from microeconomics (Glimcher, 2004); or the use of biological insights to provide better assumptions in microeconomic models (Camerer et al., 2005). However, common to these different perspectives on neuroeconomics, and to the work in this thesis, is that the main object of interest is the study of value-based decision-making. By value-based decision-making, I mean cases where an agent chooses from several alternatives based on the subjective values it places upon them.

In this section, I describe two key aspects of this biological approach: first describing choice as a process, which enables selection between models not distinguishable from behaviour alone; and second, describing how choice arises from multiple interacting systems, which may explains many “irrationalities” and “biases” as seen from the perspective of RCT.

2.3.1 Choice as a process: differentiating behavioural models

Unlike RCT, biologically-based theories provide insight into the mechanistic processes underlying choice behaviours. At a fundamental level, the purpose of a nervous system is to implement appropriate behaviours in response to environmental contingencies. Many schemas have been proposed to describe the mapping from sensory inputs to motor outputs, of which one is shown in Fig. 2.1 (Corrado and

Doya, 2007; Corrado et al., 2008). In this schema, a decision occurs when an organism, confronted by discrete options, evaluates those options and selects one to act upon. When the organism's choice is not mandated by the immediate sensory characteristics of the options, but rather by the organism's subjective preference, these can be thought of as value-based decisions. Linking option evaluation and action selection are decision variables, which are quantities internal to the subject's decision process that summarise properties of the available behavioural options relevant to guiding choice (Corrado and Doya, 2007; Corrado et al., 2008).

Different models of choice may possess markedly different internal components (i.e. decision variables), but these models may result in very similar choices. It is impossible to choose between such models on the basis of behaviour alone (e.g. as in revealed preference discussed above). However, if the models include explicit decision variables that can be calculated on every trial, we can ask if correlates of these decision variables can be identified within the brain. We use such model-based fMRI analysis (O'Doherty et al., 2007) to identify correlates of the variance of gambles in Chapter 4 and of social inequality in Chapter 6.

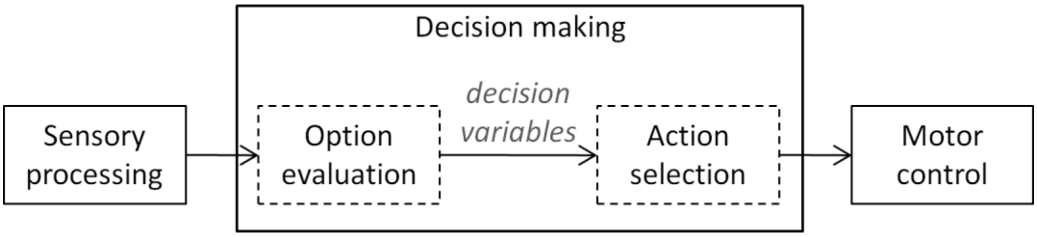

Figure 2.1 Choice is a process. *A simplified model in which sensory systems provide data to decision-making circuits, which in turn direct motor systems. The machinery of decision-making can be further subdivided into mechanisms more concerned with the evaluation of options and those more concerned with action selection: steps that can be linked by decision-variables. Adapted from (Corrado and Doya, 2007; Corrado et al., 2008).*

2.3.2 Choice results from multiple interacting decision systems

Growing evidence suggests that rather than a unitary process, as in the deliberately over-simplified sketch above, choice is actually the result of multiple interacting valuation and decision systems (Dayan, 2008; Rangel et al., 2008; Dayan and Seymour, 2009). Different weighting to the operation of these systems under different contexts may explain many of the apparent irrationalities or biases in choice seen with respect to the predictions of RCT.

How many such systems exist and how they may interact is still a matter of much debate (Dayan, 2008; Rangel et al., 2008), although there is good evidence for at least three such types of systems. First, a Pavlovian system assigns values that engage a limited set of behaviours that are evolutionarily appropriate responses to environmental stimuli. These responses include preparatory behaviours such as approaching cues that predict the delivery of food (appetitive cues), or avoidance behaviours when cues predict a punishment (aversive cues). Second are habitual systems, by which agents learn by trial-and-error to assign values to stimulus-response associations. Finally, model-based systems assign values to actions by computing action-outcome associations and then evaluating the values associated with the different outcomes.

2.4 Individual choice: risk and valence

I examine individual choices in Chapters 4 and 5 of this thesis, and specifically two important influences on choice: the risk and valence of potential outcomes. Valence is relatively easy to define, meaning whether the potential outcomes under consideration entail punishments (e.g. financial losses or painful electric shocks) or rewards (e.g. financial gains, tasty foods, or water when thirsty). However, risk is more difficult to define. In this thesis, I define risk as a state in which the decision-maker lacks knowledge about which potential outcome will follow from a choice, and I limit my enquiry to situations in which the agent knows all the probabilities associated with potential outcomes.

I acknowledge that a situation such that the decision-maker knows all probabilities reflects only one aspect of risk, albeit an important one. Another important type of risk is illustrated in the comments made by Donald Rumsfeld, United States Secretary of Defense, in 2002: "*There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know.*" As well as the known unknowns studied in this thesis and the unknown unknowns described by Rumsfeld, we could also describe a further type of risk: when clinicians or lay people identify behaviours as risky, they typically invoke a broader meaning where the behaviours may lead to harm, such as when mountain climbing (Schonberg et al., 2011). How such different aspects of risk may relate is unclear, and indeed individual differences in risky choice appear to be domain specific across domains of risk (Slovic, 1964; Weber et al., 2002).

Risk in situations where the decision-maker knows all probabilities, as examined in this thesis, has the advantage of being well defined mathematically. In the seventeenth century Pascal began to describe risky options in games of chance. Pascal proposed that one could choose the option that carried the greatest combination of value and probability, by calculating its expected value (EV) where each outcome is weighted by its probability of occurring (*v x p*, where *v* is value and *p* is probability). However, calculating expected value had the unfortunate property of giving infinite expected value to certain types of gambles (an example being the St Petersburg paradox; Glimcher, 2003). One way around this is to convert objective values into subjective utilities, where the more money an individual has the less each additional unit is worth (this concept of diminishing marginal utility explains why £10 is worth less to the Chief Executive of Ford than to a pauper). Specifically, Daniel

Bernoulli suggested that choice depends on the subjective value of goods (*u*), which leads to models of choice based on expected utility (*u* x *p*, where *u* is utility). This concept lies at the heart of the Expected Utility Theory (EUT) formulated by von Neumann and Morgenstern (1944) as part of Rational Choice Theory – and EUT has become the standard model to describe risky choice in many disciplines such as economics.

In this section I will first describe Expected Utility Theory and some of its limitations; second Prospect Theory that was designed specifically to deal with some of the limitations (particularly valence effects); third, alternative models from financial economics for choice under risk (Markowitz, 1952); and fourth, I will describe biologically-based approaches to the study of risk and valence effects on choice.

2.4.1 The standard economic model of risk: Expected Utility Theory

Expected Utility Theory (EUT) as a description of risky choice lies at the very heart of the Rational Choice Theory described by Von Neumann and Morgenstern (1944), which builds on Bernoulli's idea of expected utility rather than expected value. EUT cannot account for effects of valence on choice, but can explain risk preference. EUT is implemented, amongst other models, to analyse choice data in Chapters 4 and 5.

2.4.1.1 Expected Utility Theory: Setup and axioms

In EUT, uncertain prospects are modelled as probability distributions over a given set of outcomes. That is, the probability of each potential outcome is known to the agent as it is given as part of the description of the object. Here, I use the term prospect to refer to an object. An example of a prospect is a gamble of £10 or £0 on the toss of a coin, which is a prospect with two outcomes (£10 and £0) each with probability 0.5.

As in the case of RCT described above without risk, here the same axioms apply such that preferences must be asymmetric and negatively transitive. In addition, because EUT concerns probability distributions, two further axioms are added. Firstly, the substitution axiom of EUT (also known as the independence axiom), which relates to the principle that any state of the world that results in the same outcome regardless of one's choices can be ignored. Specifically, this asserts that if a prospect *B* is preferred to prospect *A*, then any (probability) mixture (*B, p*) must be preferred to the mixture (*A, p*). Second, the Archimedean axiom (also known as the continuity axiom) guarantees that preferences can be represented by some function that attaches a real value to each prospect. Specifically, this states that for all prospects *q, r* and *s*, where $q \succ r \succ s$, there exist numbers *α* and *β*, both from the open interval (0,1), such that $\alpha q+(1-\alpha)s \succ r \succ \beta q+(1-\beta)s$.

Calculating the expected utility of a prospect is straightforward. The utility of each outcome can be determined from the utility function, and then each utility is weighted by its probability of occurring. For illustration, consider a situation in which the agent does not yet know which state of the world will occur, and in which: there are two possible states of the world (1 and 2), with two probabilities of occurring (*p1* and *p2*), and each state will provide a potential outcome amount (*a1* and *a2*). For this example the expected utility model is specified as follows.

$$u(a1, a2) = p1*u(a1) + p2*u(a2) \qquad \text{Eq. 2.3}$$

The probabilities are given, so the only element that needs to be specified here is the utility function u(.). When choosing between prospects, individuals do so on the basis of which prospect gives them the highest expected utility.

2.4.1.2 Risk preference in Expected Utility Theory

Next we can ask how risk preferences (i.e. tastes for risk) emerge as an implicit by-product of the utility function in EUT. EUT enables us to model individuals such that they can dislike risk (i.e. risk-aversion), can be indifferent to risk (i.e. risk-neutrality) or like risk (i.e. risk-seeking).

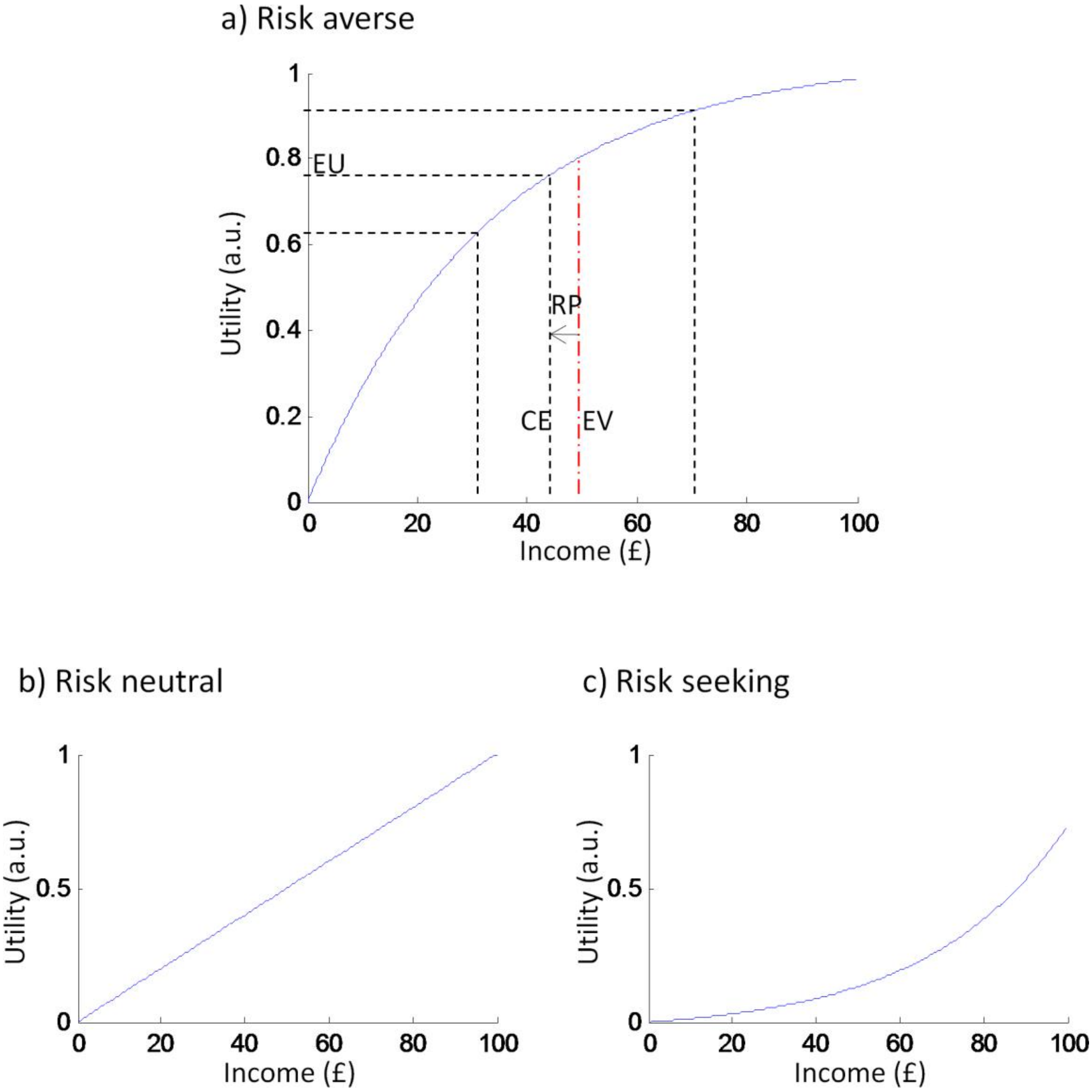

Figure 2.2 Risk preferences in Expected Utility Theory. *Panel **a)** shows how risk aversion arises as a by-product of the concave utility function. The prospect is either £70 or £30 on the toss of a coin, such that the expected value (EV) is £50. However, the utilities of £30 and £70 can be seen on the y-axis, along with the expected utility of the gamble (i.e. the average of the utilities). From the expected utility of the gamble, one can then calculate the certainty equivalent of the gamble – and this is less than the EV, with the difference being the risk premium (RP). As described in the main text, the risk averse utility function plotted here is,* $u(a) = (1\text{-}e^{-0.03a})/(1\text{-}e^{-3.3})$. *Panel **b)** shows a linear utility function, which will give risk-neutral risk preference (CE will equal EV). Panel **c)** shows a convex utility function, which will give risk-seeking risk preference (the CE of the gamble will be greater than its EV). The risk-seeking utility function plotted here is,* $u(a) = (1\text{-}e^{0.03a})/(1\text{-}e^{3.3})$.

First I will describe risk-aversion. In EUT, risk-aversion is a natural consequence of a concave utility function, such as that show in Fig. 2.2. A concave utility function seems plausible in light of the idea of diminishing marginal returns highlighted by Bernoulli (i.e. the more money you have, the less utility an extra unit of money provides). Consider such an individual with a typical concave utility function (Hey, 2003), specifically a constant absolute risk averse utility function

$$u(a) = (1-e^{-0.03a})/(1-e^{-3.3}) \qquad \text{Eq. 2.4}$$

where *a* is the argument in the utility function and *r* is the index of risk preference. The precise utility function chosen here is for illustration and many different functions would serve equally well – indeed, how risk aversion arises from any concave utilty function is simply visualised in Fig. 2.2a, which plots this specific example. The individual is offered a toss of a coin (i.e. probability of each outcome is 0.5) between £70 and £30. The expected value (EV) of the prospect is £50. However, if the individual has expected utility preferences, the evaluation is based on expected utility, rather than on expected value. Consuming £30 has utility ~0.62, whilst consuming £70 has a utility of ~0.91. The expected utility of this risky prospect is therefore (0.5*0.62 + 0.5*0.91) = 0.76 (Fig. 2.2).

We can now determine the certainty equivalent (CE) of this risky prospect, which is the amount of money that, if received with certainty, the individual regards as equivalent to the risky prospect. The utility of the certainty equivalent is the utility of the risky prospect (i.e. 0.76), which is worth approximately £44.50.

$$u(CE) = p1*u(a1) + p2*u(a2) \qquad \text{Eq. 2.5}$$

We can now also determine the risk premium, which is the maximum amount that the individual would pay to have all risk removed from the prospect. The risk premium here is £50-£44.50=£5.50. The risk premium depends on the utility function, with a more concave utility function there is a greater risk premium.

$$Risk\ premium = (p1*a1 + p2*a2) - CE = EV - CE \qquad \text{Eq. 2.6}$$

We have thus described how an individual with a concave utility function is risk-averse. Using this same utility function where *u*(*a*) is proportional to $(1-e^{ra})/(1-e^{r})$ in which *a* is the argument and *r* is the index of risk preference, we can also show how risk-neutral and risk-seeking preferences arise (Fig. 2.2). Risk-aversion arose above as the utility function was concave (equivalent to r<0 here). Risk-neutrality arises when the utility function is linear (equivalent to r=0 here), and therefore the certainty equivalent of any risky prospect is equal to its expected value. Risk-seeking arises where the utility function is convex (equivalent to r>0 here), such that the certainty equivalent of a risky prospect is greater than its expected value.

2.4.1.3 Limitations of Expected Utility Theory

The enormous influence of EUT since the 1940s is testament to its theoretical and empirical strengths. Theoretically EUT is logically internally consistent, and has normative properties such that it specifies what an agent should do (not just what she does do). Empirically, EUT captures many aspects of risk-taking behaviour in the laboratory, for example risk preferences with a variety of gambling tasks involving gain amounts (Harrison and Rustrom, 2008).

However, EUT has limitations both theoretically and as a descriptive model of behaviour. One theoretical problem is that often only a limited portion of the consumer's overall decision-problem is modelled, which can, for example, lead to portfolio effects (Kreps, 1990). A second problem concerns the validity of EUT models in complex settings such as choices over pension investments – EUT assumes unlimited rationality to understand their economic environment and the ability to perform massive calculations at no cost and instantaneously. The implausibility of unlimited rationality led Herbert Simon, for example, to propose the concept of bounded rationality (Simon, 1972).

The critical question, however, is how well EUT performs as a descriptive model of choice, even in very simple settings such as those we can implement in the lab. Unfortunately, EUT cannot account for a number of important behavioural regularities, three of which are described below.

First is the Allais Paradox, described by Maurice Allais (Allais, 1953) and which played a part in determining his Nobel memorial award. It involves two choices as follows.

> *Choice 1: Choose between two gambles. The first gives a 0.33 chance of £27,500, a 0.66 chance of £24,000, and a 0.01 chance of nothing. The second gives £24,000 for certain.*
>
> *Choice 2: Choose between two gambles. The first gives a 0.33 chance of £27,500 and a 0.67 chance of nothing. The second gives a 0.34 chance of £24,000 and a 0.66 chance of nothing.*

The modal response pattern is to take the sure thing in choice 1, and the first gamble in choice 2. This violates the substitution axiom, by which individuals should ignore any state of the world that results in the same outcome regardless of one's choices (note that a 0.66 chance of winning £24,000 removed from both options in choice 1 to create choice 2).

Second is the Ellsberg paradox, which is due to Daniel Ellsberg (Ellsberg, 1961)(Ellsberg did not win a Nobel prize, but is famous for leaking the Pentagon Papers during the Vietnam War and is still an active political campaigner). It involves two choices as follows.

> *An urn contains 300 coloured marbles; 100 are red, and 200 are some mixture of blue and green. We will select a marble from the urn at random.*
>
> *You will receive £1,000 if the marble selected is of a specified colour. Would you rather that colour be red or blue?*

You will receive £1,000 if the marble selected is not of a specified colour. Would you rather blue or red?

The modal response is to choose red in both cases. However, the red balls cannot be simultaneously more and less numerous than the blue balls! Ellsberg and others since have suggested a distinction between situations with risk (objective uncertainty) and situations with uncertainty (subjective uncertainty). This “ambiguity” relates to the different types of risk discussed above, and to the “unknown unknowns” of Donald Rumsfeld.

The third example is that EUT cannot explain why people seem to make different choices when prospects involve gains than when they involve losses. These valence effects are the subject of the next subsection, as well as Chapters 4 and 5 of this thesis.

2.4.2 Prospect Theory: introducing valence effects

Prospect Theory does not seek to replace Rational Choice Theory, but instead aims to improve upon the Expected Utility Theory model of risky choice using more psychologically realistic assumptions. Indeed, Kahneman and Tversky begin their seminal 1979 paper proposing Prospect Theory by writing that *“This paper presents a critique of expected utility theory as a descriptive model of decision making under risk, and develops an alternative model, called prospect theory”*. In order to enhance the psychological plausibility of their assumptions, Kahneman and Tversky used subjects’ responses to a series of hypothetical gambles (see Tables 2.1 and 2.2 for examples). Prospect Theory has been highly influential, and indeed for this work Daniel Kahneman received a Nobel memorial prize in 2002. Central to PT is the idea that losses and gains are treated differently, or in other words that valence impacts on choice, an insight that is important to the experiments described in Chapters 4 and 5.

Decision-making in Prospect Theory has two stages. First is an "editing phase", during which prospects are modified, for example to remove common components shared by options. Importantly, during this editing phase prospects undergo "coding" in which a "reference point" is chosen against which all the amounts in the prospect is compared. During coding everything above the reference point is a "gain" and everything below the reference point is a "loss". For example, consider a prospect in which on the toss of a coin an agent receives £10 for heads and £0 for tails. If the reference point is £5, then this prospect would be a gain of £5 or a loss of £5; whilst if the reference point were £6 then this would be a gain of £4 and a loss of £6.

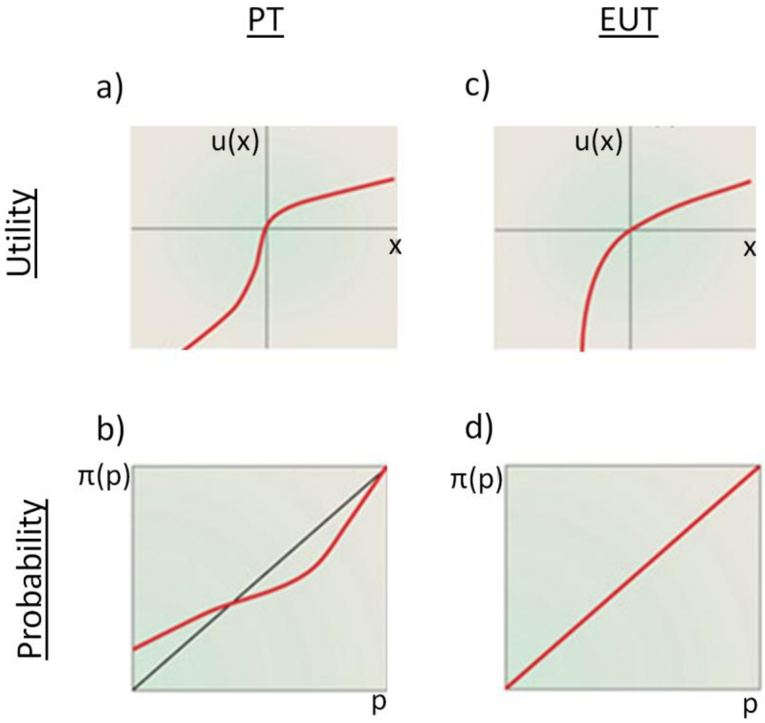

Figure 2.3 Prospect Theory utility function and probability weighting. *Panel a) shows the Prospect Theory utility function, which is concave for gains (leading to risk-aversion) and convex for losses (leading to risk seeking). Each outcome in a Prospect is assigned a utility. b) Prospect Theory employs a probability weighting function, with each objective probability, p, converted into a subjective probability, π. c) For comparison, the EUT utility function is concave throughout, and d) there is no probability weighting. Adapted from (Rangel et al., 2008)*

2.4.2.1 Evaluation in Prospect Theory: valence and probability

Following the editing phase is an "evaluation phase", during which expected utilities are calculated for each prospect. The decision-maker then chooses the prospect with the highest expected utility. The evaluation of prospects in Prospect Theory involves three components, and next I describe these as originally specified in the Kahneman and Tversky's 1979 paper proposing Prospect Theory.

The first component is the "reflection effect", which I examine experimentally in Chapters 4 and 5. This states that individuals are risk-averse with gains, but are risk-seeking with losses. To illustrate this for gains we can consider two prospects: one being a risky prospect of a gain of £10 or £0 on the toss of a coin (i.e. probability of each outcome 0.5), and the other prospect being £5 for certain (i.e. the same as the expected value of the gamble). As this decision involves gains and individuals are risk-averse for gains, then they would choose the sure option of £5. However, now consider two prospects with losses: one prospect being lose £10 or lose £0 on the toss of a coin; and the other prospect being a sure loss of £5. Individuals are risk-seeking with losses and would therefore prefer the gamble to the sure option. The reflection effect is modelled simply by having a utility function that is concave for gains (i.e. leading to risk-aversion) and convex for losses (i.e. risk-seeking) (Fig. 2.3).

The second component is "loss aversion". The idea here is that losses have a greater impact on choice than gains. Put another way, "losses loom larger than gains", so that a loss of £5 has more impact on choice than a gain of £5. Loss aversion can be illustrated by considering "gain-loss mixed gambles", which are prospects involving both gains and losses. The is illustrated by an anecdote from the influential economist Samuelson (1963), who noted that one of his colleagues prefers the status quo over a hypothetical gamble that offers an even chance to win $200 or lose $100. Follow up studies suggest that preferences of this type are rather common (e.g., Redelmeier & Tversky, 1992; Tom et al., 2007; Wedell & Bockenholt, 1994).

Loss aversion is simply modelled by rendering the utility function steeper for losses than for gains.

The third component is probability weighting. From the above, we now have a utility function that incorporates the “reflection effect” (i.e. it is concave for gains and convex for losses) and loss aversion (i.e. it is steeper for losses). Just as we did with EUT, we can therefore determine the utility of each potential outcome in a prospect – and then we must weight each outcome by the probability with which that outcome will occur. However, in Prospect Theory the probabilities are subjective, such that individuals overweight small probabilities and underweight large probabilities. This probability weighting can be achieved by having a probability weighting function such that each probability (e.g. p1, p2) is converted into a weight (e.g. π1, π2) – and it is these subjective probabilities by which the outcomes are multiplied. The precise form of probability weighting was altered by Kahneman and Tversky in their modification of Prospect Theory, called Cumulative Prospect Theory (Tversky and Kahneman, 1992).

2.4.2.2 Limitations of Prospect Theory

Prospect Theory explains choices that cannot be predicted by EUT, particularly relating to valence effects (i.e. concerning losses and gains). However, one major theoretical issue is that although the specification of a reference point is crucial, it is very difficult to know which reference point to choose in a given circumstance. For example, Kahneman and Tversky write in their 1979 paper that: *“The reference point usually corresponds to the current asset position, in which case gains and losses coincide with the actual amounts that are received or paid”*. However, others dispute this and a wide variety of methods for determining the reference point have been proposed, for example that it should be determined by rational expectations held in the recent past about outcomes (Kőszegi and Rabin, 2006). Clearly, the choice of reference point fundamentally alters the model.

Second are problems concerning the “reflection effect”, which is a focus of the experiments in Chapters 4 and 5 of this thesis. This tied relationship between risk and valence has been supported by a series of classic laboratory experiments (Kahneman and Tversky, 1979; Camerer, 1989; Battalio et al., 1990; Tversky and Kahneman, 1992). It has been used to explain important economic phenomena, such as why stock market traders hold losing stocks too long (risk-seeking) and sell winners too early (risk-aversion) (Camerer, 1998), and has been applied across diverse disciplines from international relations to political science (Levy, 2003). However, more recent findings have questioned this tied relationship. For example, risk-aversion for gains but risk-neutrality (not the predicted risk-seeking) with losses has been reported (Laury and Holt, 2005).

Finally, although loss aversion has been supported by many experiments, such as those involving mixed gambles cited above, recent work has suggested loss aversion can be attenuated by changing the format of gambles (Ert and Erev, 2008; discussed further in Chapters 4 and 5).

2.4.3 “Summary statistic” models: finance and foraging

In addition to EUT and Prospect Theory, we can also consider an alternative approach to measuring risk, which decomposes outcome distributions into “summary statistics”. Emerging shortly after EUT (Markowitz, 1952), these theories have been hugely influential in financial economics. Specifically, a distribution of outcomes can be described in terms of: the mean (i.e. expected value); variance (the dispersion of outcomes); skewness (asymmetry in outcomes); and further moments such as kurtosis. The idea here is that risk is metricated by variance (and later statistics), and that this can then be traded-off against expected value. Therefore, risk-preference can be directly generated by preference for each component.

Empirically, that humans respond to summary statistics such as variance has been shown in psychological experiments (Coombs and Pruitt, 1960; Coombs and

Huang, 1970). These summary statistic ideas have also been influential in ecological theory (Stephens, 1981). Further, these ideas have been borne out in animal behaviour experiments that have manipulated the trade-off between the mean and the variance during foraging (Real L, Ott J, Silverfine 1982; Kacelnik A, Bateson, 1996). It might provide a useful heuristic in natural stochastic environments, where it is difficult to encode each possible outcome or state of the world rapidly and with fidelity. Tracking summary statistics is also helpful for learning, as it is computationally much easier to update these estimates rather than each outcome and its associated probability separately (d' Acremont and Bossaerts, 2008).

However, there are two important limitations to this approach. Firstly, valence effects are not included in these models. Second, observations of behaviour alone cannot distinguish Expected Utility from summary statistic models since both theories make identical choice predictions, as any utility function can be approximated by preferences for summary statistics using a polynomial expansion (Scott and Horvath, 1980).

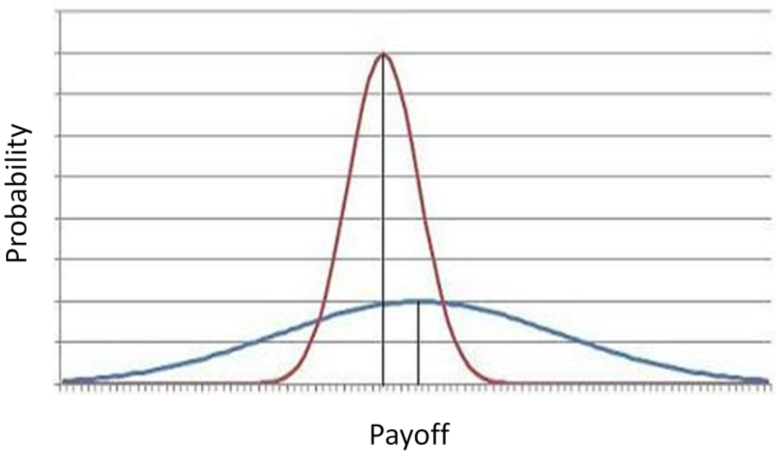

Figure 2.4 "Summary statistic" model with mean-variance trade-off. *This plot illustrates two different prospects: stocks in blue, bonds in red. The stocks have a higher mean payoff (EV), but are more risky (higher variance).*

2.4.4 Biologically-based approaches to risk and valence

Considerable biological work has examined how risk and valence influence choice, and here I review the evidence regarding each in turn.

2.4.4.1 Risk sensitive behaviour in non-human animals

Risk is ubiquitous in natural environments, and risk-sensitivity reflects a phylogenetically conserved adaptation, where maintenance of adequate nutrition and energy stores in the face of this environmental variability is critical for survival and reproduction (Real et al., 1982; Barnard and Brown, 1985; Wunderle et al., 1987; Croy and Hughes, 1991; Kacelnik and Bateson, 1996). Consistent with the idea of marginally decreasing utility (captured in EUT by a concave utility function), typically non-human animals also tend to exhibit risk aversion, which has for example been reported for animals as diverse as fish, birds and bumblebees (Stephens and Krebs, 1987; Kacelnik and Bateson, 1996).

2.4.4.2 Risk-related brain regions

Early neuroimaging studies of risky decision-making in humans identified a wide variety of cortical and subcortical structures associated with risk (Critchley et al., 2001; Paulus et al., 2001, 2003; Huettel et al., 2005; Leland and Paulus, 2005). These areas have been summarised in a recent meta-analysis of human neuroimaging studies of risk, which reported risk involving bilateral anterior insula, thalamus, dmPFC, right dlPFC, right parietal cortex, left precentral gyrus and occipital cortex (Mohr et al., 2010).

To begin parsing the contributions of these regions to the process of choice, this meta-analysis (Mohr et al., 2010) also sought to distinguish risk-related activity in two situations: “decision risk” where processing occurred before or during choice, such that it is likely to be used to guide choice; and “anticipation risk” where processing occurred after or without a choice, such that it is not used to guide choice. Both types of risk activated bilateral anterior insula, dmPFC and thalamus. Contrasting these types of risk revealed greater activity for anticipation risk in left anterior insula and left superior temporal gyrus; whilst greater activity for decision risk was seen in right anterior insula, dmPFC, dlPFC, parietal cortex and striatum.

Looking more specifically at parietal cortex, this region has shown enhanced activity during risky decision-making in both single unit and fMRI data (Platt and Glimcher, 1999; Huettel et al., 2005). Posterior parietal cortex has been implicated in executive control processes required for evaluation of uncertain choice options (Paulus et al., 2001; Huettel et al., 2005). Furthermore, in a recent study posterior parietal cortex activity scaled with the degree of risk measured as variance and reflected the choice of risky relative to sure options (Symmonds et al., 2011). This same region has also been associated with risky prospects involving “unknown unknowns” or ambiguity (Bach et al., 2009). The importance of parietal cortex in risky choice is perhaps not unexpected, given that it is known to express an interaction between number and space (Hubbard et al., 2005), in keeping with “summary statistic” related ideas in which risk may reflect the spread (variance) of an outcome distribution.

Anterior insula was one of the first brain regions to be specifically associated with risk (Critchley et al., 2001), and seems particularly related to risky choice (Platt and Huettel, 2008; Mohr et al., 2010) and subjective risk preference (Singer et al., 2010). Increased insula activity is often reported when individuals choose riskier over safer outcomes, for example during a “double-or-nothing” task, in which furthermore the magnitude of insula activation was greatest in individuals with higher neuroticism measures (Paulus et al., 2003). In a task involving choices between safer and riskier options, relative increases in insula activity before a decision preceded more risk-averse choices (Kuhnen and Knutson, 2005). A number of other studies have also implicated anterior insula in the promotion or inhibition of gamble selection based on an individual’s risk-preference (Christopoulos et al., 2009; Engelmann and Tamir, 2009; Xue et al., 2010). Insula activity has been shown not only to scale with risk when it has been manipulated by altering win probability (Preuschoff et al., 2006), but also more recently with prediction errors for risk (Preuschoff et al., 2008). Taken together with the idea that risk involves an important affective component (Schonberg et al., 2011), these various findings for risk in insula cortex are commensurate with suggestions that insula plays a more general role in arousal or subjective feeling states (Damasio et al., 1996; Craig, 2002, 2009).

Amongst subcortical structures, the dopaminergic system and its targets such as striatum have been implicated in risk processing. Dopaminergic midbrain neurons respond with a tonic increase in activity after cue presentation that reflects reward uncertainty, suggesting dopamine neurons may carry information about reward uncertainty (Fiorillo et al., 2003). Nucleus accumbens lesions in rats enhance risk aversion (Cardinal and Howes, 2005). Further, it is well known that Parkinson’s disease patients treated with dopamine agonists may develop compulsive gambling (Driver-Dunckley et al., 2003).

2.4.4.3 Risk: neural evidence for “summary statistic” models

We can also use this neural data to begin distinguishing between two classes of models, namely: EUT/Prospect Theory models in which each value is computed and then weighted by its probability; versus “summary statistic” models, in which the summary statistics of a prospect are first computed and then aggregated into a value signal (Preuschoff and Bossaerts, 2007; Bossaerts, 2010). Unlike in summary statistic models, under EUT/Prospect Theory models a separate processing pathway for risk is superfluous – and thus neural data can help adjudicate between these models by asking if the brain encodes the summary statistics of a decision.

In agreement with the summary statistic view, fMRI studies have shown expected value related signals in striatum (Preuschoff et al., 2006; Tobler et al., 2007), and the medial OFC (Rolls et al., 2008). In non-human primates, electrophysiological work has revealed signals corresponding to expected value in the midbrain dopamine system (Tobler et al., 2005), and lateral intraparietal cortex (Platt and Glimcher, 1999). Human fMRI has also revealed risk related signals (measured by variance) in striatum (Dreher et al., 2006; Preuschoff et al., 2006), insula (Dreher et al., 2006), lateral OFC (Tobler et al., 2007) and parietal cortex (Symmonds et al., 2011). In addition to variance, recent fMRI studies have also correlated neural activity with the skewness of outcome distributions in prefrontal areas (Symmonds et al., 2010, 2011; Wu et al., 2011).

2.4.4.4 Valence: neural bases of reward and punishment

Just as risk is an ubiquitous feature of the environment, so too is the need deal with both rewards and punishments, or in economic terms gains and losses. A crucial question here is: are rewards and punishments treated similarly by organisms, or are they treated differently? There are clearly commonalities in the systems neurally representing positive and negative values, for example in OFC and striatum (O'Doherty et al., 2004). However, considerable biological evidence suggests that aversion does not appear to be simply the mirror image of reward (Dayan and Seymour, 2009).

There are theoretical reasons for such reward/punishment asymmetries, for example arising from the different sampling biases introduced when learning from punishments relative to rewards (Dayan and Seymour, 2008). Empirical evidence also suggests that different systems subtend valuation of appetitive and aversive stimuli, as well as responses towards each. When studying aversion from physical pain, it is known that processing is subserved by specialised neural pathways (Craig, 2002) and a set of characteristic, involuntary responses. The basic representation and learning of aversive values implicates brainstem and midbrain structures, such as peri-acqueductal gray; whilst cortical structures such as anterior insula are associated with more complex representations (Seymour et al., 2007) and even subjective feeling states (Craig 2002, 2009). In terms of ascending neurotransmitter systems, dopamine is more associated with reward (Schultz et al., 1997, O'Doherty, 2004), whilst the identity and nature of the aversive opponent is only poorly understood despite suggestions that serotonin may play such a role (Dayan and Huys, 2009). Whilst risk tasks tend to use financial losses rather than pain, money is a conditioned reinforcer that through extensive experiential "training" is associated with reward – and losing money is therefore akin to removing a conditioned reinforcer, which is known to be aversive (Dayan and Seymour, 2009). Further, recent evidence suggests that loss aversion with tokens may be present in non-human primates (Chen et al., 2006).

Finally, a mechanistic perspective of particular relevance to the experiments in Chapters 4 and 5, there also seems to be a striking asymmetry in the actions triggered by stimuli of differing valences (Kim and Jung, 2006; Dayan and Seymour, 2008). Action and valence appear tied such that animals are disposed to approach appetitive stimuli and avoid aversive stimuli – associations that can be thought of in Pavlovian terms (discussed above). Indeed, new work by Guitart-Masip et al (2011) extends this to the monetary domain and shows that whilst the pairings of going to win and of not going to avoid punishment are naturally associated, the opposite pairings are much harder to learn for humans to learn.

2.4.4.5 Valence: neural responses to monetary gains and losses

That processing of both gains and losses at least partly involves a network of common neural regions, is suggested by findings with mixed gambles (Tom et al., 2007). Mixed gambles are stimuli containing possible losses along with gains, and behaviourally provide evidence for loss aversion (Redelmeier and Tversky, 1992; Tom et al., 2007). With mixed gambles, activity both increased with potential gains and decreased with potential losses in striatum, midbrain, ventral prefrontal cortex and anterior cingulate cortex – and furthermore in keeping with the idea of loss aversion these regions were more sensitive to the magnitude of losses than that of gains (Tom et al., 2007).

In contrast, a number of studies have found asymmetries between regions involved in the processing of gains and losses. For example, a recent meta-analysis of risk studies compared studies only involving gains and those involving losses (Mohr et al., 2010). Regions common to both gains and losses included in right anterior insula, dmPFC and thalamus. Asymmetries were seen with greater activity for losses than gains in left anterior insula, left STG, left preecnetral gyrus; and greater activity for gains than losses in dmPFC, dlPFC, right parietal cortex, thalamus, and occipital cortex.

With respect to gain/loss asymmetries, two regions that have been particularly associated with loss processing are anterior insula and amygdala. Anterior insula is involved in the representation of aversive stimuli (Calder et al., 2001; Seymour et al., 2007), and particularly in representing more complex aspects of aversive stimuli than might be represented in amygdala (Seymour et al., 2007). Experiments using framing effects have suggested a role for these regions (De Martino et al., 2006; Roiser et al., 2009; Guitart-Masip et al., 2010). Framing refers to the use of different descriptions of objectively identical outcomes such that individuals are more likely to perceive those outcomes as either gains or losses, depending on the frame (Tversky and Kahneman, 1981). Behaviourally, framing a sure option as a loss biased individuals to avoid that sure option and choose a gamble instead (De Martino et al., 2006), a bias that can also be elicited by aversive conditioned stimuli presented incidentally with the sure option (Guitart-Masip et al., 2010). Neurally, framing effects were accompanied by activity in amygdala (De Martino et al., 2006; Guitart-Masip et al., 2010) and anterior insula (Guitart-Masip et al., 2010). In addition to framing, further evidence suggests an asymmetric role for the amygdala: for example, humans with amygdala damage made poor decisions if the decisions involved potential gains, but not if they involved losses (Weller et al., 2007). It is perhaps puzzling that the study with mixed gambles discussed above (Tom et al., 2007) did not report anterior insula or amygdala activity even with a liberal threshold, and our data helping to reconcile these findings is discussed in Chapter 4.

2.5 Social choice

2.5.1 Overview

Having reviewed individual choices in the preceding section, I now turn to social choices in which decisions involve interactions with others. When von Neumann and Morgenstern (1944) proposed RCT and EUT they were in fact concerned primarily with social choice, as reflected by the title of the book introducing these concepts: the *Theory of Games and Economic Behaviour*. This RCT approach to social choice is called Game Theory, which has been hugely influential across diverse disciplines. It is important to distinguish games from Game Theory. Games are a taxonomy of strategic situations, such as the Prisoners' Dilemma Game described below; whilst analytical Game Theory is a mathematical derivation of what players with different cognitive capabilities are likely to do in games (Camerer, 2003).

In this thesis, I examine what people actually do in games, in Chapters 6, 7 and 8. Specifically, I focus here on two paradigm examples of human social motivations: fairness and cooperation. Fairness relates to how intentional agents should divide resources amongst potentially entitled recipients (Kahneman et al., 1986), and has long been of interest to sociologists (Homans, 1961), economists (Akerlof, 1979; Kahneman et al., 1986) and more recently neuroscientists (Sanfey et al., 2003). We define cooperation as the voluntary acting together of two or more individuals that brings about, or potentially brings about, ends that benefit one, both, or all, which are over and above the benefits arising from individualistic behaviour (Dugatkin, 1997; Brosnan and de Waal, 2002).

This section on social choice will follow the same pattern as that on individual choice above: first I will describe the RCT treatment of social choice; second the alternatives from behavioural economics (here involving the concept of "other-regarding preferences"); and third I will describe biologically-based approaches.

2.5.2 Rational Choice Theory in social choice: Game Theory

Naturally given its origins, Game Theory employs the same axioms described above. Players have a set of available actions, they possess preferences (that can be described by a utility function) and players choose the action that is at least as good as any other given their preferences.

Before beginning the following discussion, I raise one caveat. It may seem from the following discussion of cooperation and fairness that RCT has little descriptive power, but this is not the case in all games (e.g. the matching pennies game where individuals must keep their opponents guessing) (Camerer, 2003). Furthermore, even where RCT does not well predict behaviour it provides a useful starting point, providing a conceptual clarity and mathematically rigorous framework in which to consider social choice.

2.5.2.1 Cooperation: the Prisoners' Dilemma Game

The classic game exploring the tension between cooperation and self-interest is the Prisoners' Dilemma Game (Flood and Drescher, 1950). A typical description is as follows (see Fig. 2.5 for a payoff matrix). Two prisoners are brought in for questioning by the KGB and placed in separate cells. If both stay silent (i.e. cooperate), they both receive one year in prison. If they both accuse the other (i.e. defect) they both get four years in prison. If one stays silent and the other defects, the co-operator gets 10 years in prison and the defector gets off scot free.

Game Theory makes a clear prediction: the only rational thing for both players to do is defect. This because whatever the other player does, defection is superior. In Game Theoretic terms, mutual defection is the only Nash equilibrium. However, if the two players could cooperate, then they would receive a mutually more beneficial outcome (known as a Pareto optimal outcome).

What humans actually choose has been shown in literally thousands of experiments: subjects cooperate in one-shot PDGs about half the time (Kagel and Roth, 1995; Camerer, 2003). Whilst this goes against Game Theoretic prediction, we nevertheless observe that individuals respond rationally to incentives at least in part: for example lowering the temptation (*T*) or raising the sucker (*S*) payoffs increases cooperation; and when many rounds are played with the same partner, then cooperation tends to unravel towards the end as predicted in Game Theory (Kagel and Roth, 1995; Camerer, 2003).

	Cooperate	Defect
Cooperate	H, H	S, T
Defect	T, S	L, L

Note: assumes T > H > L > S

Figure 2.5 Payoff matrix describing the Prisoners' Dilemma Game. *The row player chooses either cooperate or defect, and the column player does likewise. The payoffs in each cell refer are written as: row, column.*

2.5.2.2 Fairness: the Ultimatum Game

We can also use a Game Theoretic approach to describe and analyse a simple game that assays fairness. Again there is tension between self-interest and a social motivation, fairness; and again Game Theory makes clear predictions. In the Ultimatum Game (UG) one player (the Proposer) is given an endowment (e.g. £10) and proposes a division (e.g. keep £6/offer £4) to a second player (the Responder), who can accept (both get the proposed split) or reject (both get nothing) the offer

(Güth et al., 1982). The Game Theoretic prediction is that if individuals are maximising only their own payoffs, then Responders should accept any amount however small (1 penny is better than nothing) and, knowing this, Proposers should offer as little as possible.

What do humans actually do in these situations? Proposers offer an average of 40% of the money (many offer half) and Responders reject small offers of 20% or so half the time (Camerer, 2003). These behaviours have been shown many times and across many cultures (Heinrich et al., 2004). Furthermore, a variant of the UG, called the Dictator Game, enables us to ask if Proposers in the UG make such high offers because they are "fair-minded" or because of fear of rejection. Dictator Games are UGs with the responder's ability to reject the offer removed – and here too Proposers do not offer zero, suggesting that behaviour is not only due to fear of rejections. However, again as with Prisoners' Dilemma Game above behaviour is not completely unpredictable, for example with Responders reacting to incentives such that they are more likely to accept higher offer proportions.

2.5.3 Other-regarding preferences and behavioural game theory

The behaviour described above raises a problem for Game Theory – why should individuals cooperate or care about fairness in an anonymous game where they will not see the other person again? Just as the assumptions of EUT were modified to create Prospect Theory, one approach to explaining behavioural regularities inexplicable using standard Game Theory is to improve its assumptions. This approach is part of behavioural economics, and it has been said that in contrast to analytic Game Theory described above, *"Behavioural game theory is about what players actually do. It expands analytical game theory by adding emotion, mistakes, limited foresight, doubts about how smart others are, and learning to analytical game theory. Behavioural game theory is one branch of behavioural economics, an* *approach to economics which uses psychological regularity to suggest ways to weaken rationality assumptions and extend theory."* (Camerer, 2003).

One way to improve the assumptions of Game Theory is to invoke the concept of "other-regarding preferences" (Fehr and Camerer, 2007). For example, in a game between me and you, my utility function (i.e. my preferences) would include not only what I personally receive, but also by what you receive (weighted in some fashion). This can be illustrated by considering the utility function of a Responder in the Ultimatum Game. Various utility functions with other regarding preferences have been proposed (Messick and McClintock, 1968; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) that make similar predictions in the UG, and therefore we can use the following simple formulation:

$$U = x_{self} - \alpha^{*}(x_{other} - x_{self}), \qquad \alpha \geq 0 \qquad \text{Eq. 2.7}$$

where x_{self} is the amount the Proposer offered in the trial and x_{other} is the amount the Proposer keeps, α is an 'envy' parameter (reflecting a tradeoff between inequality and self interest). In this model, there is no constant term as we assume a utility of 0 represents indifference between acceptance and rejection of an offer (i.e. rejection of an offer has a utility of 0). For illustration, in a single trial as Responder, the utility of the offer is calculated by combining the self-regarding component (amount to self) and the other-regarding component (the weighted impact of inequality). This utility is then compared to the utility of rejecting (zero), with the offer being accepted if greater and rejected if lesser. Thus, if I care a lot about unfairness (i.e. have a high α) then I will reject an inequitable offer in this game.

An advantage of "other-regarding preference" models is that they well capture the tension between social motivations and self-interest. They also provide quantified metrics on a trial-by-trial basis that can be used in neuroimaging analyses that use a model-based approach as implemented in Chapter 6. However, without the addition of enormous complexity such models cannot explain critical features of social behaviour, for example how the trade-off between social and self-interested motivations is dynamically modulated between different contexts. Such dynamic modulation in response to environmental contingencies is critical for success of social animals such as humans, and is the subject of the experiments in Chapters 6, 7 and 8.

2.5.4 Biological approaches

With respect to biological approaches to social choice, we first consider fairness and then cooperation.

2.5.4.1 Fairness: biology and neural correlates

As described above, fairness relates to how intentional agents should divide resources amongst potentially entitled recipients (Kahneman et al., 1986). In the UG with money humans typically reject low, “unfair”, offers even at cost to themselves (Camerer, 2003). How might such a motivation be implemented neurally? Here, we assume choice is the outcome of processes whose neural implementation may involve social computations such as prediction errors (Behrens et al., 2008; Hampton et al., 2008). Responders in the classic UG are reported to show greater activity in anterior insula and dorsolateral prefrontal cortex (DLPFC) for lower compared to higher offers, a finding interpreted as reflecting fairness and cognitive-control respectively (Sanfey et al., 2003). Alternative approaches have endeavoured to isolate components of fairness in the UG. One attempt to unconfound fairness from offer amount treated it as synonymous with offered endowment proportion, implicating lateral PFC in cognitive control (Tabibnia et al., 2008). An alternative strategy manipulated the stimuli used, where by changing Proposer intentionality anterior insula cortex was implicated in fairness responses (Güroğlu et al., 2010). Outside the UG framework others have investigated reward comparison (Fliessbach et al., 2007) and fairness in third-party decisions (Hsu et al., 2008), with the latter demonstrating that posterior (but not anterior) insula tracked an objective measure of fairness, namely inequality.

Specifically with respect to insula involvement in fairness, the precise role of different regions within this extensive (over 5cm long) and cytoarchitectonically diverse cortical region (Flynn, 1999; Varnavas and Grand, 1999) is relatively poorly understood. Hsu and colleagues (2008) asked subjects to choose between distributions of meals for African children, varying in inequality and amount, which resulted in posterior insula activity negatively correlating with inequality. Anterior insula activity has been reported as higher for rejected versus accepted offers in the UG (Sanfey et al, 2003), a result replicated in a task-matched study (Halko et al., 2009), although the same contrast in other UG studies shows little activity in this region (Guroglu et al. 2010; Tabibnia et al 2008; this study). Indeed, recent work shows anterior insula activity depends on Proposer intentionality in the UG (Guroglu et al., 2010). Finally, anterior insula activity in some UG studies may reflect processing of disgust (Sanfey et al., 2003) or aversion to norm-violation (Guroglu et al., 2010), consistent with its role in introspective awareness of emotion (Craig 2009).

With respect to the role of dlPFC in the UG, previous work has suggested a role in cognitive control (Sanfey et al. 2003; Knoch et al. 2006). Such ideas would be in keeping with broader evidence concerning dlPFC in executive function (Miller and Cohen, 2001). However, previous findings in the UG have been difficult to reconcile, as whilst bilateral DLPFC activity was seen with fMRI for lower, compared to higher, offer proportions (Sanfey et al. 2003), rTMS to right (but not left) DLPFC increased acceptance of lower proportion offers (Knoch et al. 2006). We revisit this debate about the role of dlPFC in Chapter 6, in light of our neural data.

2.5.4.2 Cooperation: hormonal modulation

There is currently much interest in the biological factors that modulate the trade-off between cooperative and more self-motivated behaviour, but in line with influential theory (Gintis et al., 2005) the focus has been on factors increasing a propensity to cooperate. Cooperative behaviours are thought to co-opt neural reward mechanisms (Rilling et al., 2002; Phan et al., 2010) and are causally promoted by the hormone oxytocin (De Dreu et al., 2010). Oxytocin has also been shown to increase measures of trust in an economic Trust Game (Kosfeld et al., 2005).

However, it is less well understood whether opponent endocrine influences exist to promote more self-orientated behaviour and reduce cooperation. One potential endocrine opponent modulator is the androgen hormone testosterone. This gonadal hormone is secreted in men and women and modulates a range of behavioural trade-offs, for example the trade-off between parenting and courtship in birds (Wingfield et al., 1990; Ketterson and Nolan, 1994), rodents (Clark and Galef, 1999) and rural Senegalese men (Alvergne et al., 2009). Socially, higher testosterone correlates with antisocial behaviour in female prisoners (Dabbs and Hargrove, 1997), while a role in fairness-related behaviours is suggested by findings from a bargaining game (Burnham, 2007), although in this bargaining paradigm administration of testosterone has provided mixed results (Zethraeus et al., 2009; Eisenegger et al., 2010).

In addition to these social effects of testosterone, it has also been implicated in a range of non-social domains. For example, endogenous testosterone in men and women has been correlated with attention (Fontani et al., 2004) and risk-taking (Sapienza et al., 2009), as well as increasing male financial traders' profit in a risky environment (Coates and Herbert, 2008). These known associations between testosterone and reward-related processing (Coates and Herbert, 2008; Sapienza et al., 2009) render it difficult to assess any potential social effects in classic tasks such as the Prisoners' Dilemma Game discussed above that use monetary rewards. We attempt to address this difficulty with our task in Chapter 8.

Chapter 3. Methods

3.1 Functional magnetic resonance imaging

Functional magnetic resonance imaging (fMRI) is a non-invasive method of brain imaging. fMRI measures local changes in cerebral blood flow that are well known to be tightly coupled to underlying neural activity, although the precise nature of this neurovascular coupling is still an active field of research (Logothetis, 2008). The main advantages of fMRI are safety and ability to image the whole brain with high spatial resolution, whilst its main limitation is a relatively poor temporal resolution, in the order of seconds.

fMRI can be used to investigate two fundamental principles of functional organisation in the brain: functional integration and functional specialisation (Friston, 2004). Functional specialisation suggests that a cortical area is specialised for some aspects of perceptual or motor processing, and that this specialisation is anatomically segregated within the cortex. A single function may then involve many specialised areas that are mediated by functional integration between them. In this thesis, fMRI is used to ask questions concerning functional specialisation.

In this chapter, first I will describe the physical principles underlying fMRI; second, the relationship between neural activity and fMRI images; third the pre-processing steps necessary to prepare fMRI data for statistical analysis; and finally the statistical analysis enabling inference about neural activity.

3.1.1 Principles of fMRI

3.1.1.1 MR signal generation

Under normal conditions, thermal energy causes the single proton in a hydrogen nucleus to spin about itself (Fig. 3.1) (Jezzard et al., 2003; Huettel et al., 2008). This spin has two effects. First, as the proton carries a positive charge its spin generates an electrical current that induces a torque when placed in a magnetic field, called the magnetic moment. Second, because the proton has an odd-numbered atomic mass, its spin also results in an angular momentum. If a nucleus has both a magnetic momentum and angular momentum it is said to have the nuclear magnetic resonance (NMR) property, and it is useful for MRI. Such nuclei can be referred to as spins.

In the absence of a strong magnetic field, the spins of the hydrogen protons are orientated randomly and tend to cancel each other out. However, when placed within an external magnetic field protons change their orientation, initiating a gyroscopic motion known as precession. Protons precess about an axis determined by the magnetic field. Precessing protons can be in two states: parallel to the magnetic field, in which they have a lower energy level; and anti-parallel, in which they have a higher energy level. The parallel, low-energy state is slightly more stable so there will be more protons in the parallel than anti-parallel state, with the relative proportion of the two states dependent on the temperature and the strength of the magnetic field. In this thesis, all fMRI experiments were conducted in scanners using a 3 Tesla (T) static magnetic field.

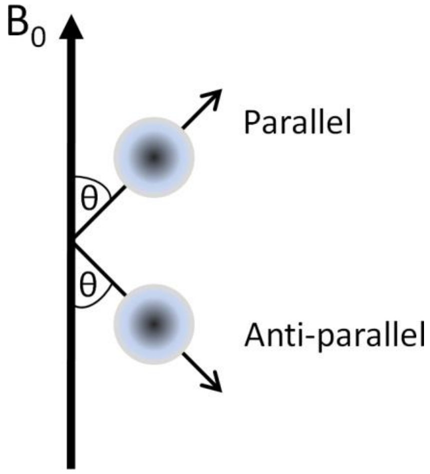

Figure 3.1 Magnetic spin. *Precessing hydrogen nuclei can be in either a parallel (low-energy) or anti-parallel (high-energy) states relative to the static magnetic field (B_0).*

When a spin in the high energy state falls into the low energy state, it emits a photon with energy equal to the energy difference between the two states. Conversely, a spin in the low energy state can jump to the high energy state by absorbing a photon with energy matching the energy difference between the two states. For a given atomic nucleus and magnetic field strength, we can calculate the frequency of the electromagnetic radiation needed to make spins change from one state to another, which is known as the Larmor frequency. These properties are used in MRI. Within MRI scanners, radiofrequency (RF) coils bombard spins in the magnetic field with photons, which are absorbed by some protons – and these are subsequently released to re-establish the equilibrium proportions of the energy states. This decaying signal can be detected by a receiver RF coil, and the signal depends on the molecular environment of the spins. By analysing this time varying signal we can learn the properties of the spins and their surrounding environment.

The process by which an MR signal, created by an excitation pulse, decays over time is known as spin relaxation. This generally occurs within a few seconds. Two primary mechanisms contribute here, namely longitudinal relaxation and transverse relaxation. For a given substance (e.g. water or fat) in a magnetic field of given strength, the rates of longitudinal and transverse relaxation are given as time constants. When the excitation pulse finishes, protons in the high energy (anti-parallel) state go back to their low energy (parallel) state: this is known as longitudinal relaxation, and the time constant associated with the longitudinal relaxation is called T1. The excitation pulse also causes coherence between spins precessing around the main field vector, as they begin their precession within the transverse plane at the same starting point. Over time, the coherence between the spins is lost and they become out of phase: this is known as transverse relaxation and the gradual loss of this coherence is characterised by a time constant T2. In addition to the T2 decay, which is caused by intrinsic spin-spin interactions, field inhomogeneities can also lead to a loss of coherence – and the combined effects of both causes lead to signal loss known as T2* decay, which is characterised by the time constant T2*.

3.1.1.2 MR image formation

We wish to acquire a three dimensional image that depicts the spatial distribution of some property of the spins, for example the T1, T2 or T2* relaxation times of the tissues in which they reside. We are able to form such an image using another field type within the MRI scanner: a magnetic gradient. The precession frequency of a spin within a magnetic field (i.e. the Larmor frequency) is determined by the magnetic field strength, which determines both the frequency of electromagnetic radiation needed during excitation to make spins change to a high energy state, and the frequency emitted by spins when they return to the low-energy state. Application of a magnetic field that varies linearly across space will cause spins at different locations to precess at different frequencies.

To generate our three dimensional image we apply three magnetic fields arranged orthogonally along the following axes: the z axis (usually superior-inferior with respect to brain anatomy; this is known as the “slice select” gradient); the x axis (left-right anatomically; known as the “frequency-encoding” or “readout” gradient); and the y axis (posterior-anterior anatomically; known as the “phase-encoding” gradient). These gradients can be stepped, enabling the partition of the image into three dimensional volume elements (voxels), for example 3x3x3mm in the functional data reported in this thesis. The size of the voxels determines the spatial resolution that can be achieved, along with anatomical constraints (discussed below).

The signal acquisition process is accomplished in two steps: first a slice is selected within the total imaging volume; and second a two dimensional encoding scheme is used within that slice to resolve the spatial distribution of the spin magnetisations. To achieve slice selection we introduce a static gradient along the slice selection axis (i.e. the z axis), such that we can exclusively tune only those spins in the slice to match the frequency of the excitation pulse. Once the spins are excited within the desired slice, they can be spatially encoded so that the MR signal from different parts of the slice can be resolved. In a typical anatomical imaging sequence this is achieved one line at a time within the slice, with each line following one of a succession of individual excitation pulses. Specifically, initially to excite the desired slice the RF and "slice select" (z-axis) field are used; then before the data acquisition period, the y-axis "phase-encoding" gradient is switched on to move the effective location of data acquisition along the y-axis (i.e. to a new line within the slice); and finally, data acquisition for that line begins, during which the x-axis "readout gradient" is switched on. Following acquisition of the slice data, a two-dimensional inverse Fourier transform can convert the raw data into image space.

3.1.1.3 MR scan types

It is possible to optimise the scanning parameters to acquire different types of data. Two such parameters govern the time at which MR images are collected: first the repetition time (TR), which is the time interval between successive excitation pulses; and second is the echo time (TE), which is the time interval between excitation and data acquisition.

An example of such optimisation in this thesis is the collection of T1-weighted structural images, for which we want to optimise our ability to differentiate between different tissues. At very short TRs there is no time for longitudinal magnetization (which is related to T1) to recover, whilst at very long TRs longitudinal magnetization recovers similarly for different tissues – and therefore an intermediate TR (e.g. 400msecs) will enable maximum differentiation between tissues. We must also have a very short TE (e.g. 20msecs) to minimise T2 contrast and have exclusive T1 contrast. All structural scans reported in this thesis are T1 weighted.

A key challenge in the acquisition of functional images is that images must be acquired very rapidly, typically every 2-3 seconds. To achieve this we use Echo-planar imaging (EPI), developed in the 1970s by Peter Mansfield. This technique allows the collection of an entire slice by changing spatial gradients rapidly following a single RF pulse, which allows a back and forth trajectory to be used to acquire the slice.

3.1.2 Blood-oxygenation-level dependent (BOLD) fMRI

Haemoglobin (Hb) is an iron-containing protein found in red blood cells, which transports oxygen in the blood in order to meet the metabolic demands of tissues in the body. The magnetic properties of Hb differ according to whether it is bound to oxygen (oxyHb), in which case it has zero magnetic moment, or whether it is not bound to oxygen (deoxyHb), in which case it is paramagnetic. The presence of deoxyHb therefore affects magnetic susceptibility, which in turn affects the T2* time constant described above. Blood-oxygenation-level dependent (BOLD) contrast is the difference in signal on T2*-weighted images as a function of the amount of deoxyHb. Work in both animals (Ogawa et al., 1990) and humans (Ogawa et al., 1992) has demonstrated that this BOLD contrast can be reliably detected.

How changes in BOLD relate to neural activity can be characterised by the haemodynamic response function (HRF; Fig. 3.2). First, increased neuronal activity increases metabolic demand and transiently increases the concentration of deoxyHb in the local vasculature, which may cause an “initial dip” in the BOLD response (Menon et al., 1995; Duong et al., 2001). Second, this is followed after a delay of around 1-2 seconds by a large increase in local blood flow, which peaks at around 6 seconds after onset of activity. This MR signal increase during neuronal activity occurs because more oxygen is supplied to the brain region than is consumed, with excess oxygenated blood flowing though active regions flushing the deoxygenated blood from the capillaries supporting the active neural tissue and from downstream venules. Third, following the “initial dip” and “peak” there is an undershoot that lasts for several seconds. fMRI relies upon identifying the clear peak in the BOLD response as the “initial dip” is smaller and difficult to identify (Heeger and Ress, 2002).

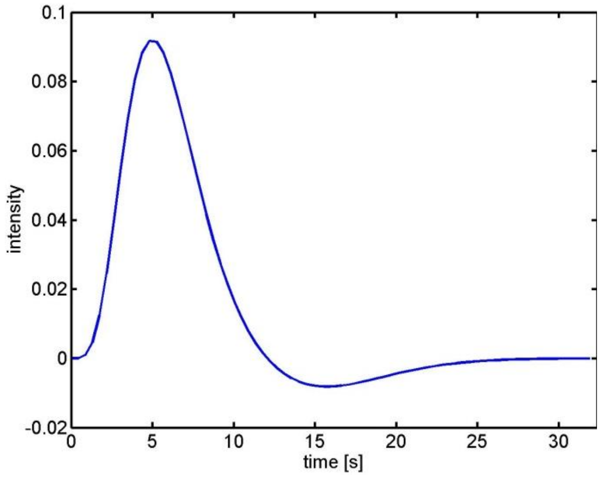

Figure 3.2 Canonical haemodynamic response function.

The relationship between the BOLD response and specific patterns of neural activity is still an area of active research (Logothetis, 2008). BOLD appears to be more related to inputs to cortical regions (i.e. synaptic activity) rather than outputs (i.e. cell firing) (Heeger and Ress, 2002; Logothetis, 2008). However, for example the relationship of the BOLD signal to inhibitory relative to excitatory activity remains unclear (Logothetis, 2008), as does the interpretation of decreases in BOLD signal relative to a resting baseline (Lin et al., 2011).

Further characteristics of the HRF also impact on the interpretation and acquisition of fMRI data. Temporal resolution is limited by the delayed nature of the HRF peak described above. The extended temporal nature of the HRF also requires that the length of each trial is not the same as the TR or a multiple of the TR in order to ensure adequate sampling of the haemodynamic response (Frackowiak et al., 2004). Spatial resolution is limited by the geometry of the cerebral microvasculature, with the spatial scale of the haemodynamic response being about 2-5mm according to high resolution optical imaging experiments (Friston, 2004). The HRF also varies between areas and between subjects (Handwerker et al., 2004).

3.1.3 Preprocessing of fMRI data

The measured BOLD signal change is small compared with the total intensity of the MR signal, in the order of a few percent. Furthermore, there are multiple sources of noise in the data, including: artefacts from head movement, heart rate or respiration; thermal noise; system noise from imperfections in scanner hardware; and variability in neuronal activity associated with non-task-related brain processes. Therefore, fMRI data undergo a number of preprocessing steps to improve the signal to noise ratio. Analyses reported here were carried out in either SPM 5 (Chapter 6) or SPM 8 (Chapter 4) (Wellcome Trust Centre for Neuroimaging, www.fil.ion.ucl.ac.uk/spm). The preprocessing steps are summarised in Fig 3.3.

To further improve the signal to noise ratio, prior to preprocessing the first few images were removed to allow for T1-equilibriation (6 such “dummies” were removed in all experiments in this thesis). Additionally, all experiments reported here were conducted at 3 Tesla (T), with higher static field strength known to improve signal to noise ratio (Wright et al., 2008).

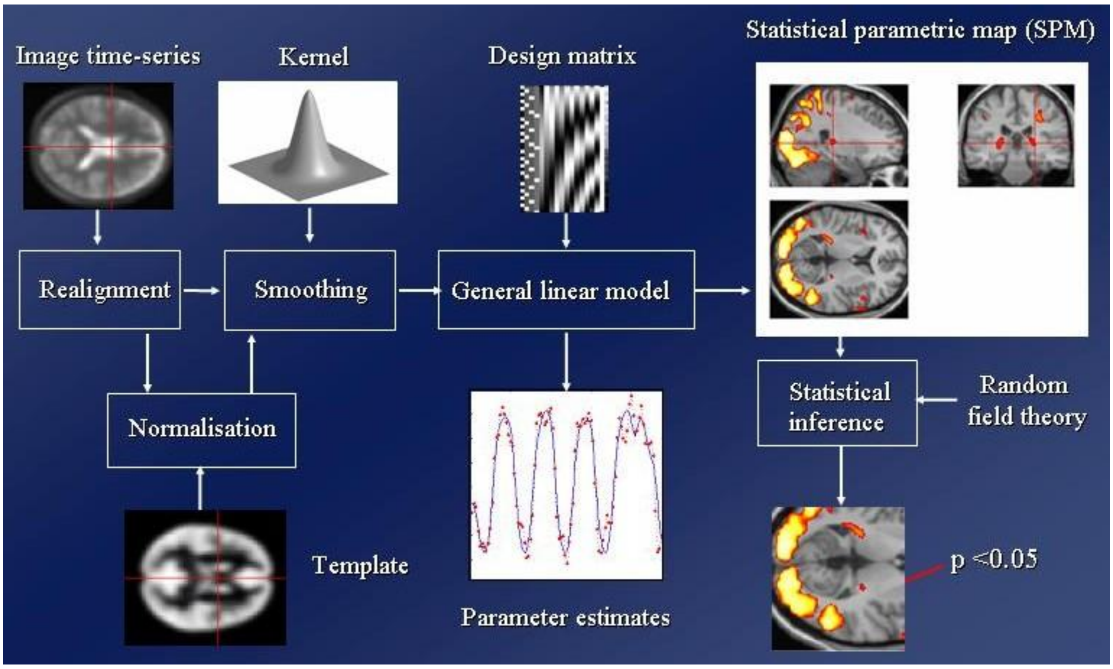

Figure 3.3 Overview of pre-processing and statistical analysis *This schematic depicts the transformations that start with an imaging data sequence and end with a statistical parametric map (SPM). During pre-processing, the data undergo realignment into the same anatomical space; are then normalised into standard space; and undergo spatial smoothing. Next, the general linear model is used to estimate the parameters of a design matrix and derive test statistic for each voxel. The test statistics (usually t or F-statistics) constitute the SPM. Finally, statistical inferences are made on the basis of the SPM and Random Field Theory. Reproduced from (Flandin and Friston, 2008).*

3.1.3.1 Realignment and unwarping

The images are first realigned into a common reference frame, in order to correct for any head movements during scanning. Head movement was also minimised during data acquisition by the use of foam pads around a participant’s head. Such realignment removes variance from a time series that would otherwise introduce error (and thus reduce sensitivity) or could, more problematically, introduce evoked effects (if movement were correlated with the cognitive task). Realignment was performed using a rigid-body affine transformation, with the reference frame as the first image in the time series. Six parameters are used in the transformation: three translations and three rotations about orthogonal axes.

However, despite realignments, considerable movement correlated variance can remain, for example that related to interactions between movements and magnetic field inhomogeneities (Andersson et al., 2001). In all studies in this thesis we militated against this problem in two ways. First, we included the movement parameters as regressors of no interest in the statistical model. Second, we use unwarping, which involved acquiring images used to estimate subject and session specific inhomogeneities in the magnetic field (called fieldmaps) – and we used these to generate a forward model of movement-by-inhomogeneity interactions (Andersson et al., 2001).

3.1.3.2 Spatial normalisation

Next, images were transformed into a standard space. This has two purposes. First, we intended to conduct analyses at the group level (see below), rendering it important to ensure that the same coordinate referred to the same anatomical area in all subjects. Second, normalisation helps standardisation and interpretation between studies. Here, normalisation was to Montreal Neurological Institute (MNI) space. Spatial normalisation was achieved by geometrically distorting each subject's brain into a standard shape (Friston et al., 1995a), which here involved the following process. First, the mean realigned and unwarped image was coregistered with the subject's T1-weighted structural image. Next, the subject's structural image was segmented into grey and white matter images and mapped onto template tissue probability maps. Finally, this mapping was applied to both the structural and functional images to create spatially normalised images.

3.1.3.3 Spatial smoothing

In the last stage of preprocessing, the fMRI data were smoothed by applying a Gaussian kernel of 8mm full width half maximum (FWHM). The motivations for smoothing the data are fourfold (Friston, 2004). Firstly, by the matched filter theorem, the optimum smoothing kernel corresponds to the size of the effect one anticipates.

Second, by the central limit theorem, smoothing the data renders the errors more normal in their distribution and ensures the validity of inferences based on parametric tests. Third, smooth data is an assumption in Gaussian random field theory (see below). Finally, when averaging across subjects, it is often necessary to smooth more (e.g. 8mm FWHM) to project the data onto a spatial scale where homologies in functional anatomy are expressed amongst subjects.

3.1.4 Statistical analysis of fMRI data

The most widespread method of analysing fMRI data is a mass univariate approach, in which the time series for each voxel is analysed independently. Within each voxel, the most common approach to analysing the time series is to use a variant of the General Linear Model (GLM, see below). The same GLM is applied to each voxel – and then the resulting statistics can be assembled into an image that is known as a statistical parametric map (SPM). Finally, using this SPM one can employ classical inference to ask if there are regionally specific effects related to the experimental factors included in the GLM. These steps are summarised in Fig 3.3. All analyses reported here were carried out in either SPM 5 (Chapter 6) or SPM 8 (Chapter 4) (Wellcome Trust Centre for Neuroimaging, www.fil.ion.ucl.ac.uk/spm).

3.1.4.1 The General Linear Model (GLM)

The GLM is an equation that expresses the observed response variable *Y* in terms of a linear combination of explanatory variables contained in a design matrix, *X,* plus an error term (Friston et al., 1995b)

$$Y = X \beta + \varepsilon \qquad \text{Eq. 3.1}$$

where *β* is a vector containing the parameters to be estimated. In this analysis of our fMRI data, for each voxel the observed response variable, *Y*, is the time series of observed BOLD signal in that voxel. Many commonly used statistical approaches are special cases of the GLM, including linear regression, t-tests and analyses of variance (ANOVAs). The GLM approach assumes that the residuals are independently and identically distributed, which is not the case for fMRI time series and therefore a correction is applied to impose sphericity (Glaser and Friston, 2004).

The design matrix, *X*, consists of columns, which are referred to as regressors. The regressors included in the design matrix represent the experimental manipulations, confounds and covariates of no interest. All experiments in this thesis use an event-related design, with the events modelled as delta (stick) or boxcar functions, which are then convolved with a canonical haemodynamic response function (Friston et al., 1998) (HRF, see Fig 3.2). Regressors in the design matrix can be categorical, such as 1 or 0 to represent the onset of a particular stimulus condition. Regressors can also be parametric, such that they modulate the height of an onset regressor. This thesis includes both types of regressors. The β parameters are then estimated using a restricted maximum likelihood algorithm.

Inferences about the effects of interest are made using the estimated β parameters (Friston, 2004). Two statistical tests are employed in my analysis. First, to test the null hypotheses that all estimates are zero, which gives an F-statistic. Second, to test the null hypothesis that some particular linear combination (e.g. a subtraction) of the estimates is zero, which gives a T-statistic. The T-statistic is obtained by dividing a contrast or compound (specified by contrast weights) of the ensuing parameter estimates by the standard error of the compound. The latter is estimated using the variance of the residuals around the least-squares fit. By applying the test at each voxel in the brain, an image of F or T statistics across the brain is produced, which is the “SPM”.

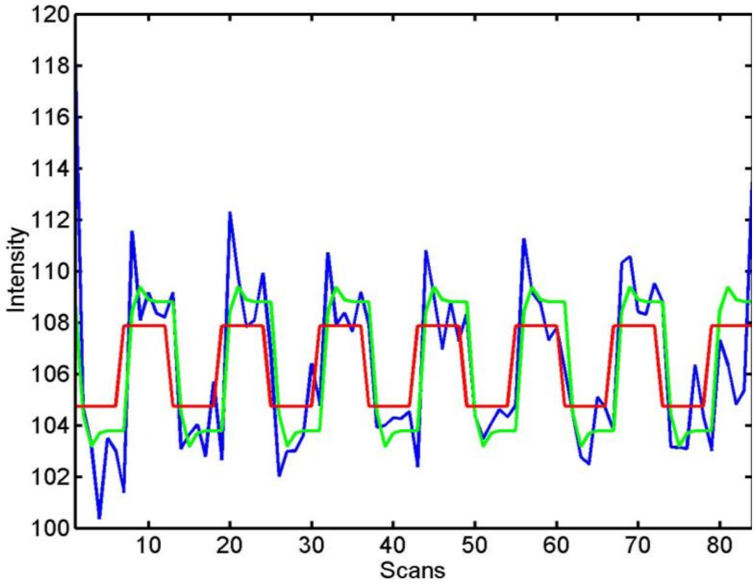

Figure 3.4 Example of regressor and convolution. *In red is the model of the stimulus, in green is the model after convolution with the canonical HRF; and in blue is the observed data.*

3.1.4.2 Multiple comparisons

The mass univariate approach described above involves many thousands of separate tests across the whole brain volume. Indeed, in a typical fMRI experiment there are approximately 20,000 voxels in the brain. If we wish to test an anatomically open hypothesis, a correction for multiple comparisons is necessary. When correcting for multiple comparisons, the adjusted type I error rate, *α*, derives from the number of independent statistical tests. One classical approach to this problem of multiple comparisons is the Bonferroni correction, in which the acceptable *α* is divided by the number of statistical tests being carried out. However, because of the large number of voxels (e.g. 20,000), whilst this minimises the chances of a type I error it also increases the chances of a type II error, that of a false negative i.e. it is an over conservative correction.

However, parameter estimates are highly correlated across adjacent voxels, for example because activity often spans large regions or large vessels and because we apply spatial smoothing. To determine a better correction factor than the number of voxels we can apply Random Field Theory, which provides a way of adjusting the p value that takes account of the fact that neighbouring voxels are not independent by virtue of continuity in the original data (Friston et al., 1995c; Friston, 2004). The strong degree of spatial correlation in functional images leads to covarying clusters of voxels (resolution elements or resells), and we can therefore control for false positives at the level of these clusters, rather than at the level of individual voxels. Random Field Theory controls the expected number of false positive regions, and because a region can contain many voxels the corrected threshold under a Random Field Theory correction is much lower than Bonferroni.

Specifically, to make inferences about regionally specific effects, the SPM is thresholded using some height and spatial extent thresholds. Corrected p-values can then be derived that pertain to: “cluster level inferences” concerning the number of activated voxels (i.e. volume) comprising a particular region; and voxel-level inferences, relating to the p-value for each voxel within that cluster. Cluster level inferences require a “cluster defining threshold”, in this thesis P<0.005 uncorrected. Although the cluster-defining threshold is somewhat arbitrary, simulations show that the assumptions behind cluster level correction hold for defining thresholds above T=2.5, and allow inferences to be based on a combination of peak height and spatial extent (Friston et al., 1993).

Whilst the above relates to anatomically open hypotheses, one can also test anatomically closed hypotheses based on a priori regions of interest (e.g. the amygdala in Chapter 4). This can be implemented using small volume correction (SVC), in which one restricts the analysis to a particular region and then applies the Random Field Theory correction. Such a region of interest can be defined in three main ways. First, as an anatomical region (e.g. the amygdala). Second, from peaks of regions identified in in previous studies (e.g. functional imaging or neuropsychological). Third, one can use orthogonal contrasts to restrict the search volume to task-relevant activations (Kriegeskorte et al., 2009; Vul and Kanwisher, 2010).

3.1.4.3 Group level analyses

The analysis described above operates at the single subject level to provide parameter estimates and contrast estimates for each subject. However, in this thesis we wish to combine data from multiple subjects and conduct analyses at the group level.

One method is to use a fixed effects analysis. This assumes that the effect of the experimental manipulation is fixed across subjects, with differences between subjects caused by random noise. However, this restricts statistical inferences to the particular sample of subjects used in the study. For example, the effects could be driven by a small proportion of the subjects, whilst the remaining subjects showed no effect.

Instead, we wish to make inferences about the population from which subjects are drawn, and therefore analyses must include information about the distribution of effect across subjects. To achieve this, a random effects analysis can be used, which treats the effect of the experimental manipulation as variable across subjects such that it could have a different effect on different subjects. This is implemented in a two-stage “summary statistic” procedure. At the “first level” (fixed effects) the contrast estimates are calculated for each individual, as described in the previous section. Then, at the “second level” (random effects) these are treated as new response variable, Y, in a GLM. The second level design matrix can be used in the same way as at the first level e.g. to test the null hypothesis that the contrasts are zero, using a column of ones and a single sample T test. At this second level, we can also look across subjects and ask whether inter-individual differences at a behavioural level covary with inter-individual differences at the neural level.

Chapter 4. Individual choice: Dissociable neural processes bias approach to risk and loss

4.1 Introduction

In Chapter 2, I outlined how the degree of risk in potential outcomes acts as a powerful influence on economic choice in humans (Harrison and Rutström, 2008) and other animals (Real et al., 1982; Barnard and Brown, 1985; Kacelnik and Bateson, 1996). This influence of risk can be captured by Expected Utility Theory (EUT). However, whether outcomes entail gains or losses (i.e. their valence) also powerfully biases behaviour (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992, 1981; Camerer, 1998) – and this is not captured by EUT.

The prevailing view in behavioural economic theory is of a tied relationship between these influences of risk and valence – the “reflection effect” – which specifies that individuals prefer riskier options with potential losses and safer options with potential gains (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). Thus, when individuals evaluate an option greater risks have more impact on choice than smaller risks, whilst valence determines whether this risk makes an option more or less desirable. This tied relationship is a foundation stone in Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) and has been supported by a series of classic laboratory experiments (Kahneman and Tversky, 1979; Camerer, 1989; Battalio et al., 1990; Tversky and Kahneman, 1992). It has been used to explain important economic phenomena, such as why stock market traders hold losing stocks too long (risk-seeking) and sell winners too early (risk-aversion) (Camerer, 1998), and has been applied across diverse disciplines from international relations to political science (Levy, 2003).

However, more recent findings have questioned this tied relationship. For example, risk-aversion for gains but risk-neutrality (not the predicted risk-seeking) with losses has been reported (Laury and Holt, 2005). Thus, our aim was to build on the insight incorporated in Prospect Theory that both valence and risk influence choice, but instead of a tied relationship seek a more general account of the relationship between these variables.

One hypothesis is that when individuals consider an economic stimulus, its valence and degree of risk independently influence choice. In other words, rather than a tied relationship, such independence would allow greater gambling with losses than with gains as classically reported (Kahneman and Tversky, 1979; Camerer, 1989; Battalio et al., 1990; Tversky and Kahneman, 1992), but also accommodate more similar gambling for each (Laury and Holt, 2005), and even the opposite finding of more gambling for gains than losses. This hypothesis also enables us to bring together insights incorporated in Prospect Theory that valence influences choice (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), with ideas derived from financial economics that individuals respond to risk as measured by the variance in potential outcomes (Markowitz, 1952; Bossaerts, 2010). Further motivating our hypothesis is mounting biological evidence that competing neural valuation systems each influence choice (Dayan, 2008; Rangel et al., 2008; Guitart-Masip et al., 2010): if risk and valence each influenced choice through distinct neural systems this would be more consistent with behavioural independence rather than tied effects. Candidate neural regions to mediate the effects of valence and risk include the orbitofontal cortex and striatum previously related to loss aversion (Tom et al., 2007), as well as insula (Preuschoff et al., 2006; Bossaerts, 2010) and parietal cortex (Platt and Glimcher, 1999; Huettel et al., 2005; Mohr et al., 2010) previously related to risk.

We tested an hypothesis of independence between biases exerted by risk and valence, by designing a new choice task that independently manipulated outcome valence (gains or losses) and degree of risk (defined as outcome variance; Figs. 4.1, 4.2). In a series of experiments, we first established that risk and valence biased choice in our paradigm (Experiment 1). Second, we then asked if these biases result from stable and independent processes (Experiment 2): we predicted that within individuals each bias should be consistent over time, but that an individual's sensitivity to one source of bias would not predict their sensitivity to the other bias. Next, we used fMRI to ask if risk and valence are processed by distinct neural systems (Experiment 3). Finally, we altered our task structure in a manner we predicted would reverse the direction of the valence-induced bias but leave the risk-induced bias unaffected (Experiment 4). We go on to suggest that behavioural and neural dissociations in the impacts of risk and valence are explicable within a biologically-based account of choice.

4.2 Materials and Methods

Our study comprised four independent experiments. In all four experiments we used a choice task that independently manipulated the degree of risk in outcomes (defined as outcome variance) and their valence (gains or losses). We developed two variants of our task, using the "accept/reject" task in Experiments 1, 2 and 3 and using the "selection" task in Experiment 4. Experiment 1 assayed behaviour in the "accept/reject" task. In Experiment 2 participants undertook the "accept/reject" task on two separate days (1-3 days apart, mean 2 days), receiving feedback and payment on the second day. In Experiment 3 participants undertook the "accept/reject" task during fMRI scanning. Experiment 4 assayed behaviour in the "selection" task. The study was approved by the Institute of Neurology (University College, London) Research Ethics Committee.

4.2.1 Participants

All participants were recruited using institutional mailing lists, were healthy and provided informed consent. 16 participants took part in Experiment 1 (mean age 26 years, range 19-70; 6 male). 28 participants took part in Experiment 2 (mean age 27 years, range 19-62; 13 male). 22 right-handed participants took part in Experiment 3 (age mean 22 years, range 18-32; 6 male), with three further participants excluded due to artefacts during acquisition of the fMRI data. 24 participants took part in Experiment 4 (age mean 23 years, range 18-34; 3 male).

4.2.2 Task

“Accept/reject” task (Experiments 1, 2 and 3): In the “accept/reject” task (Fig. 4.1) there were 200 trials presented in a random order, of which 100 were “gain trials” (all possible outcomes ≥ 0) and 100 were “loss trials” (all outcomes ≤0). In each trial participants chose to accept or reject a lottery (four possible outcomes) compared to a sure option (£6 in “gain trials”; £-6 in loss trials). Each trial began with a fixation cross presented for 1-2secs (mean 1.5secs), followed by viewing the options for 4020msec; and finally a black square appeared to indicate participants had 1500msec to input their choice by button press (the black square turned white when they chose). If participants failed to make a choice, they received zero on a “gain trial” and the maximum loss possible on a “loss trial” (£-12).

Our decision-variables of interest were risk and valence. We manipulated risk by using a set of 100 lotteries (four possible outcomes, all ≥ 0; Fig. 4.1, 4.2) in which we parametrically and orthogonally manipulated the degree of risk (variance; 10 levels) and expected value (EV; 10 levels). We presented each lottery in this set once to give 100 “gain trials”. To manipulate valence and keep all else matched including risk, we multiplied all amounts by -1 to give 100 “loss trials” (i.e. all outcomes ≤0, and a sure option of £-6).

Participants began the day with an endowment of £12. At the end of the experiment, one “gain trial” and one “loss trial” were picked at random and the outcome of both were added to the endowment to determine payment. Participants could receive between £0-24 in the task. In Experiment 2 where participants undertook the task on two separate days, they received feedback and payment on the second attendance. In Experiment 3 using fMRI, all amounts were doubled to provide a more typical level of payment for fMRI scanning.

“Selection task” (Experiment 4): This variant of our task aimed to change the route through which individuals were biased by loss aversion. It was identical to the “accept/reject” task except that, whereas on every trial in the “accept/reject” task individuals evaluated a lottery and accepted or rejected; in this new “selection” task individuals evaluated two lotteries and selected between them. To manipulate risk we again generated a set of 100 “gain trials”, in which we parametrically and orthogonally manipulated the difference in risk (10 levels of variance) and EV (10 levels) between two lotteries (each with two possible outcomes, all ≥ 0). To manipulate valence, we again simply multiplied all amounts by -1 to give 100 “loss trials”.

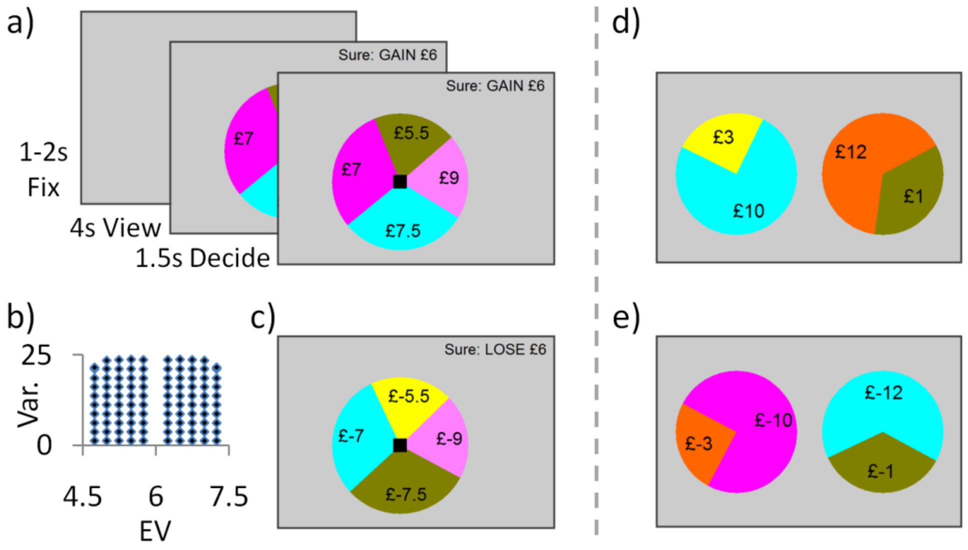

Figure 4.1 Dissociating valence and risk related biases using task design. *Panels **a-c** refer to the “accept/reject” task. **a)** In each “gain trial” individuals chose to accept a lottery (4 possible outcomes, all ≥ 0) or reject and so receive £6 for certain. **b)** We created a set of 100 “gain trials” that parametrically and orthogonally manipulated the degree of risk (defined as outcome variance; 10 levels) and expected value (EV; 10 levels) of the lotteries. Half the lotteries had an EV above the sure amount and half below, metricating risk preference as the proportion of riskier*

choices (PropRisk; risk-averse<0.5; risk-neutral=0.5; risk-seeking>0.5). ***c)*** *Multiplying all "gain trial" amounts by -1 gave 100 "loss trials" with identical parametric manipulations. All 200 trials were presented in random order. Panels d-****e*** *refer to the "selection" task, in which again there were:* ***d)*** *100 "gain trials" with parametric and orthogonal manipulation of difference in risk and EV between the two options; and* ***e)*** *100 "loss trials" created by multiplying the "gain trial" amounts by -1. However, here in each trial individuals were presented with two lotteries to consider and select between.*

4.2.3 Stimulus sets

"Accept/reject" task: For our "accept/reject" task we generated a set of 100 "gain trials" ($AR_{MainList}$), where we manipulated the difference in variance (ΔVar;10 levels) and EV (ΔEV;10 levels) of the lottery relative to the sure option of £6 (Fig. 4.2). We created this stimulus set in two stages. First, we generated a list of every possible trial within the following constraints: each lottery had four outcomes (i.e. four pie chart segments); outcomes were between £0-£12; the smallest allowable probability was 0.1, in order to militate against possible probability distortion effects at small probabilities (KT, 1979, 1992); the smallest allowable probability increment was 0.05; and we controlled for lottery skewness. Second, from within this very large number of potential trials, we selected our set of 100 trials that were the closest match to our desired 10 levels of ΔVar and 10 levels of ΔEV.

We used $AR_{MainList}$ in Experiments 1 (behavioural) and 3 (fMRI). However, to check our behavioural findings were not caused by this specific lottery set, or because we had controlled for skewness, in Experiment 2 we also compared the $AR_{MainList}$ to two alternative sets. In Experiment 2, 11 of 28 subjects used $AR_{MainList}$ (maximum ΔEV 1.25, and maximum ΔVar 23.9), and the remainder used one of two stimulus sets generated in the same way but with new lotteries and without skewness controlled (11 participants used $AR_{AlternateList1}$ [maximum ΔEV 1.35, and maximum ΔVar 23.8] and 6 participants used $AR_{AlternateList2}$ [maximum ΔEV 2.70, and maximum ΔVar 23.8]).

Importantly, the same behavioural effects were seen regardless of lottery set and therefore in our main analysis we collapse across lottery sets in Experiment 2.

“**Selection” task:** For the “selection” task we generated a set of 100 “gain trials” in the same way, although here manipulating the difference in EV (10 levels) and variance (10 levels) between two lotteries (each with two possible outcomes, ≥ 0). The difference in EV and variance between the options (maximum ΔEV 1.9 and maximum ΔVar 18.3) was similar to that used in the “accept/reject” task.

Calculation of EV, Variance and Skewness: For a given lottery with N potential outcomes *(m_1, m_2,… m_N)*, with probabilities *$p = p_1, p_2, \ldots p_N$*, we define the EV, variance (Var) and standardised skewness (Skw) of the outcome distribution as follows:

$$EV = \sum_{n=1}^{N} m_n p_n \tag{4.1}$$

$$Var = \sum_{n=1}^{N} (m_n - EV)^2 p_n \tag{4.2}$$

$$Skw = \frac{\sum_{n=1}^{N} (m_n - EV)^3 p_n}{Var^{3/2}} \tag{4.3}$$

4.2.4 Statistical analysis

All statistical tests used were two tailed.

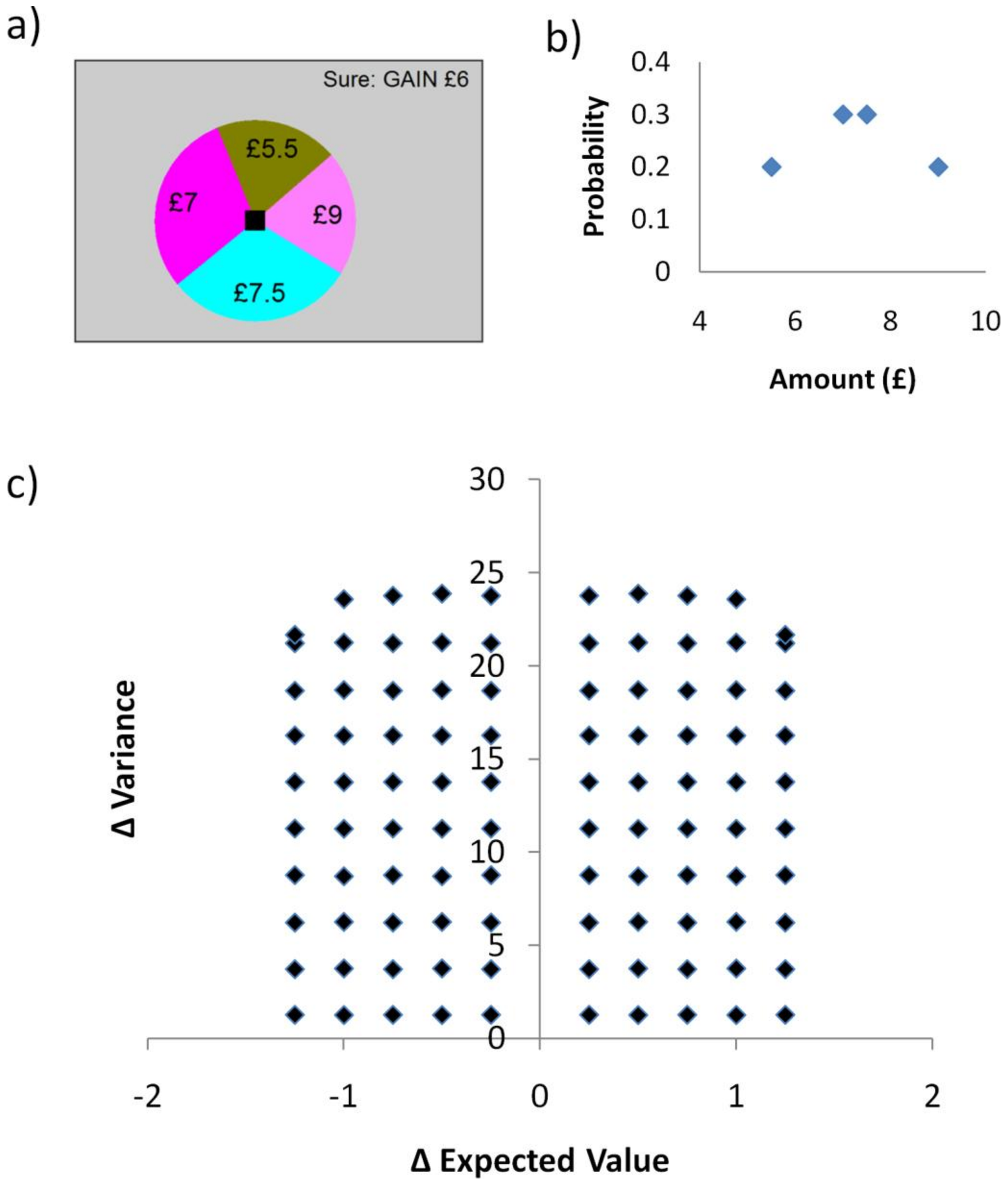

Figure 4.2 Design of the stimulus sets used in our choice task. *In each trial participants chose between two options that differed in their risk (variance) and expected value (EV). We constructed a set of 100 trials (all amounts ≥ 0) in which we parametrically and orthogonally manipulated the difference in risk (ΔVariance, 10 levels) and EV (ΔEV, 10 levels) between the two options. We presented this set of 100 trials once to give 100 "gain trials", and to create 100 "loss trials" whilst keeping all else matched we simply multiplied all amounts by -1 (not shown in this figure).* ***Panel a)*** *shows an example "gain trial" from the "accept/reject" task.* ***Panel b)*** *illustrates this lottery's outcome distribution, for which we can calculate the lottery's variance (in this case 1.3) and EV (7.25). By comparing the lottery Variance and EV to those of the other option (here £6 for sure) we determine the difference in risk (ΔVariance) and EV (ΔEV) between the two options, which could be plotted in panel* ***c****. For illustration,* ***panel c)*** *plots the stimulus set ($AR_{mainlist}$) used in the "accept/reject" task.*

4.2.5 Behavioural modelling

We used behavioural modelling of our “accept/reject” task to ask three questions: first, did both our decision-variables of interest, risk and valence, influence choice; second, can we identify a trial-by-trial metric of risk for use in our fMRI analysis; and third, can our behavioural findings be explained by probability distortion or choice randomness? We analysed the data separately from each of the three experiments using the “accept/reject” task Experiment 1 (n=16), Experiment 2 (n=28, Day 1 and Day 2), and Experiment 3 (n=22, fMRI). We also analysed the combined dataset in which we included the data from Day 1 in Experiment 2, giving a combined dataset with n=66.

In all our models, on each trial the subjective value, or utility (U), of the lottery was computed using a utility function (see below). This lottery value (U) was then compared to the value of the sure amount (*S*).

Impacts of risk and valence on choice: We compared three models to ask if behaviour was biased by risk and valence. First, in a very simple **Mean-Only model (Mn_Only)**, individuals only cared about the mean of the options.

$$U = Mean \tag{4.4}$$

Second, we asked if choice was also biased by risk, using a **Mean-variance model (Mn_Var)**. Specifically, risk is measured as variance. Here, ρ is a free parameter reflecting an individual’s preference for variance, where a risk-neutral individual has $\rho=0$, risk-averse $\rho<0$, and risk-seeking $\rho>0$.

$$U = Mean + \rho*Variance \tag{4.5}$$

Third, we asked if both risk and valence bias choice, using a **Mean-variance-valence model (Mn_Var_Val)**. There is a ρ_{gain} parameter that reflects risk preference in gain trials and a ρ_{loss} parameter reflecting risk preference in loss trials.

$$U = Mean + \rho*Variance \tag{4.6}$$

where, ρ=ρ_{gain} for Mean>0; ρ=ρ_{loss} for Mean<0;

Expected Utility model (EUT): In addition to these models described above, we also asked if our data could be explained with a standard power utility model commonly used to model expected utility (Camerer, 2003). This model incorporates the impact of risk on choice, using a free parameter, κ, that reflects the concavity of the utility function and therefore the degree of risk aversion.

$$U = \sum_{n=1}^{N} \frac{|m_n|^{1-\kappa} p_n}{1-\kappa} \qquad (4.7)$$

Prospetic model: The final utility function we tested used a model derived from Prospect Theory (Kahneman and Tversky, 1979), which in addition to the power utility function described above also incorporates the effects of valence and probability weighting. Here, the parameter λ reflects the degree of loss aversion, and the parameter π reflects probability distortion implemented with the Prelec probability weighting function (Prelec, 1998).

$$U = \sum_{n=1}^{N} \frac{|m_n|^{1-\kappa} \pi_n}{1-\kappa} \qquad \text{if } m \geq 0 \qquad (4.8)$$

$$U = \sum_{n=1}^{N} \frac{|m_n|^{1-\kappa} \pi_n \lambda}{1-\kappa} \qquad \text{if } m<0 \qquad (4.9)$$

$$\text{where} \quad \pi_n = e^{-\left(\{-\ln(p_n)\}^{\alpha}\right)} \qquad (4.10)$$

Noise in choice: In all our models, on each trial the subjective value, or utility (U), of the lottery was computed using a utility function. This lottery value (U) was then compared to the value of the sure amount (*S*) to generate a trial-by-trial probability of accepting the lottery, using a softmax function with a free parameter β (constrained between 0 and 20) that allows for noise in action selection:

$$P_{Accept} = \frac{1}{1+e^{-\beta(U-S)}} \tag{4.11}$$

Finally, we asked if valence acted by changing choice randomness. To the best fitting of the models above, we replaced the single free parameter in our softmax decision-rule with separate parameters for gain trials (β_{gain}) and loss trials (β_{loss}).

Model fitting and comparison: We fit data on an individual participant basis. We estimated best-fitting model parameters using maximum likelihood analysis. Optimisation was implemented with a non-linear Nelder-Mead simplex search algorithm in Matlab. We compared models using Group Bayes Factors, with the Bayesian Information Criterion (BIC) penalising model complexity (Schwarz, 1978).

4.2.6 Experiment 3: fMRI of the “accept/reject” task

4.2.6.1 fMRI data acquisition

Images were acquired using a 3T Allegra scanner (Siemens, Erlangen, Germany). BOLD sensitive functional images were acquired using a gradient-echo EPI sequence (46 transverse slices; TR, 2.76 secs; TE, 30 ms; 3 x 3 mm in-plane resolution; 2 mm slice thickness; 1 mm gap between adjacent slices; z-shim -0.4 mT/m; positive phase encoding direction; slice tilt -30 degrees) optimised for detecting changes in the OFC and amygdala (Weiskopf et al., 2006a). One run of 515 volumes was collected for each participant, followed by a T1-weighted anatomical scan. Local field maps were also acquired.

4.2.6.2 fMRI data analysis

Functional data were analysed using standard procedures in SPM8 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). fMRI timeseries were regressed onto a composite general linear model (GLM). The GLM contained boxcars for the length of time the lottery was displayed (5.5 seconds) to examine the decision-making process. Delta functions were also included for button presses, lottery onset to account for visual stimulus presentation, and for trials in which subjects failed to respond. We modelled our neuroimaging data using a 2 valence (gain, loss) by 2 choice (accept, reject) design. Parametric modulators were also placed on the boxcar for the EV and Variance of the lottery on that trial. The delta functions and boxcars were convolved with the canonical haemodynamic response function.

We report all activations at P<0.05 that survive whole brain correction using family-wise error at the cluster level (Friston et al., 1994), unless otherwise stated. Clusters were defined using a threshold of P <0.005 uncorrected. For presentation, images are displayed at P < 0.001 uncorrected. For the contrast of loss>gain we also used small volume correction (P<0.05) in anatomical regions of interest (amygdala and anterior insula) specified in the PickAtlas toolbox (Maldjian et al., 2003).

4.3 Results

4.3.1 Behaviour

4.3.1.1 Behaviour in the “accept/reject” task

We first established that risk and valence both biased choice within our “accept/reject” task (Experiment 1, n=16; Fig. 4.3). Our task comprised “gain trials” and “loss trials”. In each of 100 “gain trials” participants chose to accept a lottery (all outcomes ≥ 0) or reject the lottery and so receive £6 for certain. The degree of risk in the lottery, operationally defined as variance in outcomes (Bossaerts, 2010), was parametrically manipulated across the 100 gain trials. Valence was independently manipulated by multiplying all amounts in our gain trials by -1, creating 100 “loss trials” with an identical parametric manipulation of risk. We presented the 200 trials randomly ordered. Our set of lotteries orthogonally manipulated risk (variance; 10

levels) and expected value (10 levels; Fig. 4.1, 4.2), such that half the lotteries had an expected value above the sure amount and half below: providing a simple metric of risk preference indexed as the proportion of riskier choices made (*PropRisk*; risk-neutral=0.5; risk-averse<0.5; risk-seeking>0.5).

Our data showed that risk biased choice overall, with individuals being averse to risk (*PropRisk*$_{all}$ 0.40± s.d. 0.15; one-sample t-test versus risk-neutral, $t_{(15)}$=-2.9, P=0.01; Fig. 4.3d). We also extracted a simple metric for the impact of valence on choice from the difference in riskier choices in each domain (*ImpValence* = *PropRisk*$_{gain}$-*PropRisk*$_{loss}$). Individuals were also sensitive to valence (*ImpValence* 0.11±0.17; one-sample t-test versus no effect of valence, $t_{(15)}$=2.6, P=0.019). Strikingly however, against a prevailing expectation (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) individuals gambled more for gain (*PropRisk*$_{gain}$ 0.46 ± s.d.0.18) compared to loss outcomes (*PropRisk*$_{loss}$ 0.35 ± s.d.0.14; $t_{(15)}$=2.6, P=0.019).

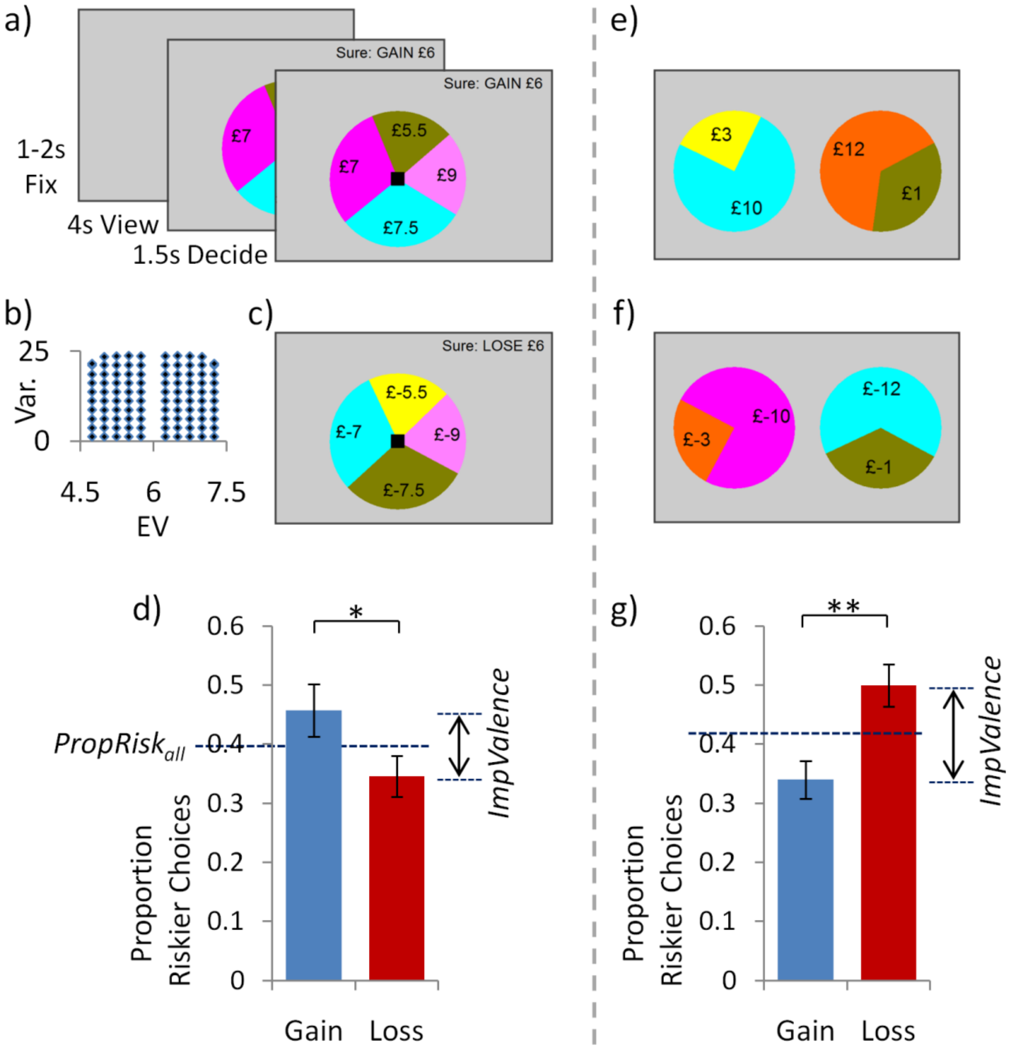

Figure 4.3 Dissociating valence and risk related biases using task design. *Panels **a-d** refer to the "accept/reject" task. **a)** In each "gain trial" individuals chose to accept a lottery (4 possible outcomes, all ≥ 0) or reject and so receive £6 for certain. **b)** We created a set of 100 "gain trials" that parametrically and orthogonally manipulated the degree of risk (defined as outcome variance; 10 levels) and expected value (EV; 10 levels) of the lotteries. Half the lotteries had an EV above the sure amount and half below, metricating risk preference as the proportion of riskier choices (PropRisk; risk-averse<0.5; risk-neutral=0.5; risk-seeking>0.5). **c)** Multiplying all "gain trial" amounts by -1 gave 100 "loss trials" with identical parametric manipulations. All 200 trials were presented in random order. **d)** Behaviour in the "accept/reject" task (Experiment 1, n=16). Individuals were risk averse overall (i.e. PropRisk$_{all}$ <0.5). Valence also biased choice, with more gambling for gains than losses (ImpValence = PropRisk$_{gain}$-PropRisk$_{loss}$). Panels **e-g** refer to the "selection" task, in which again there were: **e)** 100 "gain trials" with parametric and orthogonal manipulation of difference in risk and EV between the two options; and **f)** 100 "loss trials" created by multiplying the "gain trial" amounts by -1. However, here in each trial individuals were presented with two lotteries to consider and select between. g)*

*Behaviour in the "selection" task (Experiment 4, n=24): risk aversion overall was unaltered compared to the "accept/reject" task (i.e. PropRisk$_{all}$ <0.5), but the direction of the valence effect was completely reversed. Error bars show s.e.m., * P<0.05, ** P=0.005.*

4.3.1.2 Behaviour in the "selection" task: manipulating task design to dissociate risk and valence effects

Given that our observation of greater gambling for gains than losses differs from previously reported findings (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), we aimed to replicate those previous findings by modifying our task. One possible source for this difference is the format in which the decisions were presented, as illustrated by comparing our "accept/reject" task to the problems used in the classic paper establishing Prospect Theory (Kahneman and Tversky, 1979). In the former each trial presented a different lottery to accept or reject, whilst in the latter each problem presented two options for individuals to select between. Such a format effect is also suggested by recent work with "loss-gain mixed gambles" (gambles containing losses and gains), which were avoided more often when presented analogously to our "accept/reject" task than when presented as two options to select between (Ert and Erev, 2008).

We modified the format of our paradigm to create a new "selection" task (Fig. 4.1), aiming to selectively reverse the bias from loss aversion but leave the overall risk-induced bias unchanged relative to our "accept/reject" task. Again there were 100 "gain trials" with parametric modulation of the degree of risk between the two options in each trial, with the magnitudes of these differences similar by design to the "accept/reject" task (i.e. up to a difference in variance of approximately 20, details in SI). Valence was manipulated as before to generate 100 "loss trials". However, here in each trial individuals were presented with two lotteries to consider and select between. Again across trials we orthogonally manipulated the differences in risk (10 levels) and expected value (10 levels) between the options, giving a metric of risk preference overall as the proportion riskier choices (*PropRisk$_{all}$*), and a metric for the impact of valence from the difference in riskier choices in each domain (*ImpValence*).

As predicted, risk aversion overall was the same in the "selection" task (*PropRisk$_{all}$* 0.42±0.11) as in the "accept/reject" task (P>0.4 for independent sample t-tests against *PropRisk$_{all}$* in Experiments 1, 2 or 3, details below). Further, the magnitude of the valence effect was the same in the "selection" task (*ImpValence* -0.16±0.25) as in the "accept/reject" task (P≥0.3 for independent sample t-tests against the *ImpValence* in Experiments 1, 2 or 3, details below). However, the direction of this valence effect was completely reversed, such that now individuals selected the riskier option more for losses (*PropRisk$_{loss}$* 0.50±0.17) than gains (*PropRisk$_{gain}$* 0.34±0.16; $t_{(23)}$=3.1, P=0.005).

4.3.1.3 Behaviour: risk, valence and their relationship in all four experiments

Here we present the behavioural data from all four experiments separately. Our two decision-variables of interest were risk and valence, both of which strongly influenced behaviour in all four experiments. As shown in Figure 4.4, behaviour was strikingly consistent across the three experiments that used the "accept/reject" task (Experiments 1, 2 and 3); whilst with our "selection" task we selectively reversed the direction of the valence-induced bias but left the overall risk-induced bias unaffected (Experiment 4).

Impact of risk: In our "accept/reject" task half the lotteries had an expected value above the sure amount and half below, providing a simple metric of risk preference as the proportion of riskier choices made (*PropRisk*; risk-neutral=0.5; risk-averse<0.5; risk-seeking>0.5), which could also be used with our "selection" task. In all four experiments our participants were biased to be risk averse, choosing the risky option less than half the time overall (i.e. *PropRisk$_{all}$* <0.5). One sample ttests against the null hypothesis of risk-neutrality (i.e. *$PropRisk_{all}$* = 0.5) showed risk aversion in all datasets: Experiment 1 (*$PropRisk_{all}$* = 0.40±0.14, t(15)=-3.0, P=0.011); Experiment 2 Day 1 (*$PropRisk_{all}$* = 0.45±0.13, t(27)=-2.2, P=0.039); Experiment 2 Day 2 (*$PropRisk_{all}$* = 0.40±0.13, t(27)=-4.4, P<0.0005); Experiment 3 (*$PropRisk_{all}$* = 0.40±0.11, t(21)=-4.2, P=0.0002); and Experiment 4 (*$PropRisk_{all}$* = 0.42±0.11, t(23)=-3.7, P=0.001).

Impact of valence: We extracted a simple metric for the valence-induced bias from the difference in riskier choices in each domain (*ImpValence* = *$PropRisk_{gain}$*-*$PropRisk_{loss}$*). Valence biased choice in all four experiments, as shown by one sample ttests against the null hypothesis of no bias (i.e. *ImpValence* = 0) in all four datasets: Experiment 1 (*ImpValence* = 0.11±0.17, t(15)=2.6, P=0.019); Experiment 2 Day 1 (*ImpValence* = 0.15±0.16, t(27)=5.0, P=3.2×10^{-5}); Experiment 2 Day 2 (*ImpValence* = 0.10±0.14, t(27)=3.8, P=0.001); Experiment 3 (*ImpValence* = 0.18±0.15, t(21)=5.6, P=1.5×10^{-5}; and Experiment 4 (*ImpValence* = -0.16±0.25, t(23)=-3.1, P=0.005).

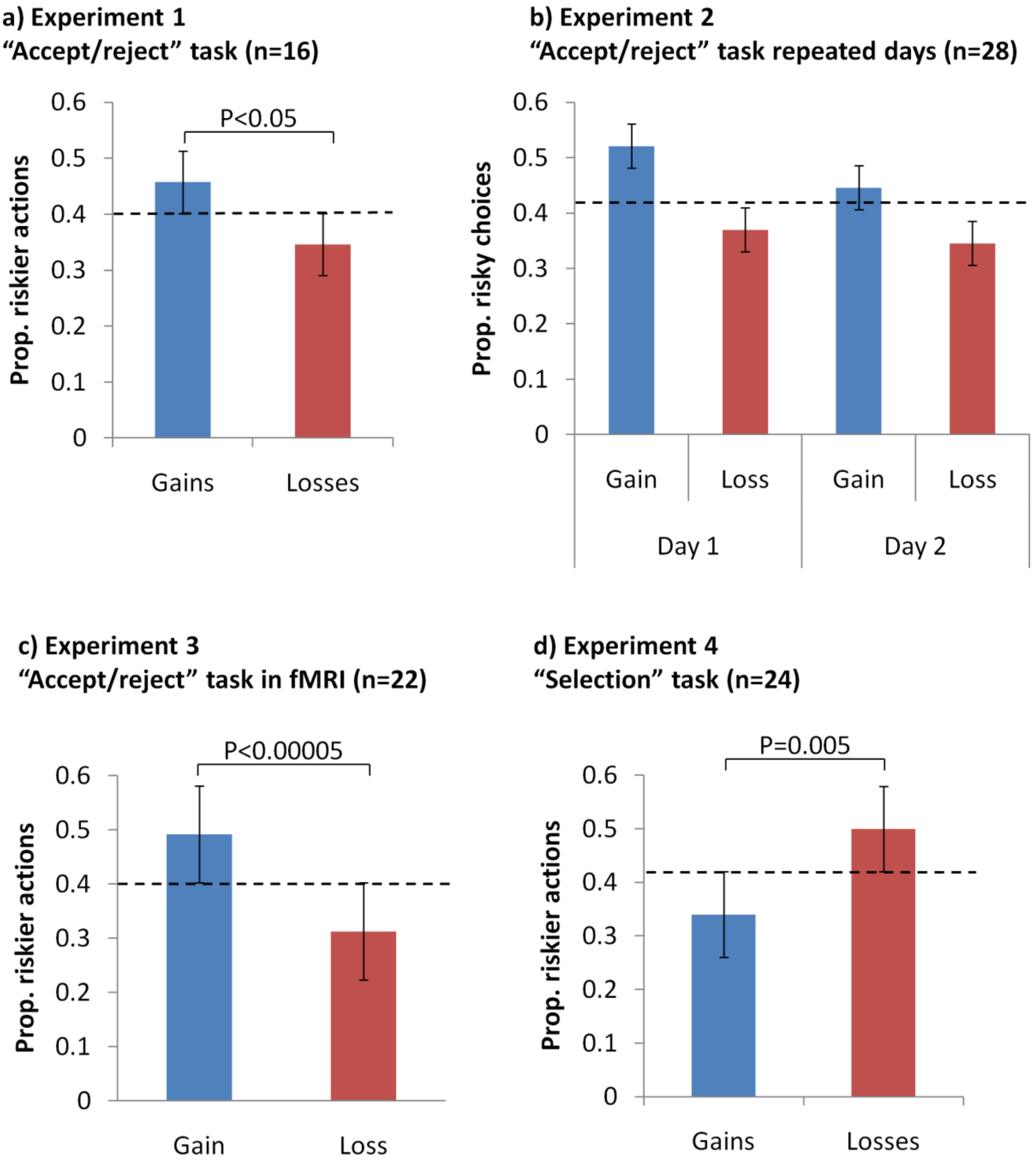

Figure 4.4 Behavioural results summary *In our “accept/reject” task (Experiments 1, 2 and 3) half the lotteries had an expected value above the sure amount and half below, providing a simple metric of risk preference as the proportion of riskier choices made (PropRisk; risk-neutral=0.5; risk-averse<0.5; risk-seeking>0.5), which could also be used with our “selection” task (Experiment 4).* ***Risk:*** *On each chart the dotted line shows the proportion of riskier choices made overall (PropRisk$_{all}$): in all four experiments participants were risk averse overall, choosing the riskier option less than half the time (i.e. PropRisk$_{all}$ <0.5).* ***Valence:*** *In each experiment there is an effect of valence, which is reversed for the “selection” task compared to the “accept/reject” task, whilst the risk-induced bias remains unaffected. We obtain the same results using parameters derived from our winning Mean-Variance-Valence model (see below).*

Relationship between the risk- and valence-induced biases: In the "accept/reject" task participants gambled more for gain than loss outcomes (Experiment 1, P=0.019; Experiment 2 Day 1, P=3.2x10^{-5}; Experiment 2 Day 2, P=0.001; Experiment 3, P=1.5x10^{-5}). In the "selection" task we reversed the direction of this valence-induced bias and showed more gambling for losses than gains (Experiment 4, P=0.005). Despite context reversing the effect of valence, context had no effect on the overall risk-induced bias (independent samples ttests comparing *PropRisk*$_{all}$ in the "selection" task to experiments using the "accept/reject" task: Experiment 1 t(38)=0.5, P=0.64; Experiment 2 Day 1 t(50)=-0.8, P=0.45; Experiment 2 Day 2 t(50)=-0.7, P=0.46; Experiment 3 t(44)=0.6, P=0.59).

This robust valence-induced bias did not result in participants becoming absolutely risk-seeking in either valence in any of the four experiments, as shown by one sample ttests against risk-neutrality (i.e. *PropRisk* = 0.5): Experiment 1 (*PropRisk*$_{gain}$ = 0.46±0.18, t(15)=-0.96, P=0.4; *PropRisk*$_{loss}$ =0.35±0.14, t(15)=-4.5, P=4.7x10^{-4}); Experiment 2 Day 1 (*PropRisk*$_{gain}$ = 0.52±0.15, t(27)=0.71, P=0.5; *PropRisk*$_{loss}$ =0.37±0.16, t(27)=-4.4, P=1.6x10^{-4}); Experiment 2 Day 2 (*PropRisk*$_{gain}$ = 0.45±0.15, t(27)=-1.89, P=0.07; *PropRisk*$_{loss}$ =0.35±0.13, t(27)=-6.1, P=1.8x10^{-6}); Experiment 3 (*PropRisk*$_{gain}$ = 0.49±0.14, t(21)=-0.30, P=0.8; *PropRisk*$_{loss}$ =0.31±0.13, t(21)=-6.7, P=1.3x10^{-6}); Experiment 4 (*PropRisk*$_{gain}$ =0.34±0.16, t(23)=-5.0, P=4.8x10^{-5}; *PropRisk*$_{loss}$ =0.50±0.17, t(23)=-0.02, P=1.0).

4.3.1.4 Stable and independent inter-individual differences for risk and valence

Having established that both risk and valence biased choice, we exploited inter-individual differences to seek evidence of their behavioural independence. If these biases result from stable and independent processes we can make two predictions: firstly, within individuals each bias should be consistent over time; and second, if they are independent then knowing an individual's sensitivity to one source of bias would not predict their sensitivity to the other bias.

We tested these conjectures in Experiment 2, where 28 participants performed the "accept/reject" task on two separate days (1-3 days apart). We found behaviour on Day 1 strongly predicted behaviour on Day 2 for both risk (*PropRisk$_{all}$* r=0.77, P=2.1x10^{-6}) and valence (*ImpValence* r=0.84, P=3.3x10^{-8}; Fig. 4.5). However, crucially these preferences were independent, with risk and valence effects showing no correlation on either Day 1 (r=-0.021, P=0.92, Fig. 4.5) or Day 2 (r=0.14, P=0.47; Fig. 4.6), or in our other datasets (Fig. 4.6).

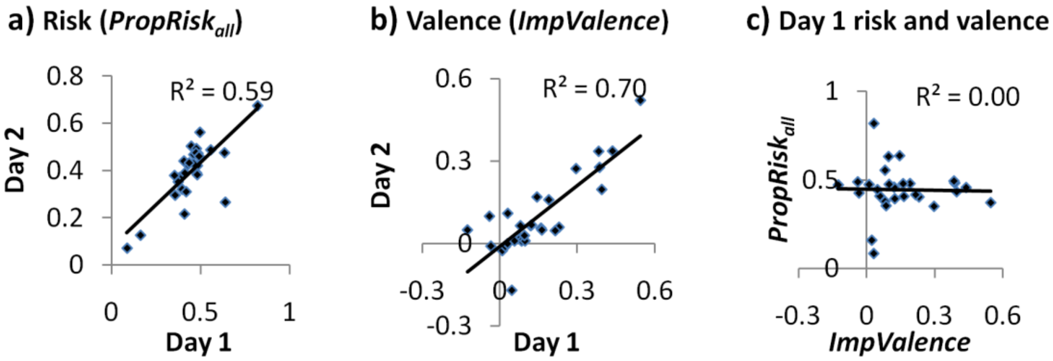

Figure 4.5 Individuals' preferences for risk and valence are consistent and independent. *In Experiment 2, 28 participants performed the "accept/reject" task on two separate days. We demonstrated a striking consistency over days in individual preferences for both risk (panel* ***a,*** *PropRisk$_{all}$ r=0.77, P=2.1x10^{-6}) and valence (panel* ***b,*** *ImpValence r=0.84, P=3.3x10^{-8}). However, crucially these preferences were independent, with risk and valence effects showing no correlation on either Day 1 (panel* ***c,*** *r=-0.021, P=0.92) or Day 2 (r=0.14, P=0.47; Fig. 4.6), or in our other datasets (Fig. 4.6).*

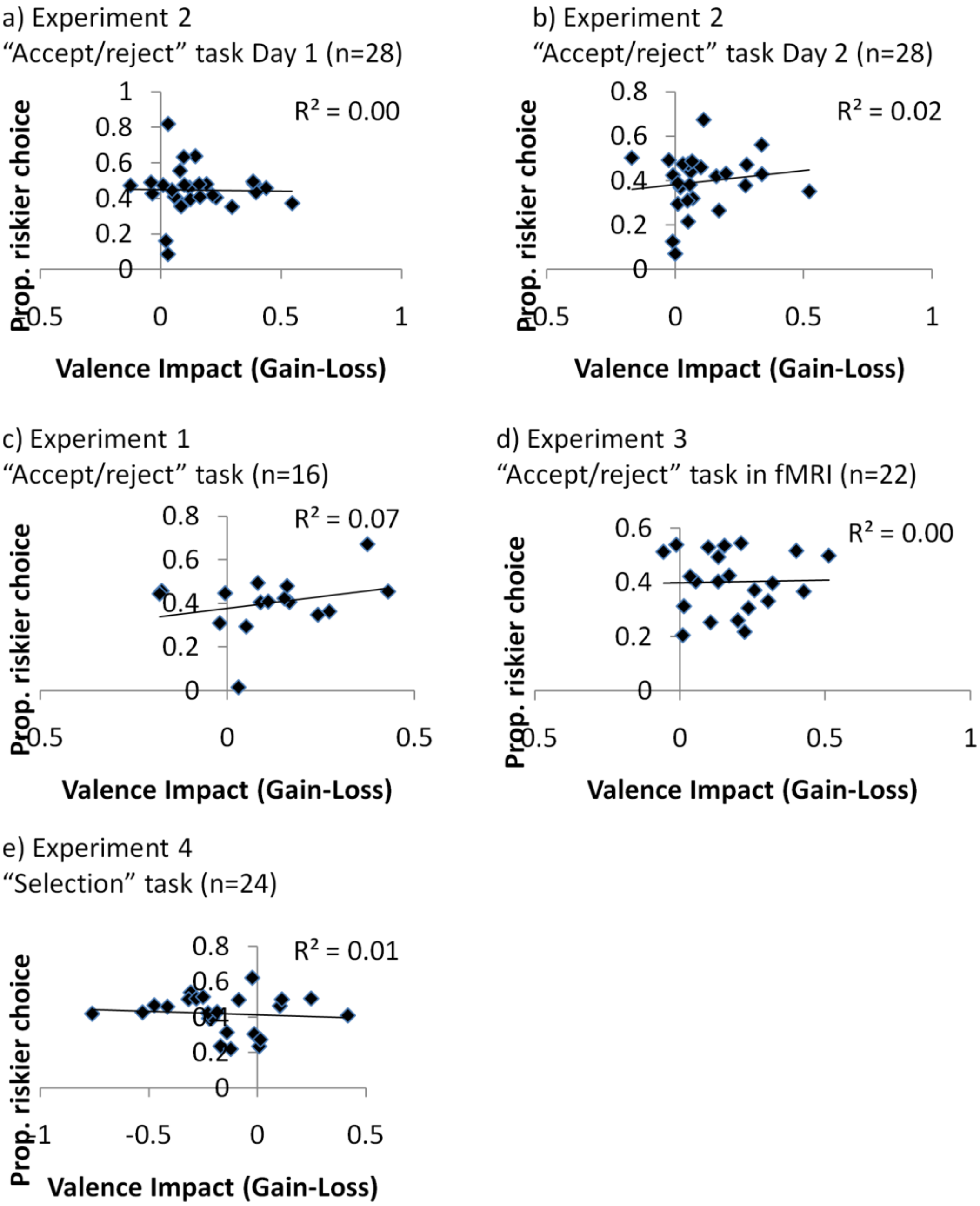

Figure 4.6 Independence in individual preferences for risk and valence. *In Experiment 2, 28 participants performed the "accept/reject" task on two separate days. First, we demonstrated that individual preferences for both risk and valence were strikingly consistent over days (Fig. 4.5). Second, independence between these preferences for risk and valence was demonstrated as they were uncorrelated on Day 1 (**panel a**) or Day 2 (**panel b**) of Experiment 2. Furthermore, we replicate this independence with the "accept/reject" task in Experiment 1 (n=16, **panel c**) and Experiment 3 (n=22, **panel d**), and also with the "selection" task in Experiment 4 (n=24, **panel e**). In this figure we demonstrate these findings using a simple metric of the risk-induced bias as the proportion of riskier choices made (PropRisk$_{all}$; risk-neutral=0.5; risk-averse<0.5; risk-seeking>0.5); and a simple metric of valence*

impact from the difference in riskier choices in each domain (ImpValence = $PropRisk_{gain}$-$PropRisk_{loss}$). We obtain the same results using parameters derived from our winning Mean-Variance-Valence model (see below).

4.3.1.5 Accept/reject" task with different stimulus sets

In Experiment 2, 28 participants performed the "accept/reject task" on two separate days. As described above, we predicted that the impact of risk and valence would be consistent over time within individuals, but that these preferences would be independent (Fig. 2). Further, to ensure our findings were not caused by the particular set of lotteries, we used three lottery sets (11 participants used $AR_{MainList}$; 11 participants used $AR_{AlternateList1}$; and 6 participants used $AR_{AlternateList2}$). As described immediately below, our results were not affected by which list was used, and therefore we collapsed across them in the analyses presented above.

Stimulus set did not alter the impacts of risk and valence: In a 2 (gains, losses) x 2 (Day1, Day2) mixed ANOVA with lottery set ($AR_{MainList}$, $AR_{AlternateList1}$, $AR_{AlternateList2}$) as a between subjects factor, lottery set did not interact with either the main effect of valence (F(2,25)=0.49, P=0.62); main effect of day (F(2,25)=0.23, P=0.12) or their interaction (F(2,25)=0.63, P=0.54). In this mixed ANOVA we found a main effect of Valence (F(1,25)=16.47, P=0.0004), no main effect of Day (F(1,25)=8.04, P=0.09), and an interaction of Valence with Day F(1,25)=7.72, P=0.01.

Independent and stable preferences for risk and valence with all three stimulus sets: We obtained the same results for all three stimulus sets, with the impacts of risk and valence being highly consistent over time within individuals, but with these biases being uncorrelated. With all three lists we found a strong correlation between the impact of risk (*$PropRisk_{mean}$*) on Day 1 and Day 2 ($AR_{MainList}$ r=0.76 P=0.007; $AR_{AlternateList1}$ r=0.88 P=4.0×10^{-4}; $AR_{AlternateList2}$ r=0.85 P=0.03) and for the impact of Valence ($AR_{MainList}$ r=0.91 P=9.0×10^{-5}; $AR_{AlternateList1}$ r=0.77 P=0.006;

$AR_{AlternateList2}$ r=0.88 P=0.022). However, there was no correlation between the measures on either day ($AR_{MainList}$ Day 1 P=0.74, Day 2 P=0.31; $AR_{AlternateList1}$ Day 1 P = 0.36 Day 2 P = 0.67, $AR_{AlternateList2}$ Day 1 = 0.49, Day 2 P=0.58).

4.3.2 Behavioural Modelling

We also conducted a model-based analysis of data from our "accept/reject" task to ask three questions: first, did both our decision-variables of interest, risk and valence, influence choice; second, could we identify a trial-by-trial metric of risk for use in our fMRI analysis; and third, could our behavioural findings be explained by probability distortion or choice randomness?

We analysed the data separately from each of the three experiments using the "accept/reject" task Experiment 1 (n=16), Experiment 2 (n=28, Day 1 and Day 2), and Experiment 3 (n=22, fMRI). We also analysed the combined dataset in which we included the data from Day 1 in Experiment 2, giving a combined dataset with n=66.

Risk and valence both influence choice: The effects of risk and valence are seen clearly by comparing our three related "summary statistic" models (Mean-Only, Mean-Variance, and Mean-Variance-Valence). The simple Mean-Only model where individuals care only about the mean value of the options (Experiment 1 Mn_Only summed BIC = 3803) is markedly improved by adding the influence of risk in the Mean-Variance model (Experiment 1 Mn_Var=3426). In turn, this Mean-Variance model is markedly improved by also accounting for valence in our Mean-Variance-Valence model (Experiment 1 Mn_Var_Val =3293), which includes separate risk parameters for each valence. We replicate these results in Experiment 2 (Day 1 Mn_Only=6028; Mn_Var =5434; Mn_Var_Val =5180; Day 2 Mn_Only=5917; Mn_Var =5087; Mn_Var_Val =4989), Experiment 3 (Mn_Only=5249; Mn_Var =4896; Mn_Var_Val =4701) and with the combined dataset (Mn_Only=15080; Mn_Var = 14571; Mn_Var_Val =13756). These marked improvements from incorporating risk and valence effects are found despite penalising for increased model complexity by using the BIC in model comparison.

The importance of risk and valence is also evident with our EUT and Prospetic models. By construction our EUT model incorporates the impact of risk, and it out-performs the Mean-Only model described above (Experiment 1 EUT = 3696). In turn, this EUT model is markedly improved by also accounting for valence in the Prospetic model (Experiment 1 Prospetic=3424). Again we replicate these results in Experiment 3 (EUT =5092; Prospetic=5067), with the combined dataset (EUT=14571; Prospetic=14128) and on the first but not second day of Experiment 2 (Day 1 EUT=5783; Prospetic=5637; Day 2 EUT=5486; Prospetic=5642),

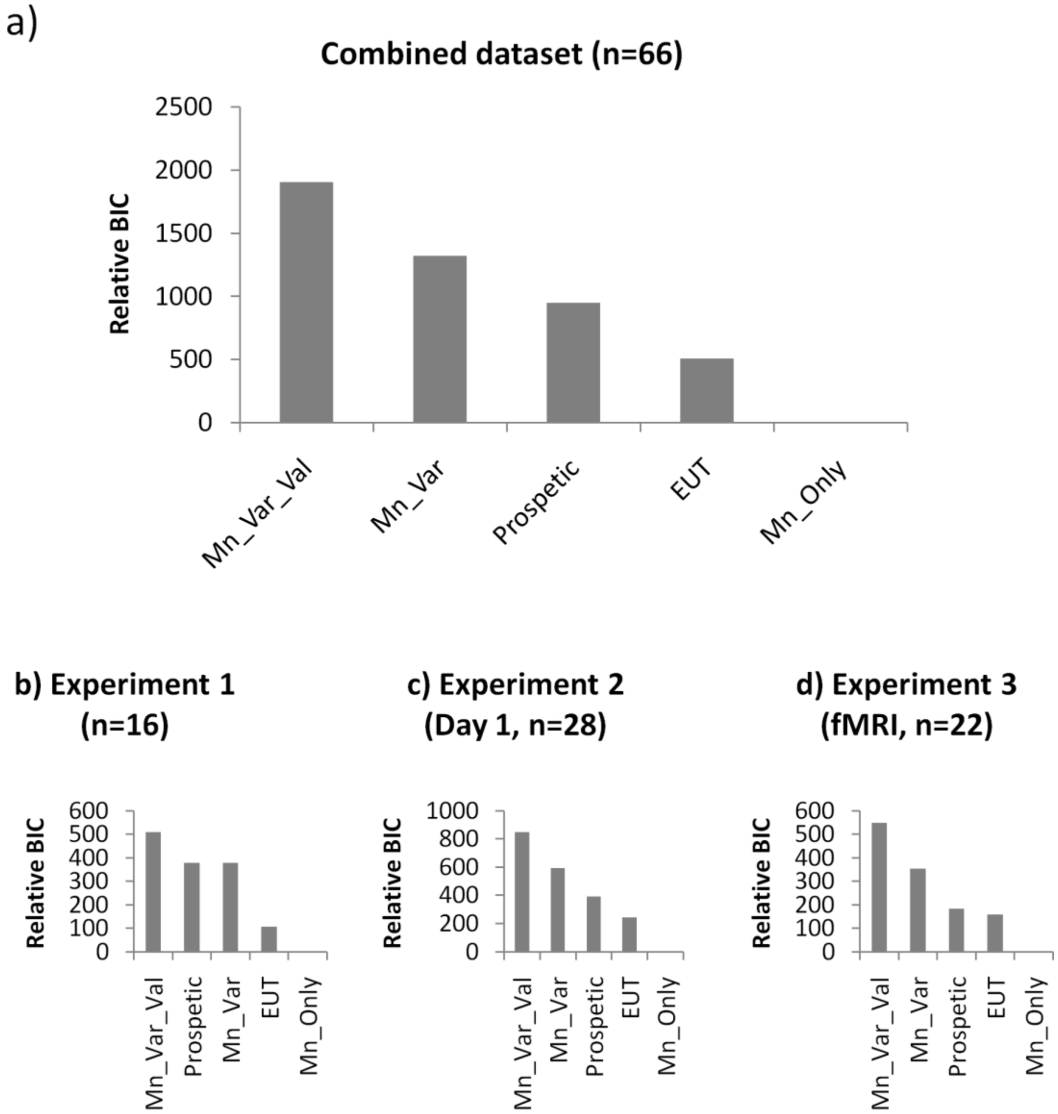

***Figure 4.7 Behavioural model comparison**. We analysed the data from each of the three experiments using the "accept/reject" task. **Panel a** shows the combined dataset (including Experiments 1, 3 and Day 1 in Experiment 2, giving n=66) and **panels b-d** each dataset separately. We plot the summed BIC for each model relative to that for the worst performing model (Mean-Only). The effects of risk and valence are seen clearly by comparing our three related "summary statistic" models: the simple Mean-Only model where individuals care only about the mean value of the options (Mn_Only) is markedly improved by adding the influence of risk in the Mean-Variance model (Mn_Var), which in turn is markedly improved by also accounting for valence in our Mean-Variance-Valence model (Mn_Var_Val) that includes separate risk parameters for each valence. We replicate these results. The importance of risk and valence is also evident with our Expected Utility (EUT) and Prospetic models: the EUT model incorporates the impact of risk and it out-performs the Mean-Only model described above, but the EUT model is itself outperformed by the Prospetic model*

that also accounts for valence. The winning model in the combined dataset and in each individual dataset is the Mean-Variance-Valence model.

Variance provides a trial-by trial metric of risk: Our task manipulated risk by parametrically altering the variance in our set of 100 lotteries (presented once as gains and once as losses). Here we asked if variance would constitute a reasonable trial-by-trial metric of risk for our fMRI analysis. Indeed, the winning model in each experiment and the combined dataset was the Mean-Variance-Valence model, which explicitly used variance as a metric of risk.

Further, in absolute terms our winning Mean-Variance-Valence model well predicted individuals' actual choices. In Experiment 1 the Mean-Variance-Valence model correctly predicted 75% (±s.d.10%) of participants choices (probability of correct choice >0.5), with similar predictive power in Experiment 2 (Day 1 80%±7% and Day 2 80%±8%) and Experiment 3 (n=22, fMRI) 74%±s.d.7%) and the combined dataset (77%±8%).

Our findings are not explained by probability distortion or choice randomness: Probability distortion, such as the overweighting of small probabilities and underweighting of large probabilities (Kahneman and Tversky, 1979), was accounted for in the Prospetic model using a Prelec probability weighting function (Prelec, 1998). However, the Mean-Variance-Valence model performed better than the Prospetic model (summed BICs reported above).

To ask if valence acted by changing choice randomness, in our winning MVV model we replaced the single free parameter in our softmax decision-rule (β) with separate parameters for gain trials (β_{gain}) and loss trials (β_{loss}). However, the Mean-Variance-Valence model with separate noise parameters (Experiment 1 summed BIC = 3314, Experiment 2 (Day 1= 5318; Day 2 = 5088); Experiment 3 = 4775) performed worse than the standard Mean-Variance-Valence model for all datasets.

Replication of our behavioural findings using model-derived parameters: Our winning Mean-Variance-Valence model provided a measure of risk preference for each valence (ρ_{gain} and ρ_{loss}) for each participant, which we could use to give a measure of the impact of risk overall (average of ρ_{gain} and ρ_{loss}) and the impact of valence (ImpValence = ρ_{gain}-ρ_{gain}). Note $\rho<0$ is risk averse, $\rho=0$ is risk neutral and $\rho>0$ is risk seeking. The model derived parameters (Experiment 1 ρ_{gain} =-0.005±0.060, ρ_{loss}=-0.045±0.035, β =1.10±1.36, one subject excluded who rejected essentially all offers; Experiment 2 Day 1 ρ_{gain} =0.010± 0.071, ρ_{loss}=-0.054±0.063, β =0.71±0.29; Experiment 2 Day 2 ρ_{gain} =-0.015±0.075, ρ_{loss}=-0.054±0.060, β =0.67±0.0.32; and Experiment 3 ρ_{gain} =-0.001±0.024, ρ_{loss}=-0.034±0.022, β =1.85±0.87) were very highly correlated with the simple metrics derived from the proportion of riskier choices (Experiment 1 gains r=0.94, P=3.1x10^{-7}, losses r=0.89, P=8.3x10^{-6}, one subject excluded who rejected essentially all offers; Experiment 2 gains [Day 1 r=0.83, P=5.5x10^{-8}; Day 2 r=0.92, P=7.2x10-^{12}], losses [Day 1 r=0.87, P=1.2x10^{-9}; Day 2 0.89, P=4.2x10^{-10}]; Experiment 3 gains r=0.94, P=1.3x10^{-10}, losses r=0.90, P=7.9x10^{-9}). Therefore, using metrics based on the model parameters instead of the proportion of riskier choices gave the same results in all preceding analyses, namely: greater gambling for gains than losses (Experiments 1, 2 and 3); and stable but independent preferences for risk and valence (Experiment 2).

4.3.3 Neuroimaging results

We next used fMRI to ask if dissociable neural processes underlie the behavioural responses to risk and valence, and the independent preferences for each, demonstrated behaviourally above (Experiment 3, n=22). We implemented a 2 valence (gain, loss) by 2 choice (accept, reject) analysis, with a trial-by-trial metric of risk as the variance of the lottery.

Our data revealed dissociable neural processing for risk and valence. The degree of risk in the lottery positively correlated with activity in posterior parietal cortex, a region associated with risk (Platt and Glimcher, 1999; Huettel et al., 2005; Mohr et al., 2010), and in middle temporal gyrus (Fig. 4.8; Table 4.1). In contrast, an effect of valence was expressed in greater activity for gains than losses in value-related (O'Doherty, 2004; Rangel et al., 2008) areas of orbitofrontal cortex and bilateral striatum (Fig.4.8), as well as left dorsolateral prefrontal cortex and right posterior insula. Further, we demonstrated an anatomical dissociation between this risk and valence related processing, by using exclusive masking with a liberal threshold (P<0.05 uncorrected). The risk-related parietal activity still survived whole brain correction having removed the valence-related voxels; as did the valence-related activity in OFC, striatum and posterior insula having removed the risk-related voxels. Interestingly, no activity for losses relative to gains survived whole brain correction, and only by taking anterior insula (Mohr et al., 2010) and amygdala (De Martino et al., 2006; Guitart-Masip et al., 2010) as a priori regions of interest did left insula survive small volume correction.

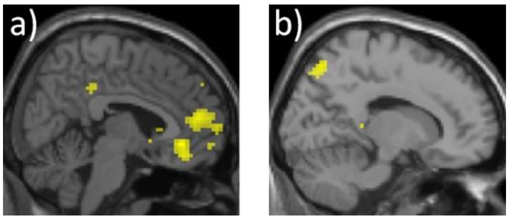

Figure 4.8 Dissociable neural encoding of stimulus risk and valence. *In Experiment 3, 22 participants underwent fMRI scanning whilst performing the "accept/reject" task, which independently manipulated the degree of risk and the valence in outcomes.* ***a)*** *For valence, greater activity was seen for gains than losses in orbitofrontal cortex and bilateral striatum.* ***b)*** *Risk was measured as the variance of the lottery, and this positively correlated with activity in posterior parietal cortex. This activity for valence and risk was neuroanatomically dissociable, as shown by exclusive masking with a liberal threshold.*

Our neural data also revealed dissociable substrates corresponding to the independent inter-individual differences found behaviourally for risk and valence.

Greater individual risk aversion (i.e. lower *PropRisky*$_{all}$) predicted enhanced activity when accepting than when rejecting a risky option, in areas including the risk-related region of posterior parietal cortex and bilateral anterior insula/inferior frontal gyrus (IFG; Fig. 4.10; Table 4.2). By contrast, the more an individual's choices were influenced by valence (*ImpValence*, defined above), the greater the enhancement of valence-related activity for gains relative to losses in right posterior insula (Fig. 4.9). Again, we used exclusive masking with a liberal threshold (P<0.05 uncorrected) to demonstrate a dissociation between these risk and valence related regions.

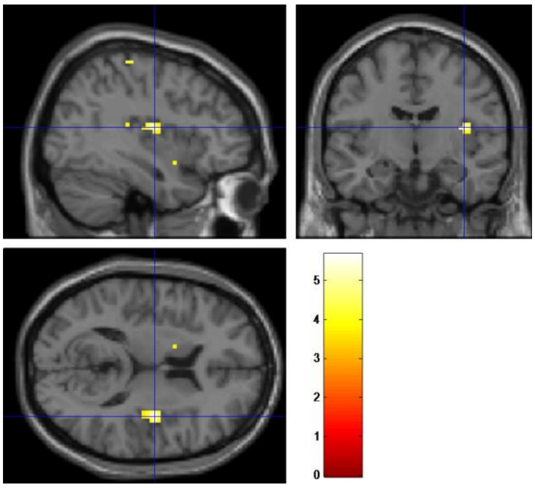

Figure 4.9 Inter-individual differences in the impact of valence. *We derive a simple metric of the impact of valence as the difference in riskier choices in each domain (ImpValence = PropRisk*$_{gain}$*-PropRisk*$_{loss}$*). The greater this metric, the greater was an individual's activity for the main effect of gains>losses in right posterior insula.*

These neural findings provide evidence that stimulus risk and valence undergo separable processing. However, this does not explain how they influence action selection to bias choice. One possibility is that they influence the individual's disposition to approach or avoid stimuli, by acting as appetitive or aversive stimulus features. Such approach/avoidance mechanisms appear to underlie a variety of biases in humans and animals (Dayan, 2008; Rangel et al., 2008; Dayan and Seymour, 2009; Guitart-Masip et al., 2010), and are consistent with the patterns of activity we observe for both risk and valence when individuals approach (accept) the lottery. Thus, the more averse an individual was to risk (i.e. lower *PropRisky*$_{all}$), the greater the activity evoked when approaching (i.e. accepting) the risky option in areas including anterior insula/IFG (Fig. 4.10), a region known to support aversive representations (Calder et al., 2001; Seymour et al., 2007). In relation to valence, actions can be parsed into approach (accept) or avoidance (reject). Of the four possible actions in our task ($\text{Gain}_{\text{accept}}$, $\text{Gain}_{\text{reject}}$, $\text{Loss}_{\text{accept}}$, $\text{Loss}_{\text{reject}}$) individuals are least disposed to choose the lottery with losses, and this specific action to which individuals were most averse ($\text{Loss}_{\text{accept}}$) was the sole action associated with increased anterior insula/IFG activity (Fig. 4.10).

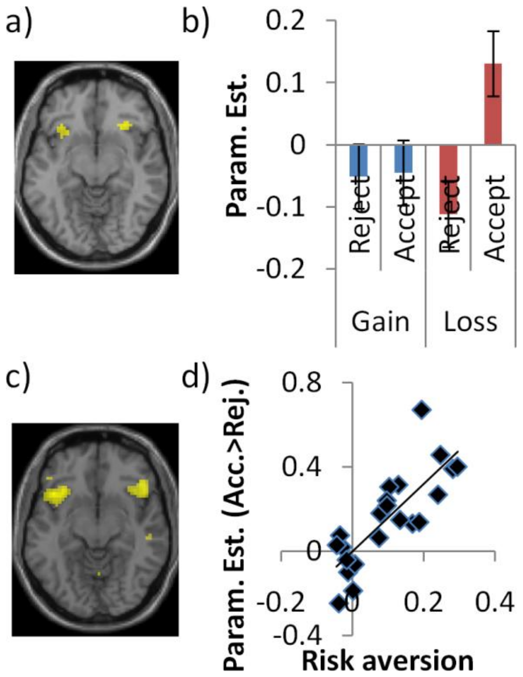

Figure 4.10 Approaching risk and loss. *A possible mechanism by which valence and risk bias choice is by influencing the disposition to approach economic stimuli. Actions can be parsed into approach (accept) or avoidance (reject).* ***a)*** *In relation to valence, anterior insula/IFG demonstrates an interaction of choice (accept, reject) and valence (gain, loss). Panel* ***b)*** *shows this interaction was driven by increased activity when approaching (accepting) the lottery with losses (Loss$_{accept}$), which was the specific action to which individuals were most averse of the four possible actions in our task (Gain$_{accept}$, Gain$_{reject}$, Loss$_{accept}$, Loss$_{reject}$). Parameter estimates are taken from the peak for this interaction in right anterior insula/IFG.* ***c)*** *For risk, the more averse an individual was to risk (i.e. lower PropRisky$_{all}$), the greater the activity when approaching (i.e. accepting) the risky option in areas including anterior insula/IFG.* ***d)*** *For illustration we plot this correlation with risk aversion (Risk aversion = 0.5 – PropRisk$_{all}$; i.e. risk-neutral = 0, risk-averse>0) at the peak for this activity in right anterior insula/IFG. Error bars indicate s.e.m..*

4.4 Discussion

We describe greater gambling for gains than losses, a finding inconsistent with a tied relationship between risk and valence that specifies a valence-induced bias in the opposite direction. Instead, we found behavioural and neural dissociations between the effects of risk and valence, consistent with an hypothesis that risk and valence exert independent influences on choice. We showed that a simple manipulation of task structure dissociated the impacts of risk and valence, by selectively reversing the effect of valence while leaving a risk-induced bias unaffected; that individual preferences for each were also independent; and further that risk and valence were encoded by distinct neural systems. These dissociations are not predicted by existing behavioural economic theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), but can be accommodated in a biologically-based account of choice in which risk and loss bias approach towards economic stimuli.

Mounting evidence suggests distinct valuation systems compete for control of action, including more reflexive systems that relate the value of particular states to innate behavioural repertoires like avoidance (Kim and Jung, 2006; Seymour et al., 2007; Dayan and Seymour, 2009); and more sophisticated goal-directed systems that use explicit models of the environment to select actions (Dayan, 2008; Rangel et al., 2008). Here, loss and risk may act through the former to bias choice by triggering avoidance or approach responses, although the degree of similarity between risk and valence related systems is a matter for further study. With respect to the insights of Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), this account is consistent with the idea of loss aversion in which losses have greater weight ("loom larger") than gains, although not with the "reflection effect" that specifies risk-seeking with losses and risk-aversion with gains. More broadly, that risk and valence bias (i.e. systematically influence) choice is not unexpected biologically, given previous work showing loss aversion in non-human primates (Chen et al., 2006) while risk sensitivity is well known to be phylogentically ancient (Real et al., 1982; Barnard and Brown, 1985; Kacelnik and Bateson, 1996).

The observation that individuals are biased to avoid a stimulus containing loss can explain behaviour in a variety of tasks. Framing a sure option as a loss biased individuals to avoid that sure option and choose a gamble instead (De Martino et al., 2006), a bias that can also be elicited by aversive conditioned stimuli presented incidentally with the sure option (Guitart-Masip et al., 2010). Avoidance of economic stimuli containing losses also explains why individuals are biased away from choosing “loss-gain mixed gambles”, which are economic stimuli containing losses along with gains (Redelmeier and Tversky, 1992; Tom et al., 2007).

A biologically-based account is also consistent with the context dependence we see in response to losses, where we reverse the direction of the loss-induced bias between our “accept/reject” and “selection” tasks (Figs. 2 and 4). Context powerfully determines how animals react to aversive stimuli, such that depending on context rats under threat respond by fleeing, freezing or even fighting (Blanchard and Blanchard, 1988; Seymour et al., 2007; Dayan and Seymour, 2009). Although losses induced avoidance in both our tasks, in the “selection” task individuals had to select between two lotteries and so could not express avoidance by withdrawal, but instead could potentially avoid losses by selecting the higher variance (riskier) option. Consistent with our data, context effects in the same direction have been shown with “loss-gain mixed gambles”, which when presented analogously to our “accept/reject” task were avoided more often than when presented analogously to our “selection” task (Ert and Erev, 2008). In the classic paper establishing Prospect Theory (Kahneman and Tversky, 1979), each problem presented two options for individuals to select between, which led to the same direction of effect as in our “selection” task.

The neural data from the “accept/reject” task also support the idea that valence biases individuals from approaching (accepting) the lottery with losses, expressed by increased anterior insula/IFG activity for this action (Fig. 4.8). Anterior insula is known to be involved in the representation of aversive stimuli (Calder et al., 2001; Seymour et al., 2007), although we recognise that fMRI data is only suggestive and that causal evidence for an approach/avoidance mechanism will depend on further experiments. Nevertheless, our data help reconcile previously discrepant neural findings concerning valence. As expected we find greater activity for gains than losses in a ventral valuation network comprising bilateral striatum and orbitofrontal cortex (O'Doherty, 2004; Tom et al., 2007; Rangel et al., 2008). However, with respect to loss-related activity, whilst some studies report activity in regions associated with aversive processing, such as amygdala (De Martino et al., 2006; Guitart-Masip et al., 2010) and anterior insula (Guitart-Masip et al., 2010), others do not (Tom et al., 2007). Crucially the loss-related activity we find in anterior insula is driven by having to approach losses, explaining why loss-related activity is reported by studies using contrasts including choice (De Martino et al., 2006; Guitart-Masip et al., 2010), such as in the interaction we see between valence and choice (Fig. 4.10). That we see loss-related activity in anterior insula rather than amygdala may reflect its involvement in representing more complex aspects of aversive stimuli (Seymour et al., 2007).

With respect to risk, the overall proportion of riskier choices was similar in the "accept/reject" and "selection" tasks ($PropRisk_{all}$ in Figs. 4.3, 4.4), where by design the magnitudes of the differences in risk between the two options in the trials was similar. Most individuals were biased (i.e. systematically influenced) to be averse to risk overall. This impact of risk was seen regardless of whether context led valence to induce greater gambling for gains than losses, or the opposite. That risk constitutes an important variable influencing choice is a long-standing idea in psychology (Coombs and Pruitt, 1960), finance (Markowitz, 1952; Bossaerts, 2010) and animal behaviour (Real et al., 1982; Barnard and Brown, 1985; Kacelnik and Bateson, 1996); which has been argued to involve an important affective component (Schonberg et al., 2011).

Our neural data revealed activity encoding the degree of stimulus risk in parietal cortex (Fig. 3), which was anatomically dissociated from activity encoding the stimulus valence. Such parietal activity concurs with single unit and fMRI data showing enhanced activity during risky decision-making (Platt and Glimcher, 1999;

Huettel et al., 2005; Mohr et al., 2010). Interestingly, we did not observe this correlation in insula, previously seen in the absence of parietal activity when risk is manipulated by altering win probability (Preuschoff et al., 2006). Instead, parietal cortex is known to express an interaction between number and space (Hubbard et al., 2005), suggesting that this parietal risk representation may reflect the spread of an outcome distribution. Our neural data do implicate anterior insula in one potential mechanism by which risk may bias choice, namely where individuals were biased from approaching (choosing) stimuli containing risk. Consistent with such a mechanism, we found greater individual aversion to risk associated with greater activity when approaching the risky option in anterior insula/IFG (Fig. 4.8).

Finally, we demonstrate stable and independent inter-individual differences for risk and valence (Figs. 4.5, 4.6), which were mirrored by dissociable neural correlates of inter-individual differences for each (Figs. 4.9, 4.10). These findings are supportive of independence between risk and valence induced biases, although alone are not inconsistent with the “reflection effect” in Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). Stability in the aversive impact of loss on choice over time has not to our knowledge been previously demonstrated, and is interesting in light of work suggesting framing effects may be genetically mediated (Roiser et al., 2009). Stability in the impact of risk concurs with work showing stability over time durations of months (Andersen et al., 2008). Our finding of functional segregation in insula for these preferences also fits recent work showing this region’s putative role in preferences (Singer et al., 2009a), and considerable functional segregation in this large cortical region (Caruana et al., 2011).

In conclusion, we find behavioural and neural dissociations between the effects of risk and valence, consistent with an hypothesis that risk and valence exert independent biases on choice. These dissociations are not predicted by existing behavioural economic theory. However, a biologically-based account of choice, in which risk and valence bias approach responses, can explain both classical (Kahneman and Tversky, 1979; Camerer, 1989; Battalio et al., 1990; Tversky and Kahneman, 1992) and our new findings. Specifically, within an account of choice proceeding from option evaluation to action selection (Corrado et al., 2009), we suggest that the risk and valence of an economic stimulus are processed by separable neural systems, and may influence action-selection partly through reflexive systems that bias approach responses. Recasting the relationship between risk and valence from a biological perspective yields testable predictions, and carries implications across the diverse disciplines to which existing theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) has been applied, including the economic (Camerer, 1998), cognitive (De Martino et al., 2006) and political sciences (Levy, 2003). In the next Chapter, we directly test the predictions of an approach avoidance mechanism using reaction times.

4.5 Tables

Regions	L/R	x y z	Z	# vox	Corr. P value
Gain>loss					
OFC / rostral ACC	R	6 38 -8	4.67	876	<0.001
		6 50 13	4.29		
		9 38 -17	4.22		
Putamen	L	-12 11 22	4.53	680	<0.001
		-24 -10 13	3.99		
Putamen	R	24 11 1	4.18		
Posterior Insula	R	30 -22 19	4.29	157	0.008
		36 -34 16	3.79		
		39 -19 16	3.75		
Loss>gain		nil			
Accept>reject					
Caudate	R	15 17 10	5.65	288	<0.001
		0 -1 16	4.02		
		6 -13 25	3.86		
Infr. Parietal lob.	R	51 -34 49	4.62	1539	<0.001
		45 -43 46	4.39		
Precuneus		21 -73 43	4.06		
Supr. Medl. gyrus	R	9 32 40	4.15	1226	<0.001
		24 14 46	4.00		
Supr. Medl. gyrus	L	3 38 31	4.15		
Reject>accept		nil			
Interaction (gain>loss, reject>accept)					
pre-SMA	R	9 20 61	4.21	285	0.003
		-3 29 52	3.62		
		0 32 43	3.58		
Anterior Insula / IFG	R	30 26 -8	4.02	97	0.025
		27 20 -20	3.39		
		39 20 -11	3.38		
Interaction (gain>loss, accept>reject)		nil			
Variance (pos. correl.)					
ITG / MTG	R	48 -61 -8	4.59	155	<0.001
		54 -55 -2	4.41		
Posterior parietal	R	39 -82 22	4.48	316	<0.001
Supr. parietal / precuneus		15 -70 55	4.25		
		33 -64 34	4.12		
Variance (neg. correl.)					
Cerebellum	L/R	-6 -76 -17	4.84	732	<0.001
		-9 -79 -26	4.61		
		6 -73 -23	4.24		

Table 4.1 fMRI results across subjects. *This table shows all activity surviving cluster level correction (P<0.05 FWE corrected; threshold of P<0.005 used to define the clusters) for contrasts involving: valence (gain versus loss); choice (accept versus reject); interaction of choice and valence; positive and negative correlations with variance; positive and negative correlations with expected value; interaction of variance in gains versus losses. For each cluster is shown: the three constituent peaks with the highest Z-scores; the number of voxels at P<0.005 (uncorrected); and the P-value of the cluster after FWE correction across the whole brain. No regions significantly correlated with expected value. In addition to these whole brain corrected results: left anterior insula showed the same interaction (gain>loss, reject>accept x=-27 y=20 z=-11 #vox=69) as right anterior insula, which was again driven by increased activity when accepting the lottery with losses; and on the left this also led to activity for loss>gain (x= -33 y=20 z=-5 Z=4.13 #vox=51). (ACC = Anterior Cingulate Cortex; IFG = Inferior Frontal Gyrus; OFC = orbitofrontal cortex; SMA = Supplementary Motor Area).*

Regions	L/R	x y z	Z	# vox	Corr. P value
PropRisk$_{all}$ (neg correl.) on accept>reject					
Infr. parietal lob.	L/R	-54 -43 49	5.20	2328	<0.001
Postcentral gyr.		33 -70 46	4.67		
		-30 -46 43	4.54		
Anterior Insula / IFG	L	-42 20 -8	5.02	384	<0.001
		-33 20 -11	4.76		
		-51 38 1	4.20		
Anterior Insula / IFG	R	39 23 -11	4.47	212	0.003
		48 23 -8	4.47		
		42 20 -2	4.25		
Middle Frontal gyr.	L	-30 2 61	4.19	194	0.022
		-33 -16 52	3.97		
		-24 -7 49	3.52		
Supr. Medl. gyr.	L/R	-3 29 49	4.17	768	<0.001
		6 23 43	4.16		
		51 14 22	4.05		
Caudate	R	15 -7 13	4.06	458	<0.001
Thalamus		15 8 13	4.01		
		9 -31 1	3.82		
Valence Impact (pos correl.) on gain>loss					
Posterior insula	R	39 -10 13	4.32	129	0.008
		63 -19 4	3.87		
		36 -31 16	3.22		

Table 4.2 fMRI results between subjects, using second level covariates related to risk and valence. *This table shows all activity surviving cluster level correction (P<0.05 FWE corrected; threshold of P<0.005 used to define the clusters) for contrasts involving: the second level covariate for risk (PropRisk$_{all}$) on activity for accept>reject; and the second level covariate for valence (ImpValence) on activity for gain>loss. The negative correlation with risk preference (PropRisk$_{all}$) indicates greater activity for accepting (approaching) the lottery with increasing risk aversion. For each cluster is shown: the three constituent peaks with the highest Z-scores; the number of voxels at P<0.005 (uncorrected); and the P-value of the cluster after FWE correction across the whole brain. (IFG = Inferior Frontal Gyrus).*

Chapter 5. Individual choice: Avoiding losses and approaching risks bias reaction times

5.1 Introduction

As described in the preceding chapter, the degree of risk in outcomes and their valence are powerful determinants of choice. In Prospect Theory the “reflection effect” specifies a tied relationship between risk and loss, such that individuals prefer gambling with losses and also safer options with gains (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). However, the preceding chapter suggests instead that risk and valence exert independent biases on choice: that is valence influences choice, but can either increase or decrease gambling dependent on context; and regardless of whether valence increases or decreases gambling, collapsing across domains there is a consistent overall effect of risk. Furthermore, our neural data, particularly in anterior insula, suggest a potential mechanism by which risk and valence bias choice, acting through systems that trigger approach towards appetitive and avoidance of aversive stimulus components. Here, in this chapter we test our hypothesis that risk and valence exert independent biases through approach/avoidance mechanisms using reaction time (RT) data.

Reaction times have previously been shown to be slower towards aversive and faster towards appetitive stimuli (Guitart-Masip et al., 2011). If valence and risk are indeed such stimulus features, this makes for the following predictions. With respect to valence, individuals will be slower to approach (choose) options containing losses than gains. With respect to risk, this stimulus feature can be aversive, neutral or appetitive depending on an individual’s risk preference. We predict that when risk-averse individuals will be slower to approach, when risk-neutral will show no RT bias; and when risk-seeking they will be faster to approach risk.

Furthermore, we can use RT data to suggest where within the choice process risk and valence might exert their biases. As described in the Literature Review (Chapter 2), the choice process can be broadly considered to consist of option evaluation and then action selection (Corrado et al., 2008). In this chapter we allow subjects to select their action at any point within the trial; whilst in the preceding chapter (Chapter 4) on each trial individuals evaluated the options during an imposed wait period before action selection. If in the free response experiments we see the same RT biases as were seen previously with the imposed wait, this would be more suggestive of a bias affecting the later action selection than option evaluation.

5.2 Methods

In the experiments reported in the previous chapter, in each trial participants were required to wait for 4020msec evaluating the options before then having 1500 msec in which to choose. Here, we report data from two new experiments in which individuals were free to respond at any time within the 5520msec that the options were displayed (Fig. 5.1). In one new experiment we used the “accept/reject” task, and we refer to this dataset as $\text{AccRej}_{\text{free}}$; n=19). The second new experiment used the “selection” task ($\text{Selection}_{\text{free}}$; n=34).

We compare behaviour in our two new free response time experiments to behaviour in the tasks with an imposed wait time. The latter uses data from two experiments from the preceding chapter, one using the “accept/reject” task ($\text{AccRej}_{\text{wait}}$; n=22; Experiment 3 in the preceding chapter) and one using the “selection” task ($\text{Selection}_{\text{wait}}$; n=24; Experiment 4 in the preceding chapter). Due to a coding error, reaction times were not accurately recorded in Experiments 1 and 2 from the preceding chapter.

Therefore, here we present results from four experiments ($\text{AccRej}_{\text{free}}$; $\text{Selection}_{\text{free}}$; $\text{AccRej}_{\text{wait}}$; and $\text{Selection}_{\text{wait}}$), in which we can examine the effect of task (“accept/reject” and “selection” tasks) and response (“free” and “wait”) on choice and reaction times. The study was approved by the Institute of Neurology (University College, London) Research Ethics Committee.

5.2.1 Participants

All participants were recruited using institutional mailing lists, were healthy and provided informed consent. In our two new experiments, 19 participants took part in AccRej$_{free}$ (mean age 23 years, range 19-31; 6 male; one further participant was excluded as they only rejected); and 34 participants took part in Selection$_{free}$ (mean age 24 years, range 19-36; 16 male; one further participant was excluded who confused the buttons). As described in the preceding chapter, 22 right-handed participants took part in the AccRej$_{wait}$ experiment (age mean 22 years, range 18-32; 6 male); and 24 participants took part in Selection$_{wait}$ experiment (age mean 23 years, range 18-34; 3 male).

5.2.2 Task

The two experiments using an imposed wait (AccRej$_{wait}$ and Selection$_{wait}$) are described in the preceding chapter. Each trial began with a fixation cross presented for 1-2secs (mean 1.5secs); followed by viewing the options for 4020msec; and finally a black square appeared to indicate participants had 1500msec to input their choice by button press (the black square turned white when they chose). The two new experiments (AccRej$_{free}$ and Selection$_{free}$) were identical, except that individuals could choose at any point during the 5520msec for which the stimuli were presented (the black square was present throughout stimulus presentation and turned white when they chose).

As described in the previous chapter, in each experiment there were 100 “gain trials” and 100 “loss trials”, with all 200 trials presented in random order. Payment was as before. Participants began the day with an endowment of £12. At the end of the experiment, one “gain trial” and one “loss trial” were picked at random and the outcome of both were added to the endowment to determine payment. Participants could receive between £0-24 in the task. The AccRej$_{\text{wait}}$ dataset was previously acquired during fMRI scanning and all amounts were doubled.

5.2.3 Stimulus sets

In the “accept/reject” task we used the set of trials described in the previous chapter (AR$_{\text{MainList}}$), where we manipulated the difference in variance (ΔVar;10 levels) and EV (ΔEV;10 levels) of the lottery relative to the sure option of £6 (maximum ΔEV 1.25, and maximum ΔVar 23.9). In the “Selection” task we used the same set of trials described in the preceding chapter (maximum ΔEV 1.9 and maximum ΔVar 18.3).

5.2.4 Statistical analysis

All statistical tests used were two tailed.

5.2.5 Choice modelling

To ensure consistency between behaviour when participants made a free response and the behaviour described in the preceding chapter, we used a model-based analysis of participants’ choices. We used identical methods to those described in preceding chapter.

5.2.6 Reaction time normalisation

We normalised each individual’s RTs by taking the natural logarithm, mean-correcting and dividing by the standard deviation. However, we note that our findings were the same irrespective of having used “raw” or normalised RTs.

5.3 Results

5.3.1 Choice behaviour with free response period

Choice behaviour in the two new experiments with a free response ($\text{AccRej}_{\text{free}}$ and $\text{Selection}_{\text{free}}$ experiments) strikingly replicate our previous results using an imposed wait period ($\text{AccRej}_{\text{wait}}$ and $\text{Selection}_{\text{wait}}$ experiments) (Fig. 5.1). As before, both risk and valence strongly influenced choice in the "accept/reject" task and the "selection" task; but with our "selection" task we selectively reverse the direction of the valence-induced bias whilst leaving the overall risk-induced bias unaffected – that is, we again dissociate risk and valence effects. Further, as before inter-individual differences for risk and valence were dissociable in both new experiments. This choice data is detailed below.

Impact of risk: In our "accept/reject" task half the lotteries had an expected value above the sure amount and half below, providing a simple metric of risk preference as the proportion of riskier choices made (*PropRisk*; risk-neutral=0.5; risk-averse<0.5; risk-seeking>0.5), which could also be used with our "selection" task. In both new experiments our participants were biased to be risk averse, choosing the risky option less than half the time overall (i.e. $PropRisk_{all}$ <0.5), shown by one sample ttests against the null hypothesis of risk-neutrality (i.e. $PropRisk_{all}$ = 0.5): $\text{AccRej}_{\text{free}}$ ($PropRisk_{all}$ 0.41±0.14; one-sample t-test versus risk-neutral, $t_{(18)}$=-2.83, P=0.01) and $\text{Selection}_{\text{free}}$ (0.39±0.11; ttest versus risk neutral $t_{(33)}$=-6.17, P=5.9x10^{-7}). There was no difference in the overall risk-aversion between the $\text{AccRej}_{\text{free}}$ and $\text{Selection}_{\text{free}}$ datasets (independent samples ttest $t_{(51)}$=-0.75, P=0.46).

For comparison, as detailed in the preceding chapter: in the $\text{AccRej}_{\text{wait}}$ experiment $PropRisk_{all}$ = 0.40±0.11, t(21)=-4.2, P=0.0002; and in the $\text{Selection}_{\text{wait}}$ experiment $PropRisk_{all}$ = 0.42±0.11, t(23)=-3.7, P=0.001.

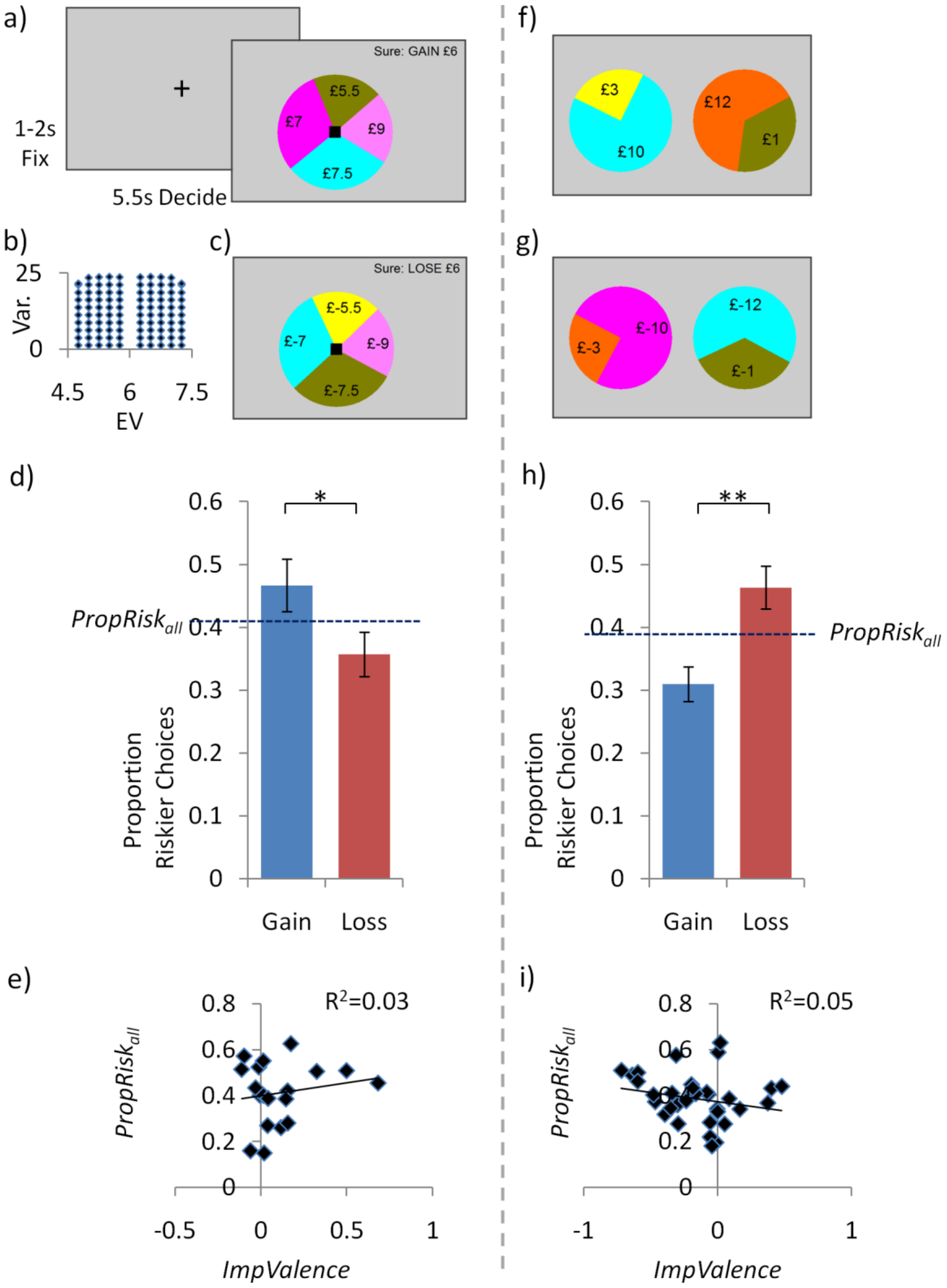

Figure 5.1 Experimental design and choice in our two new experiments with free response periods. *In two new experiments (AccRej$_{free}$ and Selection$_{free}$), both tasks were exactly as those reported previously (Chapter 4), except that here on each trial individuals could respond at any time during the 5.5 seconds for which the stimuli were displayed. Choice in these two new experiments was essentially identical to that reported in the previous chapter when there was an imposed wait period before response. Panels* ***a-e*** *refer to the "accept/reject" task.* ***a)*** *In each "gain trial"*

individuals chose to accept a lottery (4 possible outcomes, all ≥ 0) or reject and so receive £6 for certain. ***b)*** *In the 100 "gain trials" we parametrically and orthogonally manipulated the degree of risk (defined as outcome variance; 10 levels) and expected value (EV; 10 levels) of the lotteries. Half the lotteries had an EV above the sure amount and half below, metricating risk preference as the proportion of riskier choices (PropRisk; risk-averse<0.5; risk-neutral=0.5; risk-seeking>0.5).* ***c)*** *Multiplying all "gain trial" amounts by -1 gave 100 "loss trials".* ***d)*** *Behaviour in the "accept/reject" task (AccRej$_{free}$, n=19). Individuals were risk averse overall (i.e. PropRisk$_{all}$ <0.5). Valence also biased choice, with more gambling for gains than losses (ImpValence = PropRisk$_{gain}$-PropRisk$_{loss}$).* ***e)*** *Individuals' risk and valence-related preferences were independent. Panels* ***f-i*** *refer to the "selection" task, in which again there were:* ***f)*** *100 "gain trials"; and* ***g)*** *100 "loss trials". However, here in each trial individuals were presented with two lotteries to consider and select between.* ***g)*** *In the "selection" task (Selection$_{free}$, n=34) again there was overall risk aversion overall (i.e. PropRisk$_{all}$ <0.5), but the direction of the valence effect was completely reversed. i****)*** *Individuals' risk and valence-related preferences were independent. Error bars show s.e.m., * P<0.05, ** P=0.005.*

Impact of valence: As before, we extracted a simple metric for the valence-induced bias from the difference in riskier choices in each domain (*ImpValence* = *PropRisk$_{gain}$-PropRisk$_{loss}$*). Valence biased choice in both new experiments, as shown by one sample t-tests against the null hypothesis of no bias (i.e. *ImpValence* = 0): AccRej$_{free}$ (*ImpValence* = 0.11±0.20, t(18)=2.36, P=0.030); Selection$_{free}$ (*ImpValence* = -0.15±0.29, t(33)=-3.12, P=0.004). The magnitude of the valence related bias was the same in both tasks (independent samples ttest t$_{(51)}$=0.59, P=0.56).

For comparison, as detailed in the preceding chapter: in the AccRej$_{wait}$ experiment *ImpValence* = 0.18±0.15, t(21)=5.6, P=1.5x10^{-5}; and in the Selection$_{wait}$ experiment *ImpValence* = -0.16±0.25, t(23)=-3.1, P=0.005.

Relationship between the risk- and valence-induced biases: In the "accept/reject" task participants gambled more for gain than loss outcomes (AccRej$_{free}$ ttest t$_{(18)}$=2.36, P=0.03). In the "selection" task we reversed the direction of this valence-induced bias and showed more gambling for losses than gains (Selection$_{free}$, ttest t$_{(33)}$=-3.12, P=0.004). Thus, despite context reversing the effect of valence, context had no effect on the overall risk-induced bias (i.e. risk and valence effects are dissociated).

As before, this robust valence-induced bias did not result in participants becoming absolutely risk-seeking in either valence in either experiment, as shown by one sample ttests against risk-neutrality (i.e. *PropRisk* = 0.5): in the $\text{AccRej}_{\text{free}}$ experiment (*PropRisk*$_{gain}$ 0.47±0.18, t(18)=-0.79, P=0.44; *PropRisk*$_{loss}$ 0.36±0.16, $t_{(18)}$=-4.01, P=0.001); and in the $\text{Selection}_{\text{free}}$ experiment (*PropRisk*$_{gain}$ (0.31±0.16; ttest $t_{(33)}$=-6.95, P=6.1x10^{-8}; *PropRisk*$_{loss}$ 0.46±0.20 ttest $t_{(33)}$=-1.09, P=0.29).

5.3.1.1 Independent inter-individual differences for risk and valence

As before, inter-individual differences for risk and valence were dissociable in both our new experiments. There was no correlation between *PropRisk*$_{all}$ and *ImpValence* in either the $\text{AccRej}_{\text{free}}$ experiment (r = 0.17, P=0.49) or the $\text{Selection}_{\text{free}}$ experiment (r=-0.14 P=0.53; Fig. 5.2).

5.3.1.2 Behavioural modelling of choice behaviour

Model-based analysis of data from the accept/reject task replicated our previous findings (Fig. 5.3). Choice was best predicted by models incorporating both risk and valence induced biases, and as before the winning model was the Mean-Variance-Valence model described in the preceding chapter.

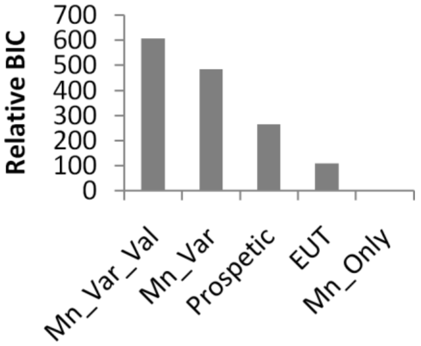

Figure 5.2 Behavioural model comparison in the new AccRej $_{free}$ ***experiment.*** *We show the same results as before, with the Mean-Variance-Valence model best predicting choice. We plot the summed BIC for each model relative to that for the worst performing model (Mean-Only). The effects of risk and valence are seen clearly by comparing our three related "summary statistic" models: the simple Mean-Only model where individuals care only about the mean value of the options (Mn_Only) is markedly improved by adding the influence of risk in the Mean-Variance model (Mn_Var), which in turn is markedly improved by also accounting for valence in our Mean-Variance-Valence model (Mn_Var_Val) that includes separate risk parameters for each valence. The importance of risk and valence is also evident with our Expected Utility (EUT) and Prospetic models: the EUT model incorporates the impact of risk and it out-performs the Mean-Only model described above, but the EUT model is itself outperformed by the Prospetic model that also accounts for valence.*

5.3.2 Reaction times with free response

5.3.2.1 Approaching valence biases reaction times

We next tested our hypothesis that individuals are slower to choose (approach) losses than gains – regardless of whether the task induces increased gambling for gains relative to losses ("accept/reject" task) or the reverse ("selection" task). As predicted individuals were slower to approach losses than gains in both new experiments, using the "accept/reject" task (AccRej$_{free}$: gains mean RT 2975± s.d.574msec; losses 3189±648; $t_{(18)}$=4.62, P=2.1x10^{-4}) and the "selection" task (Selection$_{free}$: losses 3222±525; gains 2681±472; $t_{(33)}$=13.04, P=1.4x10^{-14}) (Fig. 5.2).

This main effect of valence was also seen in a 2 valence (gain, loss) by 2 choice (riskier, surer) analysis of variance (ANOVA), with RT as the dependent variable, in both new experiments: in the AccRej$_{free}$ experiment (gain riskier 3066msec ±464;

gain surer 3016±702; loss riskier 3365±556; loss surer 3365±556; main effect of valence $F_{(1,18)}$=29.0, P=4.1x10^{-5}; no main effect choice $F_{(1,18)}$=2.87, P=0.11; and no interaction $F_{(1,18)}$=2.63, P=0.12); and in the Select$_{free}$ experiment (gain riskier 2880±604; gain surer 2674±445; loss riskier 3274±544; loss surer 3256±541; main effect of valence $F_{(1,33)}$=245.95, P=7.4x10^{-17}; main effect choice, $F_{(1,33)}$=8.17, P=0.007; and no interaction $F_{(1,33)}$=3.03, P=0.091).

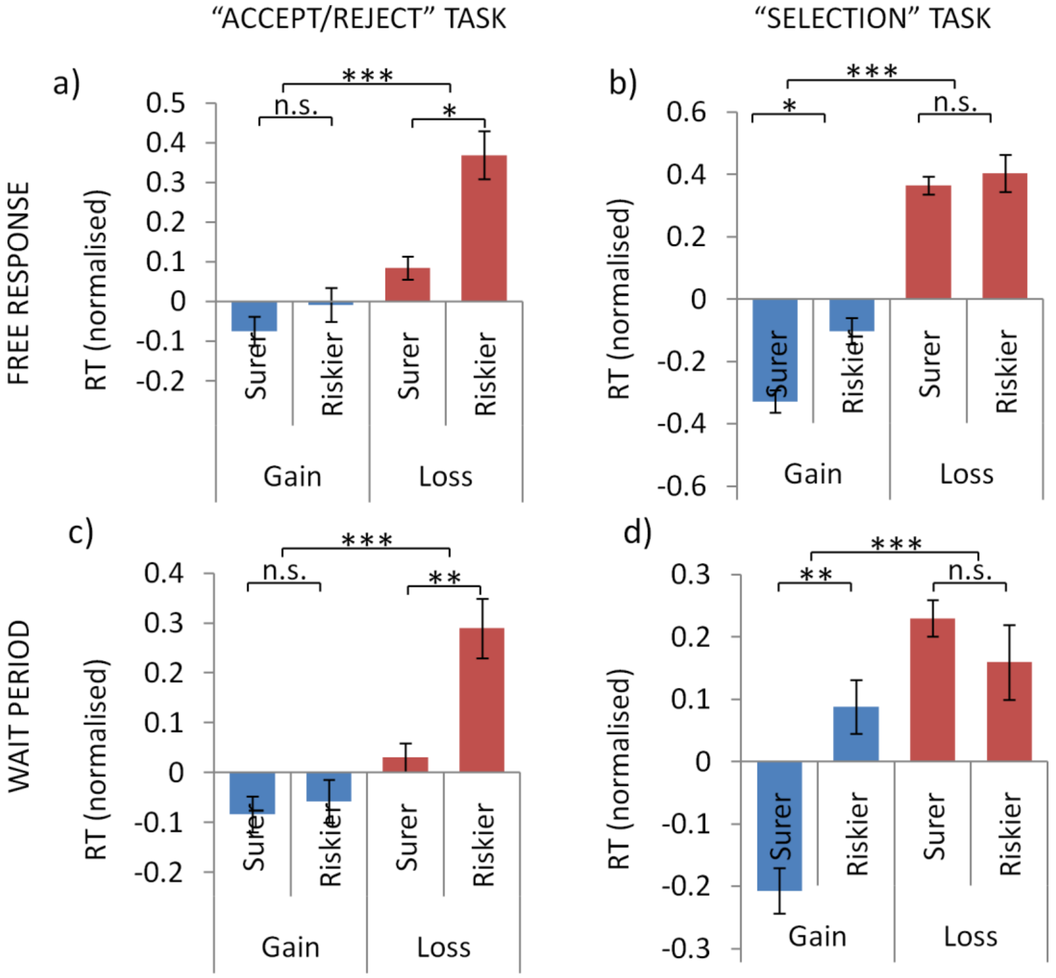

Figure 5.3 Valence biases RTs. *In all four experiments individuals were slower to approach losses than gains in all four experiments: a) the AccRej$_{free}$ experiment; b) Selection$_{free}$; c) AccRej$_{wait}$; and d) Selection$_{wait}$. RT data is normalised for each subject. In each experiment we parse trials into the four possible events: the possibilities here are a gain trial and choose the surer option; a gain trial and choose riskier; a loss trial and choose surer, a loss trial and they choose riskier.*

*Error bars show s.e.m., * P<0.05, ** P=0.005, *** P<0.0005.*

5.3.2.2 Approaching risk biases reaction times

We hypothesised that, depending on an individual's preferences, risk can be aversive, neutral or appetitive. Thus, our approach/avoidance hypothesis makes the following predictions: when individuals are risk-averse then they will be slower to choose riskier options; when individuals are risk-neutral no RT bias will be seen; and when individuals are risk-seeking then they will be faster to choose riskier options.

Analysing our data averaged across subjects, we show exactly these RT patterns for risk-aversion and risk-neutrality in both new experiments (Fig. 5.2). In the $\text{AccRej}_{\text{free}}$ experiment, risk-aversion seen in losses was associated with slower RTs for riskier than safer choices in losses (t(18)=2.49, P=0.023); and the risk-neutrality in gains was associated with no RT bias (t(18)=0.51, P=0.62). In the $\text{Selection}_{\text{free}}$ experiment, the risk-aversion seen in gains was associated with slower RTs for riskier than safer choices (t(33)=2.67, P=0.012); and risk-neutrality in losses was associated with no RT bias (t(33)=0.35, P=0.73).

However, averaging across subjects masks the degree to which RTs conform to our predictions concerning risk. For example, although on average we see risk-neutrality with gains in the "accept/reject" task, this averages across individuals from along the full spectrum of risk preferences. In the $\text{AccRej}_{\text{free}}$ experiment, an individual's risk preference with gains ($\text{PropRisk}_{\text{gain}}$) strongly predicted the RT bias ($\text{RT}_{\text{riskier}}$-$\text{RT}_{\text{surer}}$) with gains (r=-0.89, P=4.2x10^{-7}); and their risk preference with losses ($\text{PropRisk}_{\text{loss}}$) strongly predicted the RT bias with losses (r=-0.75, P=2.3x10^{-4}). In the $\text{Selection}_{\text{free}}$ we see exactly the same relationship between risk preference and RT bias with both gains (r=-0.82, P=4.6x10^{-9}) and losses (r=-0.70, P=4.8x10^{-6}). Furthermore, as shown in Fig. 5.4, in both tasks and for both gains and losses, we see exactly the predicted pattern, where: risk slowed approach when risk was aversive; risk induced no RT bias when risk was neutral; and risk speeded approach when risk was appetitive.

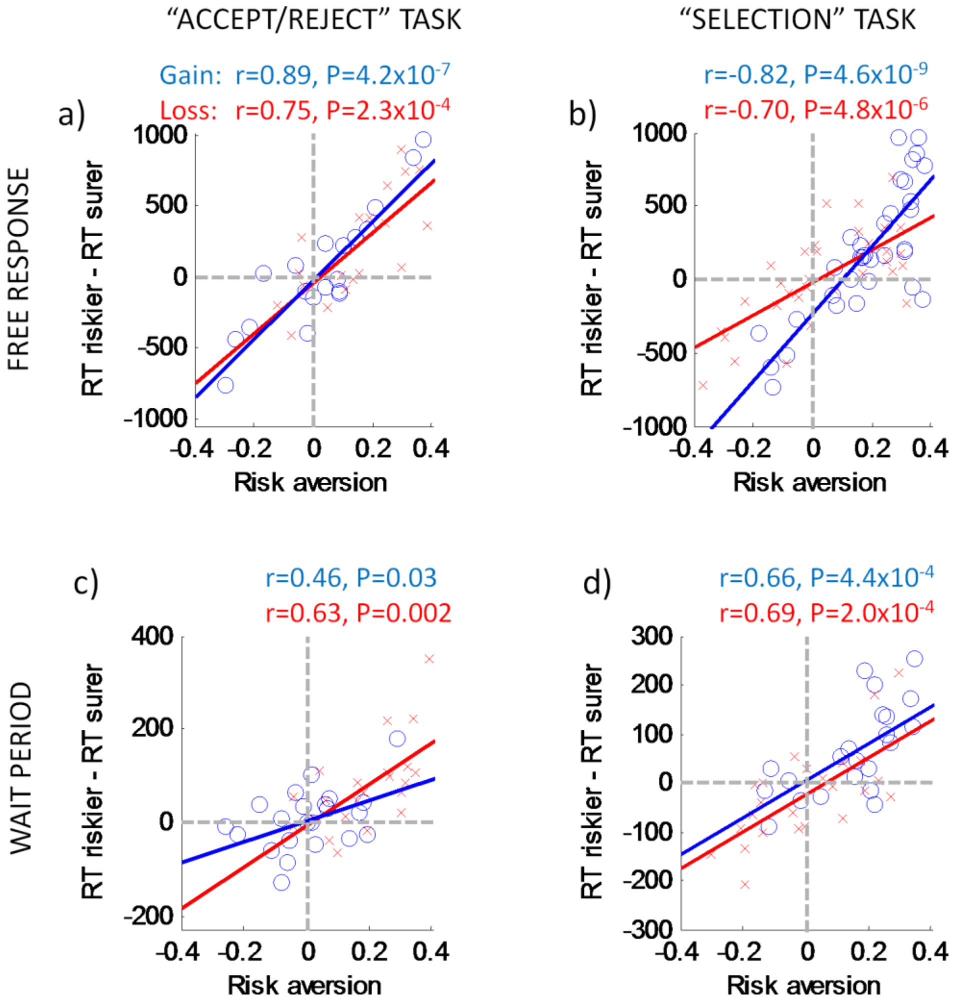

Figure 5.4 Risk biases RTs and can be aversive, neutral or appetitive. *In all four experiments, an individual's risk preference with gains strongly predicted an RT bias ($RT_{riskier}$-RT_{surer}) with gains; and their risk preference with losses strongly predicted the RT bias with losses. In both tasks we observe our predicted pattern, where: risk slowed approach when risk was aversive; risk induced no RT bias when risk was neutral; and risk speeded approach when risk was appetitive. Gains are in blue and losses are in red. Regression lines are shown, which are in not constrained in any way. For illustration as a measure of risk preference we plot risk aversion (Risk aversion = 0.5 – $PropRisk_{all}$; i.e. risk-seeking<0, risk-neutral = 0, risk-averse>0).*

5.3.3 Risk and loss bias RTs after imposed wait period

We replicated these RT biases for risk and valence in our experiments with an imposed wait period of 4 seconds in each trial before the 1.5 second choice period ($AccRej_{wait}$; $Selection_{wait}$; Fig. 5.4). This suggests that these biases affect action selection rather than option evaluation.

With respect to valence, individuals were slower to choose (approach) losses than gains in both the “accept/reject” task ($\text{AccRej}_{\text{wait}}$: losses mean RT 621msec± s.d.105; gains 575±; $t_{(21)}$=3.16, P=0.005) and the “selection” task ($\text{Selection}_{\text{wait}}$: losses 611±135; gains 531±103; $t_{(23)}$=6.58, P=1x10^{-6}). This main effect of valence was also shown in a 2 valence (gain, loss) by 2 choice (riskier, surer) analysis of variance (ANOVA) with RT as the dependent variable in both the $\text{AccRej}_{\text{wait}}$ experiment (gain riskier 584±94; gain surer 575±99; loss riskier 682±141; loss surer 603±99; main effect of valence $F_{(1,21)}$=15.11, P=8.5x10^{-4}; main effect choice $F_{(1,21)}$=14.58, P=0.001; and an interaction $F_{(1,21)}$=7.41, P=0.013); and the $\text{Selection}_{\text{wait}}$ experiment (gain riskier 584±131; gain surer 517±103; loss riskier 609±148; loss surer 634±140; main effect of valence $F_{(1,23)}$=41.41, P=1.5x10^{-6}; no main effect choice, $F_{(1,23)}$=2.56, P=0.12; and an interaction $F_{(1,23)}$=11.00, P=0.003).

With respect to risk, analysing our data averaged across subjects we again show the predicted RT pattern for risk-aversion and risk-neutrality in both experiments with an imposed wait (Fig. 5.2). In the $\text{AccRej}_{\text{wait}}$ experiment, the risk-aversion seen in losses was associated with slower RTs for riskier than safer choices in losses (t(21)=4.00, P=7.2x10^{-4}); and the risk-neutrality in gains was associated with no RT bias in gains (t(21)=0.65, P=0.52). In the $\text{Selection}_{\text{wait}}$ experiment, the risk-aversion seen in gains was associated with slower RTs for riskier than safer choices in gains (t(23)=3.62, P=0.001); and risk-neutrality in losses was associated with no RT bias in losses (t(23)=1.24, P=0.23).

Further with respect to risk, we again show that individuals’ choice biases associated with risk were strongly correlated with their RT biases induced by risk (Fig. 5.3). In the $\text{AccRej}_{\text{wait}}$ experiment, an individual’s risk preference with gains ($\text{PropRisk}_{\text{gain}}$) predicted the RT bias ($\text{RT}_{\text{riskier}}$-$\text{RT}_{\text{surer}}$) with gains (r=-0.46, P=0.03); and their risk preference with losses ($\text{PropRisk}_{\text{loss}}$) predicted the RT bias with losses (r=-0.63, P=0.002). In the $\text{Selection}_{\text{wait}}$ we see exactly the same relationship between risk preference and RT bias with both gains (r=-0.66, P=4.4x10^{-4}) and losses (r=-0.69, P=2.0x10^{-4}). Furthermore, as shown in Fig. 5.5, in both tasks and for both gains and losses, we see exactly the predicted pattern of risk slowing approach when risk is aversive, inducing no bias when risk is neutral, and speeding approach when risk is appetitive.

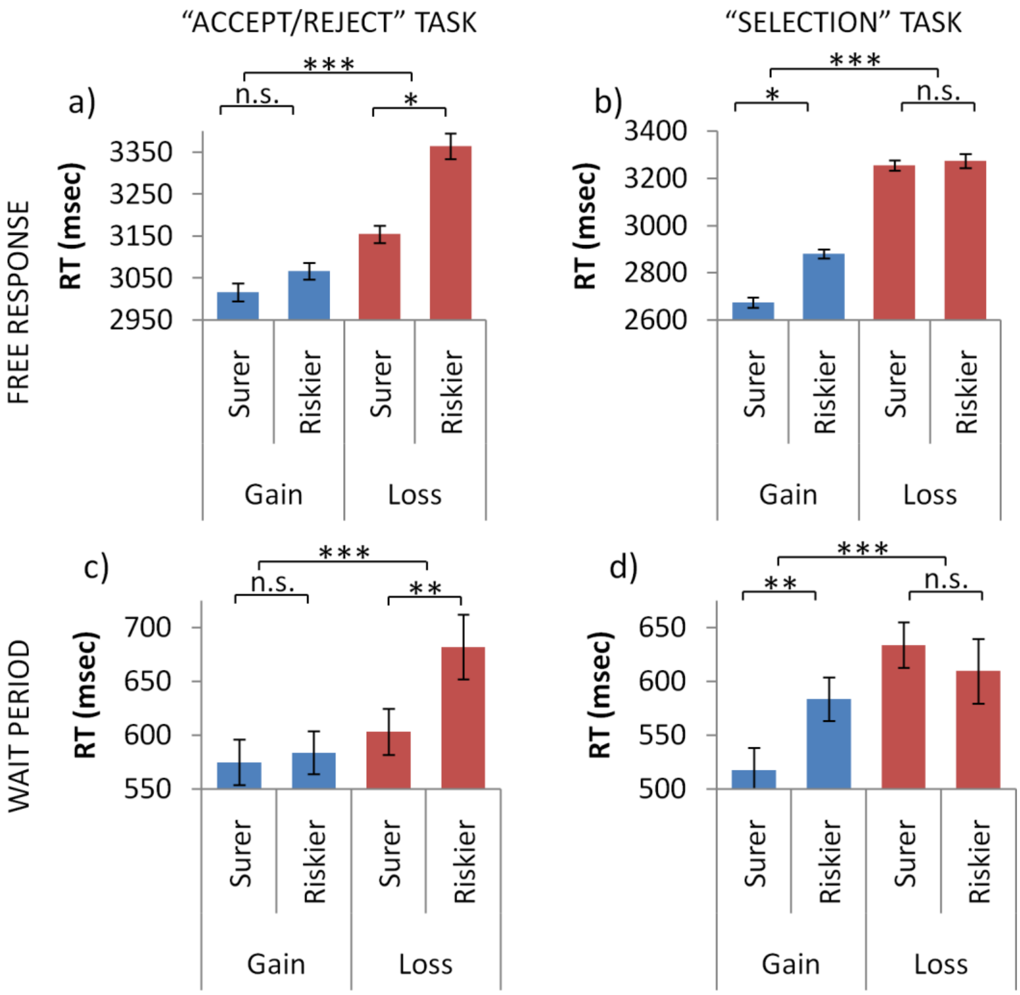

Figure 5.5 Raw RTs suggest risk and valence bias action selection rather than option evaluation. *This figure is exactly as Fig.5.3 but with raw RT displayed, to illustrate RT biases between the free response and imposed wait versions of each task – despite there being a very large difference in RT between them. Error bars show s.e.m., * P<0.05, ** P=0.005, *** P<0.0005.*

5.4 Discussion

In this chapter we tested an hypothesis that risk and valence biased choice through the approach/avoidance mechanisms suggested by our fMRI data in Chapter

4. We find RT evidence for such approach/avoidance mechanism for valence, such that individuals are slower to choose (approach) options containing loss (Fig. 5.2), regardless of whether loss induces greater gambling for gains than losses (in our “accept/reject” task) or the opposite (in our “selection” task). We also find RT evidence for such an approach/avoidance mechanism for risk, where risk can be aversive, neutral or appetitive depending upon subjective risk preference (Fig. 5.3). Furthermore, consistent these biases influencing action selection rather than option evaluation, we find the same RT biases even after an imposed wait for option evaluation that is longer than the average free response time (Fig. 5.4).

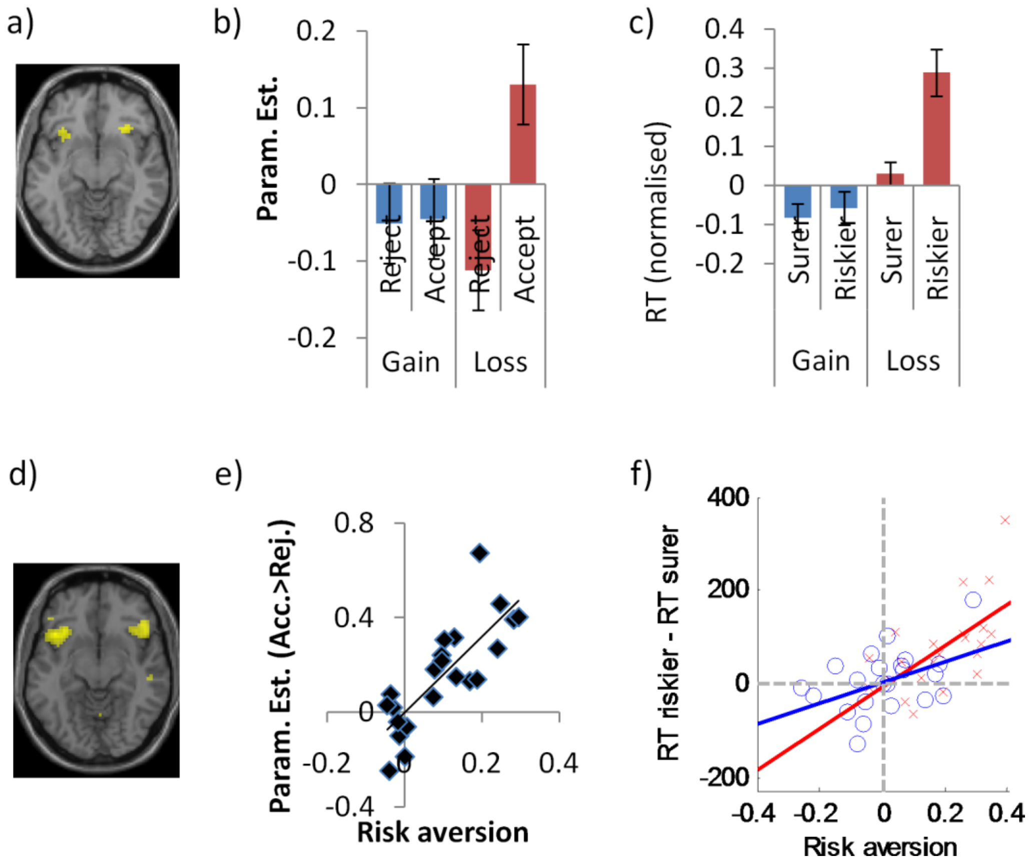

Figure 5.6 Concordant fMRI and RT evidence for approach/avoidance mechanisms. *For ease of comparison, we show fMRI data from the “accept/reject” task previously presented in the previous Chapter, and we show RT data from the “accept/reject” task previously presented in this Chapter. For valence, in anterior*

insula (panels a and b) we show increased activity when approaching the risky option with losses (the most aversive option), which is the same pattern we see in our RT data in this chapter (panel c). For risk, we see that in anterior insula the greater an individual's risk aversion the greater the activity when choosing the risky relative to sure option (panels d and e), which is the same pattern we see in our RT data in this chapter (panel f).

One potential explanation for the risk-related RT bias in the "accept/reject" task (Fig. 5.3) is that risk-preference is defined by the frequency with which the riskier option is chosen, as one might then expect more frequently performed actions to be faster. However, this cannot account for the valence related bias (Fig. 5.2). Furthermore, it cannot account for the same risk-related biases seen in the "selection" task (Fig. 5.3), where the riskier option could randomly appear on either side of the screen and was chosen by the related button press (Fig. 5.1).

Our findings can explain previously reported longer RTs for losses than gains (Dickhaut et al., 2003), although we note that a number of previous studies examining risk and valence do not report RTs (e.g. Tom et al., 2007). Interestingly, where framing is used to manipulate valence rather than actual losses, no RT difference was noted with frame (De Martino et al., 2006; Guitart-Masip et al., 2010). Finally, in support of our approach/avoidance hypothesis, we see a striking concordance between reaction time data and findings in anterior insula during the "accept/reject" task (Chapter 4 and Fig. 5.6). Together, these findings support a biologically-based explanation of choice that can account for behavioural patterns not predicted by existing economic theories.

Chapter 6. Social choice: Neural mechanisms underlying fairness in choice

6.1 Introduction

Fairness is of interest to sociologists (Homans, 1961), economists (Akerlof, 1979; Kahneman et al., 1986) and neuroscientists (Sanfey et al., 2003). Fairness reflects objective features of how people share resources, classically elicited in the Ultimatum Game (UG) where one player (the Proposer) is given an endowment (e.g. £10) and proposes a division (e.g. keep £6/offer £4) to a second player (the Responder), who can accept (both get the proposed split) or reject (both get nothing) the offer (Güth et al., 1982). In the Rational Choice Theory framework Game Theory, as described in Chapter 2 a Responder should accept any offer however low, and knowing this a Proposer should offer the lowest possible amount – but behaviour does not conform to this prediction, with low offer proportions routinely rejected (Camerer, 2003). This has been accommodated in behavioural economic models using “other-regarding preferences”. However, fairness attribution varies between contexts and individuals: labourers’ wages might not seem unfair considered alongside colleagues, yet extremely unfair alongside executives’ salaries. This contextual aspect cannot well be explained using behavioural economic models. Here, we aim to tease apart objective and contextual components of fairness, dissociating their neural substrates. Importantly, we define the contextual component of fairness as a choice bias, leaving open the question of whether subjects are subjectively aware of this shift (Pronin, 2007).

Responders in the classic UG are reported to show greater activity in anterior insula and dorsolateral prefrontal cortex (DLPFC) for lower compared to higher offers, a finding interpreted as reflecting fairness and cognitive-control respectively (Sanfey et al., 2003). However, as the classic UG cannot dissociate objective and contextual fairness, alternative approaches have endeavoured to isolate components of fairness in the UG. One attempt to unconfound fairness from offer amount treated it as synonymous with offered endowment proportion, implicating lateral PFC in cognitive control (Tabibnia et al., 2008). An alternative strategy manipulated the stimuli used, where by changing Proposer intentionality anterior insula cortex was implicated in fairness responses (Güroğlu et al., 2010). Outside the UG framework others have investigated reward comparison (Fliessbach et al., 2007) and fairness in third-party decisions (Hsu et al., 2008), with the latter demonstrating that posterior (but not anterior) insula tracked an objective measure of fairness, namely inequality. However, an isolated fairness manipulation has not as yet been reported.

Here, we contextually manipulate the fairness of a set of offers from a group of Proposers by presentation: alone; interleaved with higher offers from different Proposers; or interleaved with lower offers. Using a formal inequality aversion model (Messick and McClintock, 1968; Fehr and Schmidt, 1999) we isolated objective and contextual fairness components together with economic self-interest, and apply this model's parameters to fMRI data. Behaviourally, we predicted increased acceptance when offers were contextually perceived as fairer, despite being objectively identical. Neurally, we predicted anatomical dissociations between cognitive control in prefrontal and fairness in insula cortex. Within insula itself, a broad-based framework suggests a role for posterior insula encoding more primary quantities, mid-insula in contextual integration (Craig, 2002, 2009) and anterior insula in introspective awareness of emotion and bodily-state (Critchley et al., 2004; Paulus and Stein, 2006; Singer et al., 2009b). Also within insula, previous studies suggested objective inequality is encoded posteriorly (Hsu et al., 2008) and an integrated fairness metric anteriorly (Sanfey et al., 2003). Therefore, within insula we predicted objective and contextual aspects of fairness would map to posterior and mid/anterior regions respectively.

6.2 Materials and methods

6.2.1 Subjects

32 healthy, right-handed subjects participated in the study, 16 of whom played an Ultimatum Game (UG) during fMRI scanning (10 male; age 18-30; mean age 21.8) and 16 undertook exactly the same task outside the scanner (4 male; age 18-26; mean age 20.6). Two further subjects were excluded from the study, one because he could not tolerate the MRI scanner and the other because he did not understand the task. All subjects provided informed consent and the study was approved by the Institute of Neurology (University College, London) Research Ethics Committee.

6.2.2 Experimental design

Subjects underwent fMRI scanning or behavioural testing as Responder in an UG. The general form of the UG with two players is as follows. Initially, one player (the Proposer) is given an endowment (e.g. £8) and makes an offer to the second player (the Responder) about how to split the endowment (e.g. £6 for the Proposer and £2 for the Responder). The Responder then chooses to accept the offer (both get the split as proposed) or to reject it (both get nothing).

To reinforce the social nature of the task, two subjects always attended each experimental session and at the outset were seated together. Prior to data collection the pair were assigned to separate testing rooms. Subjects understood that they were responding to offers made by past and present participants, who had been placed into one of three coloured Proposer groups (blue, yellow and orange) based on their answers to two questionnaires. The questionnaires were chosen as they quantify the tension between self- and other-regarding motivations [Machiavellianism IV (Christie and Geis, 1970); and Social Value Orientation (Van Lange et al., 1997)]. In each trial, subjects were told which group an offer came from but were not provided with information regarding which specific member. Subjects were told each group comprised the six preceding participants classified into that group and, additionally, that one group would also contain the other attendee that day. Subjects were shown photographs of group members and asked if any were known to them, and subjects also had their own photograph taken for use in future testing with other participants.

This group format was motivated by a need to ensure that, as far as possible, subjects treated each trial individually thereby preventing temporal dependencies in choice with small numbers of individual Proposers gaining reputations over repeated proposals (e.g. earlier lower offers leading to later rejections of higher offers to punish that individual). The group format also avoided a need to present a less plausible scenario that large numbers of subjects had previously attended the experiment. However, in reality the three Proposer groups comprised three sets of 25 offer proportions. Each set spanned a full range from around 0.10 to 0.50 of the endowment with the intention that subjects consider each individual offer and not deterministically accept or reject all offers from a particular group. The behavioural regularity from experimental economics is that offers below 0.25 are rejected about half the time (Camerer, 2003). The “M set” offers were concentrated around this point to maximise our sensitivity to contextual changes in acceptance rates. The “L set” had a mean offer proportion of 0.21; the “M set” had a mean offer proportion of 0.30; and the “H set” had a mean offer proportion of “0.40”. The means of the “L set” and “H set” were chosen to induce the context effects on the “M set” described below.

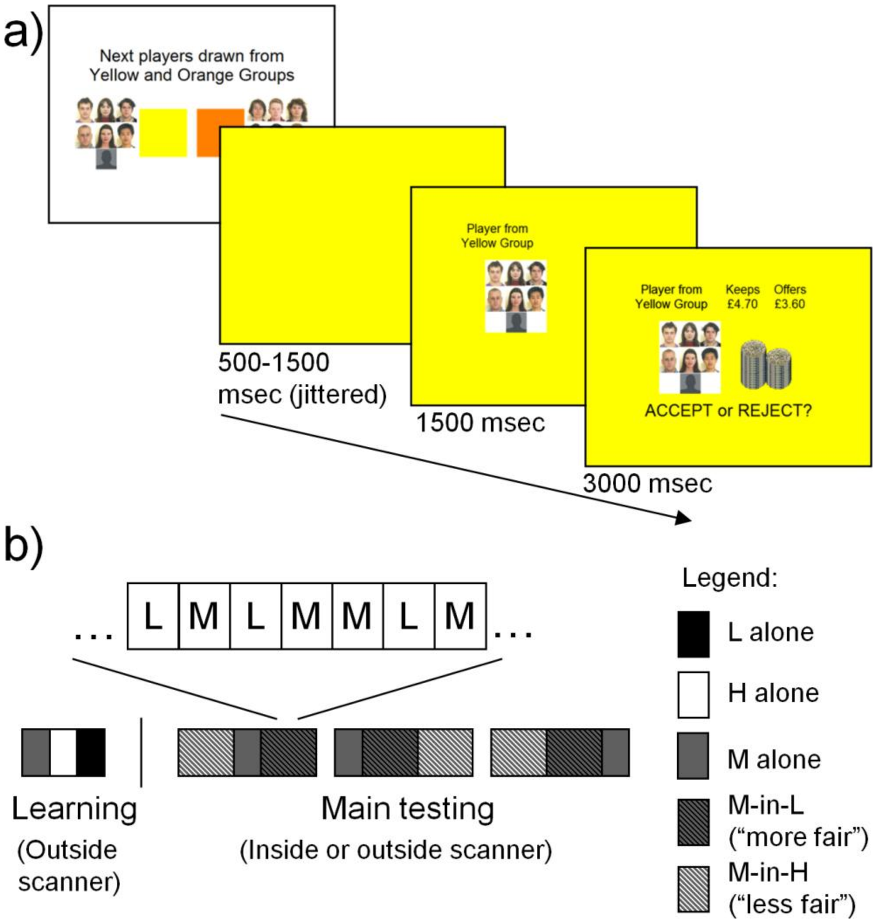

Figure 6.1 Illustration of experimental design. a) *Timeline of trials. Illustrated is the task from the perspective of the Responder: firstly a blank screen is presented for 500-1500msec (mean 1000msec) in the colour of the Proposer group (L, M or H); second, a panel containing the photographs of the Proposer group was added for 1500 msec; and third the proposal was then shown for 3000msec (denoted both numerically and visually by the height of the coin stacks) along with the instruction to accept or reject (side counterbalanced between subjects). Subjects understood that the silhouette represented the other subject attending that session, who had been placed in one of the three Proposer groups. During the 3000msec in which the proposal was shown, subjects had to decide by a button press whether to accept or reject the offer. Subjects saw a brief screen with "REST" displayed every 8-9 contiguous trials before an introductory screen announced the group or groups whose offers would be presented next (i.e. M group only; M and H groups; M and L groups).* ***b)*** *Order of conditions. Initially, in a reputation learning session performed outside the scanner subjects responded to the full set of 25 offers from the M Proposer group alone (grey in panel b), the 25 H offers alone (white in panel b) and the 25 L offers alone (black in panel b). Subjects then underwent the main testing session in the scanner or behaviourally, which comprised 3 runs. In each run of 125 trials, subjects responded to the M set of offers in 3 contexts: alone ("neutral"); interleaved with the L set (M-in-L; "more fair"); and interleaved with the H set (M-in-H; "less fair").*

The full sets of offers are as follows: L = {0.08, 0.08, 0.09, 0.09, 0.1, 0.1, 0.1, 0.15, 0.15, 0.15, 0.16, 0.16, 0.17, 0.17, 0.2, 0.21, 0.22, 0.26, 0.27, 0.3, 0.31, 0.37, 0.4, 0.5, 0.5}; M = {0.08, 0.1, 0.16, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, 0.26, 0.26, 0.27, 0.28, 0.29, 0.3, 0.31, 0.32, 0.36, 0.37, 0.4, 0.46, 0.5, 0.5, 0.5, 0.5}; and H = {0.1, 0.15, 0.21, 0.27, 0.3, 0.35, 0.36, 0.36, 0.37, 0.4, 0.4, 0.41, 0.42, 0.42, 0.45, 0.46, 0.47, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5}.

Our prime interest in this experiment related to subjects' responses to the "M set" where our key manipulation of contextual fairness meant that subjects saw the "M set" of 25 offers in three different contexts (Fig. 6.1, panel b). This change in context was expected to bias subjects' choices in relation to what was otherwise an objectively identical M set. The specific contexts were: 1) "M alone" in which the full set of 25 M offers was shown on its own in random order with no contextual manipulation; 2) "M-in-H" in which the full set of 25 M offers was shown interleaved with the full set of 25 H offers (i.e. there were 50 offers shown in random order; note that on each individual trial subjects were told which set the offer emanated from); and 3) "M-in-L" with the full set of 25 M offers interleaved with the full set of 25 L offers. The main testing session comprised three runs of trials as Responder. In each run the full set of 25 M offers was presented three times, once in each context (Fig. 6.1), and the order of these conditions was counterbalanced within and between subjects. Therefore, in this main session there were a total of 375 trials consisting of: 75 M-in-L; 75 M-in-H; 75 M-alone; 75 L-in-M; and 75 H-in-M trials.

During each trial subjects (Responders) first saw from which group a proposal emanated and then saw the proposed division of the endowment in that trial (Fig. 6.1). They then indicated, via a button press, their decision to either accept (both get the split as proposed) or reject (both get nothing) the offer. The endowment was varied in increments of £0.10 from around £7.70 to £9.80.

For each subject the experiment comprised four consecutive phases. Initially, subjects completed questionnaires and were then informed of the nature of the task. Secondly, subjects made 20 proposals. Thirdly, to learn the reputations of the three groups (L, M and H), subjects (as Responders) played against the full set of 25 offers in each group separately. Fourthly, the main testing session comprised three runs of trials as Responder and a total of 375 trials. To provide subjects with rest periods of approximately six seconds and to militate against fatigue, subjects saw a brief screen with “REST” displayed every 8-9 contiguous trials. After each rest period an introductory screen announced the group or groups whose offers would be presented next (i.e. M group only; M and H groups; M and L groups). Of the two subjects attending each experimental session the one with fewer rejections in the learning phase continued the main testing phase in the behavioural testing room whilst the second subject conducted this main phase in the MRI scanner. This was to increase power to detect fairness related activation independently of choice and also to avoid scanning subjects with a deterministic strategy of accepting all offers. Subjects were informed that this selection had been made at random. At the end of the experimental session subjects were debriefed and all reported believing that the proposals they faced were made by present and past participants. The subject’s payment was determined by responses and proposals chosen at random. Subjects received on average around £30 (~$50). All statistical tests were two-tailed.

6.2.3 Behavioural modelling

We fit data on an individual subject basis both according to a psychometric model and an economic model. In our psychometric analysis we modelled changes in the proportion of acceptance as offer proportion increases with logistic regression:

$$P(accept) = \frac{1}{1+e^{-(b_0+b_1 z)}} \qquad \text{Eq. 6.1}$$

In this model z is the offer proportion (binned into groups of 5 trials). We estimate the model separately for “M” offers in each of the three offer contexts (M-alone, M-in-L and M-in-H).

In our economic analysis, we modelled individual choices using a binary logistic regression utility model. Various utility functions with other regarding preferences have been proposed (Messick and McClintock, 1968; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) that make similar predictions in the UG and we chose the following formulation:

$$P(accept) = \frac{1}{1+e^{-\frac{U}{\lambda}}} \qquad \text{Eq. 6.2}$$

Where $U = x_{self} - \alpha^*(x_{other} - x_{self}),$ $\alpha \geq 0; \lambda \geq 0$

Eq. 6.3

x_{self} is the amount the Proposer offered in the trial and x_{other} is the amount the Proposer keeps, α is an ‘envy’ parameter (reflecting a tradeoff between inequality and self interest), and λ reflects choice randomness. In this model, there is no constant term as we assume a utility of 0 represents indifference between acceptance and rejection of an offer (i.e. rejection of an offer has a utility of 0). For illustration, in a single trial as Responder, the utility of the offer is calculated by combining the self-regarding component (amount to self) and the other-regarding component (the weighted impact of inequality). This utility is then compared to the utility of rejecting (zero), with the offer being accepted if greater and rejected if lesser (with stochasticity in action selection captured by λ). To further illustrate the other-regarding component of the utility function, we used an objective metric for unfairness (inequality) weighted by the α (“envy”) parameter. For example, if α=0 the subject is entirely self-regarding (they will accept any offer however small) and if α is large they will reject lower offers. Because in this study x_{self} did not exceed x_{other}, the "guilt" parameter contained in some formulations was not used (Fehr and Schmidt, 1999). We optimised subject-specific *α* and *λ* parameters across all trials using nonlinear optimization implemented in Matlab for maximum likelihood estimation.

We also estimated parameters for each subject in the M-in-H and in the M-in-L conditions to examine context effects. 30 of the 32 subjects were included as 2 behavioural subjects accepted essentially all offers such that parameters could not be estimated. No difference was found when a paired t-test was used to compare either *α* or *λ* from the two contexts obtained when both parameters were freely estimated for each subject in both conditions. As we were interested in the relative contributions of both parameters, we fixed *λ* as the mean *λ* in the two conditions for that subject. As an additional check, we confirmed that this result held when *λ* was fixed at the average for these two conditions across all subjects. Finally, we performed the same procedure but fixed *α* and estimated *λ*.

6.2.4 fMRI data acquisition

Images were acquired using a 3T Allegra scanner (Siemens, Erlangen, Germany). BOLD sensitive functional images were acquired using a gradient-echo EPI sequence (48 transverse slices; TR, 2.88 secs; TE, 30 ms; 3 x 3 mm in-plane resolution; 2 mm slice thickness; 1 mm gap between adjacent slices; z-shim -0.4 mT/m; positive phase encoding direction; slice tilt -30 degrees) optimised for detecting changes in the OFC and amygdala (Weiskopf et al., 2006b). Four runs of 284-286 volumes were collected for each subject, followed by a T1-weighted anatomical scan. Local field maps were also acquired.

6.2.5 fMRI data analysis

Functional data were analysed using standard procedures in SPM5 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). fMRI timeseries were regressed onto a composite general linear model (GLM) containing delta (stick) functions representing the onsets of the offer. These delta functions were convolved with the canonical HRF and its temporal derivative. The stimulus delta functions were separated into five regressors depending on the Proposer type (L-in-M, M-in-L, M-alone, M-in-H and H-in-M). Each was then parametrically modulated by two orthogonalised regressors entered in the following order: offer amount followed by inequality ($x_{other} - x_{self}$). A second GLM with a 2 (accept, reject) by 5 (Proposer type) factorial design matrix was also constructed to determine the betas for the components of the interaction of choice and context. Throughout the analysis of the imaging data the main effects were calculated using the data across all five Proposer types, whilst the effect of context was probed within the comparison of M-in-L and M-in-H. These two conditions were fully matched in terms of offer set (i.e. M offers) and task demands (unlike M-alone offers, which were not interleaved with another trial type).

Cluster-based statistics were used to define significant activations both on their intensity and spatial extent (Friston et al., 1993). Clusters were defined using a threshold of $P < 0.005$ and corrected for multiple comparisons using family-wise error correction and a threshold of $P < 0.05$. For presentation purposes, images are displayed at $P < 0.001$ uncorrected unless otherwise stated. We report all activations at $P<0.05$ that survive whole brain correction using family-wise error at the cluster level, unless otherwise stated.

6.3 Results

6.3.1 Responder behaviour

To confirm consistency of our results with well known behavioural regularities in the UG (Camerer, 2003), we first collapsed across all Proposer group sets. In both the learning and main sessions all subjects accepted almost all offers of half the endowment, and 29 of the 32 subjects rejected almost all offers under one tenth of the endowment. In the main session the mean acceptance rate was 0.50 (s.d. 0.20) (Fig. 6.2). Using an analytic framework derived from psychophysics we show a graded relationship between increasing offer size and increasing acceptance rate, at both individual (Fig. 6.3) and group levels (Supplemental Fig. 6.2). Moreover, such a framework predicts that reaction times (RTs) should be greatest at the point of equality in value as this represents the point of maximum decision uncertainty (Grinband et al., 2006), which is exactly what we observed (Fig. 6.3).

To test our key behavioural prediction of a contextual bias in fairness perception, we focused on the critical M set trials from the main session. Here the set of M offers was presented alone ("M-alone"); interleaved with H offers ("M-in-H"; contextually less fair) or interleaved with the L offers ("M-in-L"; contextually more fair). When all 32 subjects were included in the analysis we observed effects of social context on the acceptance rate of otherwise identical offers (one way ANOVA; F(2,62)=4.88, *P* =0.013), driven by a highly significant difference between M-in-L trials (seen as "fairest"; mean 48.5%) relative to M-in-H trials (seen as "less fair"; mean 45.8%; paired two-tailed t-test, t(31)=2.86; *P* =0.007; Fig. 6.2).

16 of the 32 subjects who comprised our behavioural sample also performed the task in the fMRI scanner. The same pattern was seen when the 16 scanned subjects were analysed separately for the key comparison of M-in-L with M-in-H (t(15)=2.64, *P*=0.019; M-in-L mean 39.3%; M-in-H mean 36.3%) with trend-level significance for a

one-way ANOVA across all three contexts (F(2,30)=3.21, *P* =0.074). This contextual bias was primarily driven by a lower rate of acceptance in the “less fair” (M-in-H) condition, with a significant difference between M-in-H and M-alone (all subjects t(31)=2.18, *P*=0.037; scanned subjects t(15)=2.54, *P* =0.023) but not between M-in-L and M-alone (*P* >0.1).

6.3.2 Proposals and questionnaires

The proposals our subjects made were consistent with those seen in previous studies (Camerer, 2003). The mean proportion offered was 0.44 (s.d. 0.09) over all 32 subjects, and there was no significant difference (two-tailed ttest, p>0.1) between the group of 16 scanned subjects (mean 0.46; s.d. 0.10) and the 16 subjects who solely underwent behavioural testing (0.42; 0.07). There was also only small variation within each subject’s offers and no significant effect of trial number on the mean offer.

Mach IV questionnaire scores (Christie and Geis, 1970) were typical of normal populations (mean 95.4; stdev 11.9; n=32); with no difference between scanning group and behavioural groups (97.8 (11.4) vs. 93.1 (12.3), two-tailed ttest, p>0.1). Using the Van Lange social value orientation questionnaire more subjects were classified as prosocial in the scanning than in the behavioural group (Van Lange et al., 1997). In the scanning group 8 were classified as prosocial, 5 as individualist, 2 as competitive and 1 was not classifiable, whilst in the behavioural group 4 were prosocial, 4 individualist, 4 competitive and 4 not classifiable. During the learning session subjects learnt the reputations of the three opponent groups (L, M and H), with only 6 of the 32 subjects (2 in the scanning group) incorrectly ranking the L, M and H groups on a visual analogue scale from “most unfair” to “most fair”.

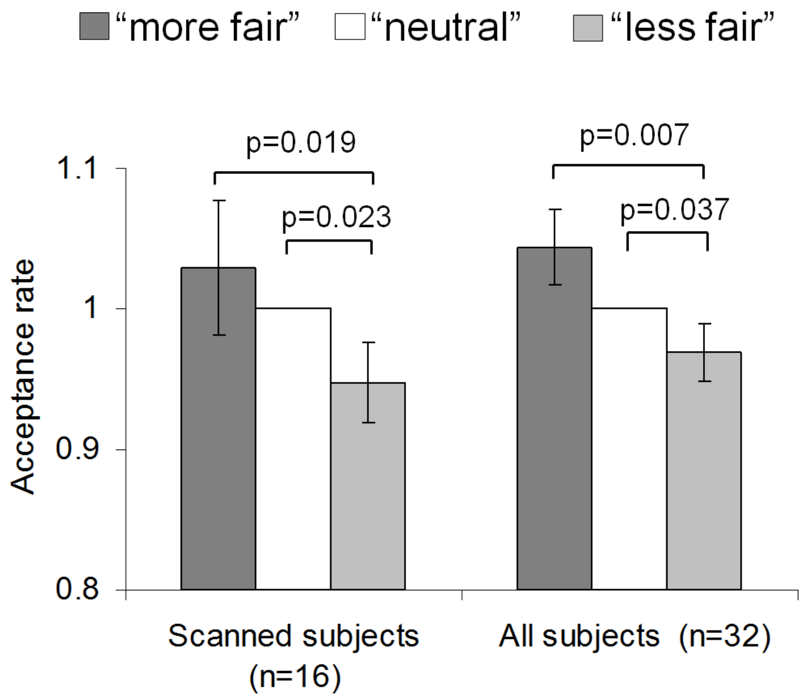

Figure 6.2 Biased acceptance of objectively identical offers by contextual manipulation. *Each bar represents rate of acceptance (+/- s.e.m) of the “M” set of offers that spans the full range of offer proportions from 0.08 to 0.50. The M set is presented in three different contexts, representing a manipulation of contextual fairness: M-in-L (“more fair”, interleaved with lower offers); M-alone (“neutral”, presented alone); and M-in-H (“less fair”, interleaved with higher offers). The rate of acceptance is normalised with respect to the rate of acceptance in the neutral M-alone condition. Data is shown for the main session for the scanned group (n=16) and the combined data from all subjects including both the scanned subjects and those who solely underwent behavioural testing (n=32).*

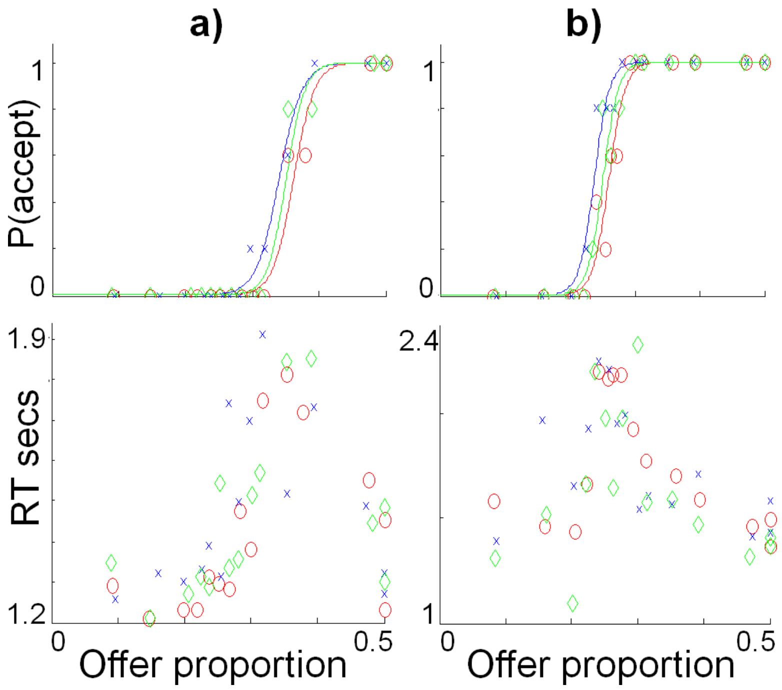

***Figure 6.3 Psychometric analysis of individual subject data.** Data from the scanning session is shown for two exemplar subjects (a and b). For each subject the upper panels show the probability of acceptance and the lower panels show reaction times (RTs), both plotted against offer proportion (each data point being the mean of five trials). Data is shown for the "M" offers in the three contexts: M-alone (green; "neutral"); M-in-L (blue; "more fair"); and M-in-H (red; "less fair"). The upper panels also show probability of acceptance as a logistic function fitted to those points, demonstrating a contextual bias evident in a shift to the left for contextually more fair (M-in-L) and a shift to the right for the less fair (M-in-H) relative to the control condition (M-alone). In the lower panels it can be seen that the point of indifference (probability of acceptance is 0.5) corresponds to peak RTs, consistent with choice difficulty being greatest at this point and arguing against either acceptance or rejection as being a "default" choice .*

6.3.3 Modelling Responder behaviour

We next tested if a formal economic model predicted subjects' choices as Responder (Bolton and Ockenfels, 2000;Charness and Rabin, 2002;Fehr and Schmidt, 1999;Messick and McClintock, 1968). The intention here was to derive model parameters as a bridge to underlying neural mechanisms. The model is specified by the following formalism:

$$U = x_{self} – \alpha^*(x_{other} – x_{self})$$

In this model, the utility (U) a subject derives from a proposal is given by the subject's payoff (x_{self}) minus the weighted inequality of the proposal, with inequality being the amount kept by the Proposer (x_{other}) less the amount offered. The α ("envy") parameter quantifies how much a particular subject cares about inequality (i.e. how much they weight this social component, e.g. if α=0 they are entirely self-regarding). We combined this model of fairness preference with a logistic model of stochastic choice with a noise parameter, λ, and estimated both parameters using maximum likelihood estimation.

We first collapsed across all trial types. Across all 32 subjects mean α was 0.89 (s.d. 0.54; range 0-2.59) and mean λ was 0.54 (s.d. 0.31; range 0-1.26). For the 16 scanned subjects, mean α was 1.08 (s.d. 0.55; range 0.38-2.59) and mean λ was 0.51 (st. dev. 0.31; range 0.20-1.26). These observations are consistent with previous studies using similar model(Fehr and Schmidt, 1999; Krajbich et al., 2009). A statistical measure of how well the model predicted subjects' choice is provided by a likelihood ratio test: comparing the full to a reduced model without the inequality term was highly significant (all 32 subjects $X^2(1)>10^4$, P <0.001; 16 scanned subjects $X^2(1)>6000$, P <0.001).

We next examined the contextual bias in fairness perception by estimating α and λ in the M-in-H and M-in-L conditions for 30 subjects (2 subjects in the behavioural group accepted essentially all offers; the remaining 30 subjects had mean α=0.95, mean λ=0.56 when collapsed across all trial types). There was a significant difference in *α* between the M-in-H and M-in-L conditions when the λ parameter used for estimation was fixed at either the mean λ in the two conditions for that subject (mean $\alpha_{\text{M-in-H}}$=1.01; mean $\alpha_{\text{M-in-L}}$=0.91; t(29)=2.246, *P*=0.032) or the average for these conditions across all subjects (mean $_{\text{M-in-H}}$=0.97; mean $\alpha_{\text{M-in-L}}$=0.89; t(29)=2.157, *P*=0.039). λ did not significantly differ between conditions when *α* was fixed in this fashion.

6.3.4 Neuroimaging

16 of the subjects performed the task in the fMRI scanner as Responder and we report only activation surviving whole brain correction at the cluster level unless otherwise stated. Initially, in a conventional factorial analysis we determined whether our behavioural bias in choice (accept or reject) driven by a change in social context (M-in-L or M-in-H) was reflected in brain activity. We computed the interaction term [(Acc$_{\text{M-in-H}}$ + Rej$_{\text{M-in-L}}$) – (Acc$_{\text{M-in-L}}$ + Rej$_{\text{M-in-H}}$)], which revealed an effect in right DLPFC (middle frontal gyrus; Table 1). Examination of simple effects showed that this interaction was driven by greater activity for rejecting contextually fairer offers (i.e. M-in-L) than rejecting contextually less fair offers (i.e. M-in-H; Fig. 6.4; paired two-tailed ttests rejection t(15)=3.3, P=0.005, acceptance t(15)=0.04, P>0.9).

In this factorial analysis, we also observed a main effect of choice in increased activation for accepting, relative to rejecting, offers in bilateral supplementary motor area (SMA) and pre-SMA (Table 1). The pre-SMA and SMA have distinct anatomical connections and in imaging studies the vertical commissure anterior (VCA) line is often used to distinguish the precise source of activation (Nachev et al., 2008). Interestingly in light of previous findings, the peak voxel of this cluster was anterior to the VCA in pre-SMA, a region with strong connections to DLPFC (Nachev et al., 2008). No activation survived cluster level correction for the reverse contrast (reject>accept) or for the main effects of context (M-in-H > M-in-L or M-in-L > M-in-H).

We next used a parametric analysis to isolate activity correlating with specific components of our formal economic model and their relationship to social context. Offer amount and inequality were included as parametric regressors, and the subject-specific envy (α) parameter was included as a second level covariate on the inequality regressor. We orthogonalised inequality with respect to offer amount in order to identify its independent contribution to the BOLD signal having accounted for activity related to the offer magnitude. We examined the main effect of inequality, our objective metric of fairness, by collapsing across all trials (L; H; M alone; M-in-L; M-in-H) and this analysis showed a negative correlation with inequality in right posterior insula (i.e. greater activation for a more equal allocation; Table 6.2 and Fig. 6.5). A similar pattern was evident on the left which did not survive cluster level correction for the whole brain (x=-45, y=-15, z=3, Z=4.01, 58 voxels at *P*<0.005 uncorrected).

We next asked where in the brain inequality is integrated with context to produce changes in contextual fairness. The key contrast in this analysis was to compute the interaction between the parametric regressor for inequality in the more (M-in-L) versus the less fair (M-in-H) context [$\text{inequality}_{\text{M-in-L}}$ > $\text{inequality}_{\text{M-in-H}}$]. This interaction showed activation in bilateral mid-insula and rolandic operculum, extending on the right into posterior insula and inferior parietal cortex (Table 6.2 and Fig. 6.5). Intuitively, this interaction can be expressed as a difference in regression slope of activity under both levels of the categorical factor (Toga and Mazziotta, 2002) (see Fig. 6.5). Interestingly, in light of our inequality aversion model, inequality modulated insula activity in the condition when offers are viewed as more aversive (M-in-H; Fig. 6.5). In fact, the observation that the M-in-H condition drove the interaction effect exactly mirrors our behavioural finding where this factor was the principal driver of a contextual bias (Fig. 6.2).

For comprehensiveness we also tested for areas correlating with inequality in the matched M-in-H and M-in-L conditions alone. Here, we demonstrated a similar negative correlation with activation in posterior insula to that seen when including all trials ($P < 0.05$, cluster-level whole-brain corrected on the left, Table 2; 68 vox. at $P<0.005$ unc. on the right at x=42 y=-24 z=27, Z=3.51) with additional activation seen in STS close to the location of activation reported in “theory of mind” (TOM) tasks (Grèzes et al., 2004; Gobbini et al., 2007). Separately, we observed an interaction of offer amount with social context in right DLPFC [offer amount$_{\text{M-in-L}}$ – offer amount$_{\text{M-in-H}}$] Table 2), which in light of the fact that choice is highly correlated with offer magnitude converges with the pattern identified by the interaction term in our factorial design (Fig. 6.4).

Finally, we examined neural correlates of how much a particular subject cares about what others receive, regardless of the particular contextual manipulation. In our economic model this is captured by the α (“envy”) parameter derived from each subject’s behaviour, which specifically weights the inequality component of the utility function. The magnitude of the α parameter correlated across subjects with activity in the precuneus ($P < 0.05$, voxel-level whole brain corrected), left frontopolar region ($P < 0.05$, voxel-level whole brain corrected) and left temporo-parietal junction (cluster level corrected as above).

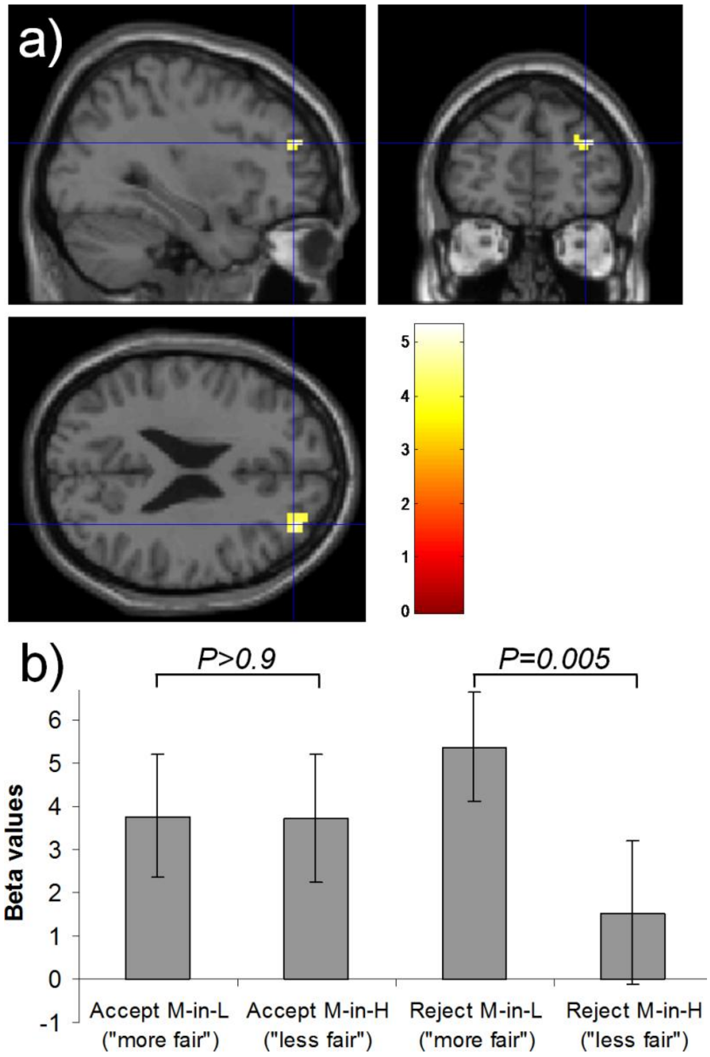

Figure 6.4. Interaction of choice and context in right dorsolateral prefrontal cortex a) *In our factorial analysis of the fMRI data there was an effect in right DLPFC for the interaction of choice (accept > reject) and context (M-in-H > M-in-L). This activation survived whole-brain cluster-level correction (P<0.05 FWE corrected; threshold of P<0.005 used to define the cluster) and is displayed at P<0.001 (uncorrected) on slices through the peak voxel (x=33, y=48, z=24).* ***b)*** *To illustrate which differences drive the interaction we plot the regression coefficients (beta values +/- s.e.m.) for each condition versus baseline at the peak voxel for the interaction. Greater activation is seen for rejecting offers perceived as more fair than for those perceived as less fair (t(15)=3.3, P=0.005, two-tailed), with no difference during acceptance (t(15)=0.04, P>0.9, two-tailed).*

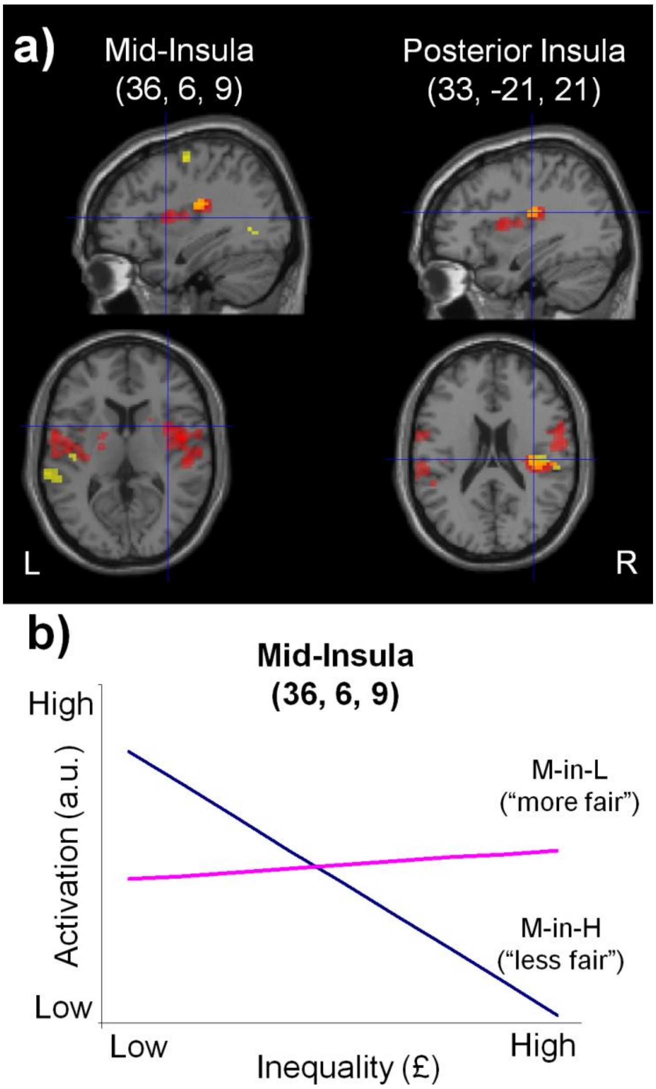

***Figure 6.5 Neural responses to inequality and its interaction with social context in insular cortex**. a) Sections showing activation related to processing of inequality in the insula. In yellow we show posterior insula is negatively correlated with inequality when collapsed across all trials (L, M-in-L, M-alone, M-in-H, H). In red are shown areas of differential activity in mid and posterior insula when contrasting responses to inequality in the more fair (M-in-L) versus the less fair (M-in-H) context. Areas of overlap are shown in orange. The right panels are at the peak of activation for the negative correlation, whilst the left panels show slices at the peak for the interaction. The activations including these peaks survive whole-brain cluster-level correction (P<0.05 FWE corrected; threshold of p<0.005 used to define the clusters) and are displayed at p<0.005 (uncorrected, 10 voxel threshold). **b)** Visualisation of the effects driving the interaction of inequality with social context at the peak voxel in mid-insula (x=36, y=6, z=9). The slope of the regression line for activation (in arbitrary units) against inequality (£) is shown for each social context (M-in-H and M-in-L). The difference in slopes illustrates that inequality modulates insular activity only when offers are viewed more aversively (M-in-H). The slope of each regression line is the parameter estimate (beta) for each condition obtained from the general linear model*

used to analyse the fMRI data, averaged across subjects (M-in-H beta=-2.92 with s.e.m.=0.64; M-in-L beta=0.32 with s.e.m.=0.63).

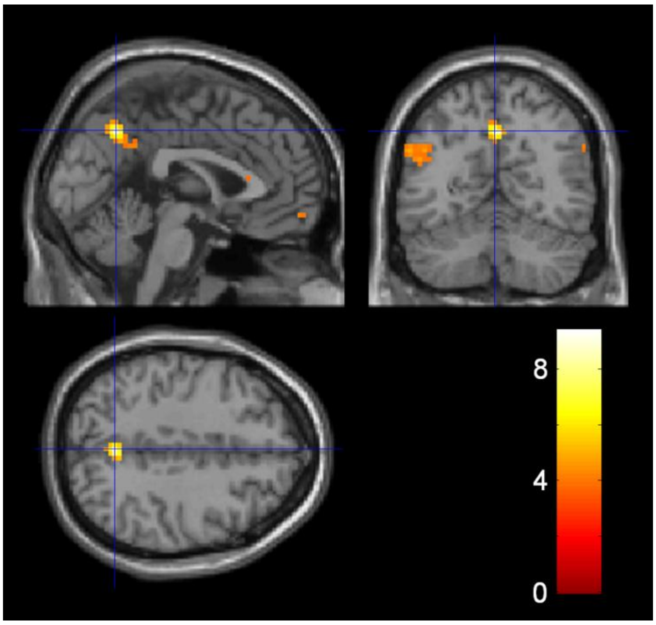

Figure 6.6 Inequality and envy. *For each subject we calculated an α (envy) parameter by fitting an inequality aversion model to behavioural data of the Responder. In this economic model, utility is the offer amount minus inequality (defined as amount to other minus amount to self) and inequality is weighted by α. Given their close relationship, the envy parameter was used as a second level covariate on the main effect of inequality. Areas highlighted are those that correlate between subjects with the degree to which the Responder cared about how much money the other person stood to gain relative to themselves. Activation surviving whole brain correction at the voxel level (P<0.05 FWE) was seen in precuneus (x=0, y=-60, z=42) and L frontopolar cortex (x=-18, y=60, z=18), whilst activation surviving cluster level correction (P<0.05 FWE corrected; threshold of P<0.005 used to define the clusters) is seen in the L angular gyrus (inferior parietal cortex) (x=-54, y=-60, z=30). The data are displayed at P<0.001 (uncorrected) at the peak voxel of the precuneus activation.*

6.4 Discussion

Our principal aim in this study was a behavioural and neural characterisation of objective and contextual aspects of fairness. We defined the contextual component of fairness as a shift in choices in response to otherwise identical offers, while remaining agnostic to the question of conscious awareness of this shift. Our finding of a marked context-dependence provides a perspective on fairness as a relative rather than absolute quantity, echoing findings in relation to other high-level quantities such as valuation (Ariely et al., 2006; Seymour and McClure, 2008; Vlaev et al., 2009). However, our neural data also highlight a fundamental role for objective social inequality that accords with effects seen in the UG across diverse cultures (Henrich, 2004), in human infants (Fehr et al., 2008) and in similar tasks in non-human primates [(Brosnan and De Waal, 2003), but note(Jensen et al., 2007)]. Our data highlights how these objective and contextual aspects interact to construct a fairness motivation with sufficient flexibility to enable appropriate responses to the social environment.

Fairness relates to how intentional agents should divide resources amongst potentially entitled recipients (Kahneman et al., 1986). Inequality aversion quantifies how this motivation influences choice (Messick and McClintock, 1968). Here, we assume choice is the outcome of processes whose neural implementation may involve social computations such as prediction errors (Behrens et al., 2008; Hampton et al., 2008). We also characterise objective and contextual components of fairness in decision-making by combining a standard economic inequality-aversion model (Fehr and Schmidt, 1999;Messick and McClintock, 1968) with psychophysical methods and the psychological concept of cognitive control (Miller and Cohen, 2001; Gilbert and Burgess, 2008).

Our neuroimaging data strongly support inequality aversion models: first, we find a main effect of inequality in posterior insula; second, between subjects the envy parameter correlates with activity in the precuneus, left TPJ and frontopolar cortex; third, inequality modulated posterior and mid-insula activity more strongly when inequality is psychologically more aversive (M-in-H; Fig. 6.5). Our findings extend current inequality aversion models, demonstrating the behavioural and neural flexibility to avoid knee-jerk aversion to inequality.

We found no neural correlate for a self-interested component of the utility function, a result that accords with previous UG studies that, as indeed was also the case here, did not provide reward-related feedback during the task (Sanfey et al., 2003; Halko et al., 2009; Güroğlu et al., 2010). Conversely, robust reward-related activity has been seen in social comparison tasks when feedback was given in an estimation task (Fleissbach et al., 2007) or where, analogous to feedback, subjects were given an outcome in every trial (Tricomi et al., 2010). Our data tentatively link the behavioural economic concept of “other regarding preferences” (Fehr and Camerer, 2007) and the psychological concept of “Theory of Mind” (TOM) (Premack et al., 1978; Frith and Frith, 2006). Subjects’ α (envy) parameter correlated with activity in a subset of TOM-related areas including TPJ, implicated in representing others’ intentions, and precuneus involved in perspective taking (Van Overwalle and Baetens, 2009).

Neurally, our results implicate insula cortex in fairness motivation and, combined with previous work, suggest functional segregation in this extensive (over 5cm long) and cytoarchitectonically diverse cortical region (Flynn, 1999; Varnavas and Grand, 1999). Both here, and in Hsu et al. (2008), posterior insula activity negatively correlated with inequality (see Fig 4 in Hsu et al., 2008). Hsu and colleagues asked subjects to choose between distributions of meals for African children, varying in inequality and amount. Our concordant findings are striking, as Hsu used decisions about third-parties rather than first-party decisions (e.g. in the UG), a difference markedly affecting choice in behavioural experiments (Camerer, 2003). We also find a mid-insula peak for integration of context with inequality in our UG. Anterior insula activity has been reported as higher for rejected versus accepted offers in the UG (Sanfey et al, 2003), a result replicated in a task-matched study (Halko et al., 2009), although the same contrast in other UG studies shows little activity in this region (Guroglu et al. 2010; Tabibnia et al 2008; this study). Indeed, recent work shows anterior insula activity depends on Proposer intentionality in the UG (Guroglu et al., 2010).

One resolution to these diverse findings is that distinct fairness-related processes are expressed in segregated regions of the insula. Thus, a negative correlation with inequality in posterior insula occurs when subjects can form predictions about inequality, having previously experienced group offers (in our study) or the distribution of experimental allocations (in Hsu et al., 2008). Given predictions for inequality, a more equal division than expected could engender a positive prediction error and a more unequal division a negative prediction error (see also Paulus and Stein 2006; Singer et al., 2009). This explanation for the observed negative correlation can also explain why it may not have been seen when each UG Proposer is encountered only once and fewer trials are played, as predictions cannot be formed (Halko et al., 2009;Sanfey et al., 2003). Capacity of posterior insula for high-level computation is suggested by involvement in other high-level tasks, albeit different to fairness, for example inter-temporal choice (Tanaka et al., 2004; Wittmann et al., 2007) and language perception (Jones et al., 2010). Note the peak activity for our contextual manipulation is in bilateral mid-insula, which has a role integrating representations of physiological state or feelings from the body with activity associated with awareness(Craig, 2002, 2009; Farrer et al., 2003; Tsakiris et al., 2007). Finally, anterior insula activity in some UG studies may reflect processing of disgust (Sanfey et al., 2003) or aversion to norm-violation (Guroglu et al., 2010), consistent with its role in introspective awareness of emotion (Craig 2009). Such emotional impact may be attenuated with many more trials (hundreds versus 10 human Proposals in Sanfey et al., 2003), with unfairness directed at third-parties (Hsu et al., 2008), or where much shorter trials reduce scope for introspection (36secs in Sanfey et al., 2003).

In DLPFC we noted an interaction between choice and context, driven by greater activity for rejecting contextually fairer offers than less fair offers (Fig. 6.4, Tables 6.1 and 6.2). Against a background of multiple competing motivations in the UG, our results appear consistent with previous suggestions of a role for DLPFC in cognitive control during the UG (Sanfey et al. 2003; Knoch et al. 2006). In light of this, one interpretation of the data is that increased activity in DLPFC during rejection of contextually more fair offers reflects enhanced difficulty of implementing these rejections. Such an interpretation is consistent with evidence that rTMS to right (but not left) DLPFC increases acceptance of lower proportion offers (Knoch et al. 2006), where such disruption would be expected to diminish rejection of contextually fairer (M-in-L) offers. However, this interpretation has to be tempered by consideration of the fact that if indeed DLPFC activity is specific to rejection, then this would predict uniformly greater activity in DLPFC for rejection compared to acceptance, which was not what we observed (Fig. 6.4). Understanding the dynamics of DLPFC activity in the UG is clearly highly complex and an issue for further investigation using more refined paradigms.

In conclusion, we provide evidence that objective inequality of social distributions fundamentally influences fairness behaviour; and that this inequality is flexibly integrated with social context. This account might explain otherwise hard to reconcile social phenomena. A role for objective inequality helps explain the trans-cultural phenomenon of UG rejections (Henrich et al., 2004), the importance of workplace inequality (Akerlof, 1982) and the historical attraction of political ideas stressing equality (Rousseau, 1754). However, whilst within a particular context inequality is key, relativity of fairness helps explain the importance of comparator groups in labour negotiations (Akerlof and Shiller, 2009); and the muted public response as US executive pay gradually increased over the past 20 years from around 60 to 160 times median US income (The Economist, 2007). That fairness inherently involves both objective and contextual aspects can also inform wide-ranging social debates, from labour disputes to fair structuring of tax systems. However, although these data suggest a biological basis for the fairness motivation, whilst humans bargaining over money tend to reject unfair offers, in contrast chimpanzees bargaining over primary rewards of food do not show this motivation to reject (Jensen et al., 2007). Whether fairness represents a uniquely human motivation, or whether humans would also ignore the unfairness of offers of primary rewards, such as food, water and sex, is the subject of the next chapter.

6.5 Tables

Area	**L/R**	**x**	**y**	**z**	**Z score**	**# vox, p<0.005**	**Corr. p-value**
Choice (accept > reject): Main effect across all trial types							
SMA and pre-SMA	L	-9	3	54	3.66	158	0.001
		-9	12	57	3.52		
SFG	R	24	15	54	3.65		
Choice (reject > accept): Main effect across all trial types							
Nil whole brain corrected.							
Interaction of choice and social context							
MFG	R	33	48	24	3.92	108	0.005
		27	45	15	3.54		
		15	54	30	3.44		
SFG	R	18	15	51	3.68	79	0.028
		24	9	51	3.60		
SMA	R	6	6	66	3.41		

Table 6.1 Results using a factorial model for analysis of the fMRI data. *This table shows all activation that survived cluster level correction (P<0.05 FWE corrected; threshold of P<0.005 used to define the clusters) for the following contrasts: main effects of choice (collapsed across all trials; M, L and H); main effects of context (between the two matched conditions; M-in-L and M-in-H); and interactions between choice and context [($Acc_{M\text{-}in\text{-}H}$ + $Rej_{M\text{-}in\text{-}L}$) – ($Acc_{M\text{-}in\text{-}L}$ + $Rej_{M\text{-}in\text{-}H}$)]. For each cluster is shown: the three constituent peaks with the highest Z-scores; the number of voxels at P<0.005 (uncorrected); and the P-value of the cluster after FWE correction across the whole brain. (SMA = Supplementary Motor Area; SFG = Superior Frontal Gyrus; MFG = Middle Frontal Gyrus).*

Area	L/R	x	y	z	Z score	# vox, p<0.005	Corr. p-value
Offer amount: Main effect across all trial types							
Postcentral gyrus	R	39	-30	45	4.91	536	<0.001
		33	-33	36	4.16		
SMA and pre-SMA	L	-9	3	51	4.41		
Offer amount: Interaction with social context (M-in-H > M-in-L)							
MFG	R	30	45	18	4.44	113	0.006
Inequality: Negative main effect across all trial types							
Posterior Insula	R	33	-21	21	3.73	114	0.006
		42	-24	24	3.43		
IPC	R	54	-18	30	3.38		
Inequality: Negative main effect, matched trials (M-in-H & M-in-L)							
Posterior Insula / STG	L	-45	-18	0	3.87	80	0.043
		-36	-12	3	3.25		
		-54	-12	6	3.24		
STS / MTG	R	66	-33	-6	3.55	137	0.002
		57	-30	6	3.49		
		63	-12	0	3.36		
Inequality: Interaction with social context (M-in-L > M-in-H)							
Mid-Insula / Rolandic operc.	R	48	-3	9	5.05	728	<0.001
		36	6	9	4.12		
Precentral gyrus	R	60	3	18	4.00		
MCC	L	-12	-6	39	4.46	135	0.002
SMA / MCC	R	9	-27	51	3.89		
		9	-3	45	3.45		
Supramarginal gyrus	L	-51	-27	27	4.38	512	<0.001
Mid-Insula / Rolandic operc.	L	-48	-9	12	3.93		
Precentral gyrus	L	-60	0	9	3.91		

Table 6.2 Results using a parametric model for analysis of the fMRI data. *This table shows all activation that survived cluster level correction (P<0.05 FWE corrected; threshold of P<0.005 used to define the clusters) for the following contrasts: main effects of offer amount (collapsed across all trials; M, L and H); main effects of inequality orthogonalised with respect to offer amount (firstly collapsed across all trials and secondly collapsed across the matched M-in-L and M-in-H trials); and the effects of context (M-in-L v. M-in-H) on each of these parametric regressors. For each cluster is shown: the three constituent peaks with the highest Z-scores; the number of voxels at P<0.005 (uncorrected); and the P-value of the cluster after FWE correction across the whole brain. (SMA = Supplementary Motor Area; SFG = Superior Frontal Gyrus; MFG = Middle Frontal Gyrus; MCC = Middle Cingulate*

Cortex; STS = Superior Temporal Sulcus; MTG = Middle Temporal Gyrus; STG = Superior Temporal Gyrus).

Chapter 7. Social choice: The biological limits of human responses to unfairness

7.1 Introduction

As described in the preceding Chapter, in humans fairness has been studied extensively using games played for money (Camerer, 2003). The paradigmatic example is the Ultimatum Game (UG;(Güth et al., 1982)), which to recap involves one player (the Proposer) being given an endowment (e.g. £10) and proposing a division (e.g. keep £6/offer £4) to a second player (the Responder), who then accepts (both get the proposed split) or rejects (both get nothing) the offer. In the UG with money humans typically reject low, "unfair", offers even at cost to themselves (Camerer, 2003). In contrast, with a food primary reward chimpanzees behave solely as self-interested maximisers in an UG, accepting unfair offers (Jensen et al., 2007). Here, we asked if thirsty humans make similar self-interested maximising responses to unfair offers with a primary reward of water. To maximise our power to induce self-interested behaviour we physically presented the water, which has been shown recently to increase food's propensity to trigger appetitive responses (Bushong et al., 2010), and we amplified this effect by an experimental induction of thirst. Further, we manipulated water's value by inducing different levels of thirst, and asked if fairness was traded-off against the self-interested motivation to slake that thirst.

7.2 Methods

7.2.1 Participants

21 healthy participants provided informed consent (11 male, mean age 25 (range 20-32) years; 2 further participants did not complete testing) for a study approved by University College London Ethics Committee.

7.2.2 Thirst manipulation

On the testing day, participants were asked to refrain from drinking after 08:00am and arrived at 09:00am. We manipulated thirst using saline administered via an intravenous line for 50 minutes, at a rate 0.15ml/kg/min for males and 0.12ml/kg/min for females. In a double-blind, randomised design. 11 participants received isotonic saline (0.9% NaCl) similar to normal human osmolarity, with a minimal impact on thirst; and10 received hypertonic saline (5% NaCl) that markedly increases blood osmolarity and, as a consequence, thirst (Denton et al., 1999). After infusion subjects performed one hour of non-social tasks (not reported here); then the UG; and finally waited a further hour without water.

At pre-infusion baseline ($t_{baseline}$) and the time of testing (t_{UG}) we measured subjective thirst (visual analogue scale from 0-10) and blood osmolarity (analysis by freezing point depression osmometer). Participants completed a similar session 5-7 days before but without the UG (and receiving the alternative infusion), and were unaware of the prospect of the UG until it was conducted.

7.2.3 Behavioural task

Three participants attended each session, met each other, and were then tested in separate rooms. At time of testing, t_{UG}, participants first received written instructions stating that two participants (one Proposer and one Responder) would be randomly selected to play an UG, in this case dividing 500ml of water for immediate consumption. Next, all participants were informed they were the Responder. The experimenter then brought a covered tray, removed the cover and left the room. For all participants the tray contained two straight-sided 500ml capacity glasses, one holding 62.5ml (12.5%) with “they offer” written below, and the other holding 437.5ml (87.5%) next to “they keep”. Participants had 15 seconds to circle “accept” or “reject” on a piece of paper. Participants who accepted then drank the 62.5ml, and all participants waited one further hour without water.

All statistical tests are two-tailed.

7.3 Results

7.3.1 Thirst manipulation

As predicted, administering hypertonic saline markedly altered objective and subjective measures relating to thirst. Osmolarity at $t_{baseline}$ did not differ between treatments (hypertonic 293 $mOsmL^{-1}$ ± s.d. 4; isotonic 295±7; t(19)=1.27, P=0.22), but at t_{UG} was higher for the hypertonic (310±5) than isotonic group (295±5; t(19)=7.58, $P=3.7x10^{-7}$). Similarly, subjective thirst was no different between treatments at $t_{baseline}$ (hypertonic 2.5±1.9 and isotonic 2.5±1.7; t(19)=0.057, P=0.96), but differed at t_{UG}(7.3±1.6; 3.5±2.0; t(19)=4.68, P<0.0005).

7.3.2 Fairness influences choice

Our data show fairness powerfully influenced responses in the UG despite the use of primary rewards, with 13 of 21 individuals rejecting the unfair offer (binomial test versus no influence of fairness, P<0.0001). Fairness influenced choice in both treatment groups (5 of 10 hypertonic and 8 of 11 isotonic individuals rejected; likelihood ratio test between groups, X^2=1.16, P>0.25).

7.3.3 Fairness is traded-off against subjective self-interest

Next, we asked whether this fairness was traded-off against self-interest: and this was the case for subjective thirst (measured by rating scale). Crucially, subjectively thirstier individuals at t_{UG} were more likely to accept the unfair offered water, indicated by a main effect of choice in a 2 choice (accept, reject) by 2 treatment (isotonic, hypertonic) mixed-effects analysis of variance (ANOVA) with subjective thirst as the dependent variable (main effects of choice $F_{(1,17)}$=9.37, P=0.007; and treatment $F_{(1,17)}$= 46.12, P<0.0005; with no interaction, $F_{(1,17)}$=0.15, P=0.7).

Further, our data revealed that the degree to which hypertonic infusion increased subjective thirst was related to choice. This was shown by the significant interaction of choice (accept, reject) and treatment (isotonic, hypertonic) in an ANOVA with change in subjective thirst as the dependent variable (interaction $F_{(1,17)}$=7.19, P=0.016; main effect of treatment $F_{(1,17)}$= 27.40, P<0.0001; no main effect of choice, $F_{(1,17)}$=3.52, P=0.078). This interaction was driven by the degree to which hypertonic saline increased subjective thirst (Fig. 7.1b). However, our objective measure of thirst (blood osmolarity) was not related to choice, either when used as the dependent variable or as a covariate in the previous ANOVAs (see below). Together, these results suggest the primary driver of the self-interested motivation was subjective, rather than objective, thirst.

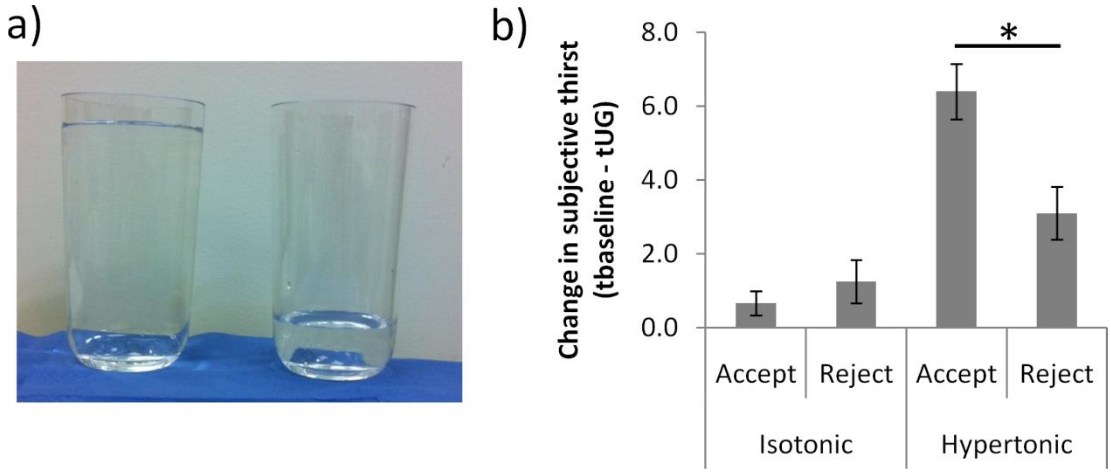

Figure 7.1 Change in subjective thirst and responses to unfairness a) *All participants were assigned the Responder role in an UG and faced this proposed division of water. They received an intravenous infusion of either isotonic or hypertonic saline.* ***b)*** *The change in subjective thirst induced by the saline infusion is calculated from the difference in subjective thirst at baseline and at testing. In the hypertonic group the degree to which the infusion increased subjective thirst was related to choice, such that a greater increase was seen in those who accepted (6.4±1.7) relative to those who rejected (3.1±1.6; independent samples ttest, $t_{(8)}$=3.19, P=0.013).*

7.3.4 Objective measures of thirst (osmolarity):

Osmolarity measures were related to treatment, but not choice. When osmolarity was included as a covariate in the preceding ANOVAs this did not alter the findings. For example, this was evident in a choice (accept, reject) by treatment (isotonic, hypertonic) mixed-effects analysis of variance (ANOVA) with subjective thirst at t_{UG} as the dependent variable and including osmolarity as a covariate (main effect of treatment F(1,16)=6.46, P=0.22; main effect of choice F(1,16)=8.94, P=0.009; no interaction of treatment and choice F(1,16)=0.11, P=0.74; no effect of osmolarity F(1,16)=0.15, P=0.70). When osmolarity at t_{UG} was used as the dependent variable in a 2 choice (accept, reject) by 2 treatment (isotonic, hypertonic) ANOVA there was a main effect of treatment ($F_{(1,17)}$=47.79, P=2.5x10^{-6}), no main effect of choice ($F_{(1,17)}$=1.08, P=0.31) and no interaction ($F_{(1,17)}$=0.26, P=0.62). When change in osmolarity (t_{UG}-$t_{baseline}$) was used as the dependent variable in a 2 choice (accept, reject) by 2 treatment (isotonic, hypertonic) ANOVA there was a main effect of treatment ($F_{(1,17)}$=54.63, P=1.1x10^{-6}), no main effect of choice ($F_{(1,17)}$=1.64, P=0.69) and no interaction ($F_{(1,17)}$=0.83, P=0.38).

7.4 Discussion

Humans' closest relatives, chimpanzees, appear to be rational self-interested maximisers who do not reject unfair offers in the canonical fairness task, the UG (Jensen et al., 2007). Such behaviour suggests the motivation to reject unfair treatment is uniquely human. Indeed, here we show that humans remain interested in fairness even with primary rewards and in a deprived state.

However, whilst our human participants were not solely self-interested, neither were they solely motivated by fairness. Instead they exhibited a trade-off between these motivations. In terms of behavioural economic theory, such a trade-off maps conceptually onto economic models where choice is determined by utility functions containing both self and other regarding components (Fehr and Schmidt, 1999). Further, our data speak to previously discrepant behavioural economic results as to whether individuals' choices are influenced by the size of the stakes relative to their wealth (Cameron, 1999; Camerer, 2003). Thus, here we show that raising the stakes by increasing osmolarity (analogous to reducing wealth) does matter, but only when these stakes impact upon the individual's subjective motivational state. Speculatively, with even more profound hunger, thirst or sexual deprivation than is ethically possible we might infer that the expression of fairness may be even more severely attenuated or abolished. Overall, our data isolate fairness during bargaining as a distinctively human motivation, but crucially one with biologically determined limits.

Chapter 8. Social choice: Testosterone disrupts human cooperation by increasing egocentric behaviour

8.1 Introduction

In the two preceding chapters we have seen that fairness is traded-off against a more self-interested motivation, and that this trade-off is modualted by social context (Chapter 6) and thirst (Chapter 7). Here we examine the trade-off between cooperation and self-interest, and ask how this is biologically modulated by the endocrine system. The evolutionary conservation of cooperation, be it lions hunting in prides (Dugatkin, 1997) or human scientists toiling together in the lab, ultimately derives from benefits accruing to the individual and the wider social group (Axelrod, 1984; Dugatkin, 1997; Gintis et al., 2005). Cooperation in humans is especially striking as it extends to total strangers (Gintis et al., 2005). There is currently much interest in the biological factors that modulate the trade-off between cooperative and more self-motivated behaviour, but in line with influential theory (Gintis et al., 2005) the focus has been on factors increasing a propensity to cooperate. Cooperative behaviours are thought to co-opt neural reward mechanisms (Rilling et al., 2002; Phan et al., 2010) and are causally promoted by the hormone oxytocin (De Dreu et al., 2010). However, whether equivalent influences drive the trade-off in the other direction, to promote more self-orientated behaviour and negatively impact on cooperation, is unknown.

Here, we examined testosterone as a candidate agent. This gonadal hormone, secreted in men and women, modulates a range of behavioural trade-offs, for example the trade-off between parenting and courtship in birds (Wingfield et al., 1990; Ketterson and Nolan, 1994), rodents (Clark and Galef, 1999) and rural

Senegalese men (Alvergne et al., 2009). Socially, higher testosterone correlates with antisocial behaviour in female prisoners (Dabbs and Hargrove, 1997) while a role in fairness-related behaviours is suggested by findings from a bargaining game (Burnham, 2007), although in this bargaining paradigm administration of testosterone has provided mixed results (Zethraeus et al., 2009; Eisenegger et al., 2010). However, parsing the precise role of any neurohumoral agent in social choice is difficult because of an imperative to dissociate social from non-social effects. For example, endogenous testosterone in men and women has been correlated with attention (Fontani et al., 2004) and risk-taking (Sapienza et al., 2009), as well as increasing male financial traders' profit in a risky environment (Coates and Herbert, 2008).

To isolate the impact of testosterone on cooperative and individual decision-making, we exploited a task that assays each of these components independently (Bahrami et al., 2010). In this task individuals must share information, and actively cooperate, to gain a performance benefit in a visual perceptual decision task. Here, we define cooperation as the voluntary acting together of two or more individuals that brings about, or potentially brings about, ends that benefit one, both, or all, which are over and above the benefits arising from individualistic behaviour (Dugatkin, 1997; Brosnan and de Waal, 2002). Collaborative efforts underlie many examples of such cooperative behaviour (Dugatkin, 1997; Brosnan and de Waal, 2002), and indeed they are essentially synonymous with paradigm cases of cooperation such as group hunting (Boesch and Boesch, 1989; Dugatkin, 1997). Importantly, our task enables us to avoid known associations between testosterone and reward-related processing (Coates and Herbert, 2008; Sapienza et al., 2009), for example in the economic bargaining game mentioned above testosterone affected proposers whose choices had uncertain reward-related outcomes but not responders whose choices had certain outcomes (Eisenegger et al., 2010). We predicted testosterone would leave individual decisions unaffected, and would causally disrupt cooperation.

8.2 Methods

We administered testosterone in a randomised, placebo-controlled, double-blind, cross-over design (Fig. 8.1a). Pairs of healthy participants (dyads) comprised our study sample. In our task, both dyad members sat in a room and performed a 2-alternative forced choice task on identical stimuli presented on separate monitors (Fig. 8.1c). On each trial there were two intervals and participants initially decided alone in which interval a target (a higher contrast grating) appeared. Target contrast varied between trials, enabling us to measure the sensitivity of each individual's non-social decision-making by estimating the slope (S_{indiv}) of their psychometric function: where a large slope indicated highly sensitive performance. After these initial individual decisions, participants then saw their partner's choice. In trials where the dyad's initial responses diverged, one participant was randomly selected to announce a cooperative decision reached after free discussion. As was the case for individuals we derived a psychometric function for the dyad, where cooperative success was reflected in the slope ($S_{collective}$). Feedback either followed the individual decision if they initially agreed, or alternatively followed their joint decision.

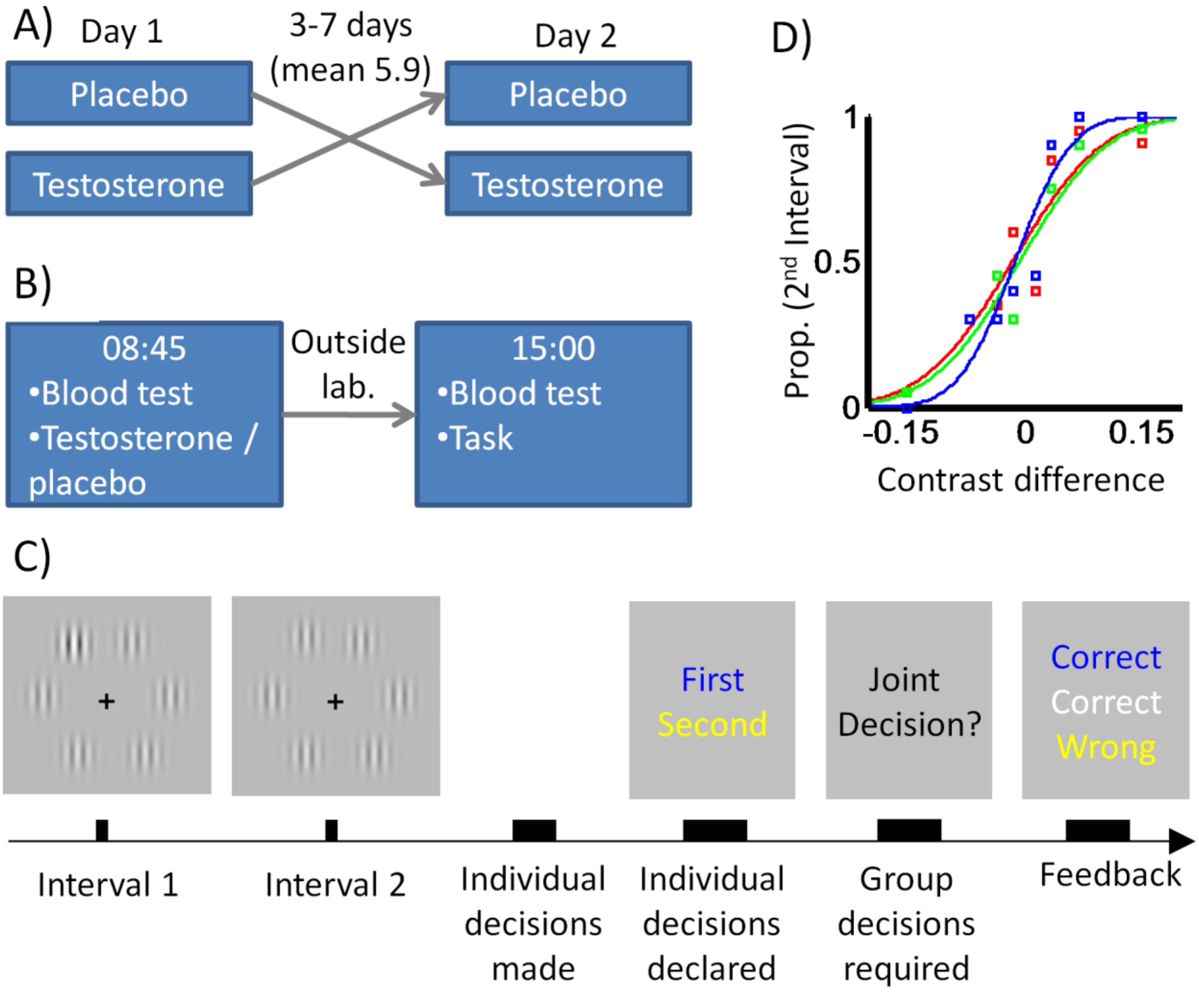

Figure 8.1 Experimental design a) *Pairs of female participants (dyads) attended on two separate days in a blinded, randomised, placebo-controlled cross-over design. Both dyad members received identical treatment order.* ***b)*** *Participants had blood taken before treatment and testing.* ***c)*** *During testing dyad members sat in the same room viewing separate monitors. In a 2-alternative forced choice design gratings were presented at two intervals, one containing a target grating with increased contrast. Each participant initially responded without consultation, providing measures of individual decision-making (S_{indiv}). If they disagreed a joint decision was requested, which provided a measure of cooperative decision-making ($S_{collective}$).* ***d)*** *Example psychometric function for Dyad 1 under placebo. Proportion of trials reported as second interval is plotted against target contrast difference. Highly sensitive observers give steep functions with large slope (S). Here individuals (S_{indiv}) are red and green, and the dyad ($S_{collective}$) blue.*

8.2.1 Participants

34 female participants completed the study (mean age 21.7 years, range 18-30). All participants were part of a dyad and had the same partner throughout. Dyad members did not know each other beforehand. In addition to these 17 dyads, two further dyads were excluded (one participant performed below chance and a second failed to attend both sessions). All were healthy females with normal or corrected to normal visual acuity, and took no medication other than long-standing contraceptives (7 participants took combined oestrogen and progestogen contraception; one took progestogen only contraception). All reported regular menstrual cycles (29.1 ± s.d. 2.2 days, range 29 to 35 days) and were tested between days 1-14 of their cycle. All gave informed consent and the experiment was approved by the local ethics committee.

8.2.2 Experimental procedure

In a randomised, placebo-controlled, double blind, cross-over design 80mg testosterone undecanoate was administered orally (Restandol® testocaps™). The unit of randomisation was the dyad, i.e. both participants received testosterone on one occasion and both participants received placebo on the other occasion. Oral testosterone undecanoate has long been in widespread clinical use and its pharmacokinetics are well known (Geere et al., 1980; Katz et al., 1993). Therefore, to provide a sufficient washout period each dyad attended the laboratory on two separate days 3 to 7 days apart (mean=5.9 days± s.d.=1.1); all had consumed or were given breakfast to aid drug absorption; and the gap between drug administration and the start of behavioural testing was 6-7 hours.

Given that testosterone has a circadian rhythm (highest in the morning), all participants attended the laboratory at the same times on each of the two testing days: 08:45 and 15:00. On each testing day, at 08:45 the pair of participants had a blood sample taken and then received testosterone or placebo. Participants then left the laboratory and returned at 15:00 to undergo venepuncture and then perform our behavioural task.

8.2.3 Hormonal measurement

Total testosterone was measured with a standard, commercially available Roche Modular testosterone assay using electrochemiluminescence immunoassay methods in the University College London Hospitals biochemistry laboratory. Biochemical data was available from 14 of the 17 dyads, with hormonal data from the remaining 3 dyads incomplete due to administrative errors in the biochemistry laboratory.

8.2.4 Behavioural methods

8.2.4.1 Display parameters and Response Mode

During the behavioural testing (Bahrami et al., 2010) dyad members sat in the same testing room and each viewed her own visual display. Display screens were placed on separate tables at right angle to each other. Participants could see each other by turning around. The two displays were connected to the same graphic card via a video amplifier splitter and controlled by the Cogent toolbox (www.vislab.ucl.ac.uk/Cogent/) for MATLAB (Mathworks Inc). Each participant viewed an LCD display at a distance of approximately 60cm (resolution = 800×600 – Dell Ultra Sharp, 22") for which a look-up table linearized the output luminance. Background luminance was 62.5 Cd/m2 in both displays. The displays were connected to a personal computer through an output splitter that sent identical outputs to both of them. Within each session of the experiment, one participant responded with the keyboard and the other with the mouse. Both participants used their right hand.

8.2.4.2 Task, Stimuli and Procedure

A 2-Alternative temporal Forced Choice (2AFC) design was employed with two successive observation intervals. A target stimulus always occurred either in the first or the second interval and participants were instructed to choose the interval most likely to have contained the target. In each interval stimuli comprised 6 vertically oriented Gabor patches (standard deviation of the Gaussian envelope: 0.45 degrees; spatial frequency: 1.5 cycles/degree; contrast: 10%) placed equidistant from each other around an imaginary circle (radius: 8 degrees). The target stimulus was generated by increasing the contrast of one of the six patches. The target location and interval were randomized across the experimental session. The stimulus duration in each interval was 85 ms. Target contrast was determined by adding one of 4 possible values 1.5%, 3.5%, 7.0% or 15% to the 10% contrast of the non-target items.

Each trial was initiated by the participant responding with the keyboard after coordinating with their partner (see Fig. 8.1). A black central fixation cross (width: 0.75 degrees visual angle) appeared on the screen for a variable period, drawn uniformly from the range 500-1000 ms. The two observation intervals were separated by a blank display lasting 1000 ms. The fixation cross turned into a question mark after the second interval to prompt the participants to respond. The question mark stayed on the screen until both participants had responded. Each participant initially responded without consulting the other. The participant who used the keyboard responded by pressing “N” and “M” for the first and second interval, respectively; the participant who used the mouse responded with a left and right click for first and second interval, respectively. Individual decisions were then displayed on the monitor (Fig. 8.1), so both participants were informed about their own and their partner’s choice of the target interval. Colour codes were used to denote keyboard (blue) and mouse (yellow) responses. Vertical locations of the blue and yellow text were randomised to avoid spatial biasing. If the partners disagreed, a joint decision was requested, with the request made in blue if the keyboard participant was to announce the decision and in yellow if the mouse participant was to announce the decision. The keyboard participant announced the joint decision in odd trials; the mouse participant on even trials. Participants were free to verbally discuss their choice for as long as they wanted and to choose any strategy they wished.

The participants received feedback either immediately after they made their decision, in cases where they initially agreed, or after the joint decision was announced, in cases where they initially disagreed. The feedback word was either “CORRECT” or “WRONG”, one for each participant (keyboard: blue; mouse: yellow) and one for the dyad (white), and it remained on the screen until the next trial was initiated by the keyboard (Figure 1, main text). Vertical order of the blue and yellow was randomized and the dyad feedback always appeared in the centre.

On Day 1 participants completed one practice block of 16 trials and then on both days completed 192 trials as 12 blocks of 16 trials (the first three dyads completed fewer trials, with a minimum of 128 trials per day). The experiment was self-paced.

8.2.5 Data Analysis

Psychometric functions were constructed for each participant and for each dyad by plotting the proportion of trials in which the target was seen in the second interval against the contrast difference at the target location (the contrast in the second interval minus the contrast in the first). The psychometric curves were fit to a cumulative Gaussian function, whose parameters were bias, *b*, and variance, σ^2.To estimate these parameters a probit regression model was employed using the *glmfit* function in Matlab (Mathworks Inc). A participant with bias *b* and variance σ^2 would have a psychometric curve, denoted $P(\Delta c)$ where Δc is the contrast difference between the second and first presentations, given by

$$P(\Delta c) = H\left(\frac{\Delta c + b}{\sigma}\right), \qquad \text{Eq 8.1}$$

where *H*(*z*) is the cumulative Normal function,

$$H(z) \equiv \int_{-\infty}^{z} \frac{dt}{(2\pi)^{1/2}} \exp\left[-t^2/2\right]. \qquad \text{Eq 8.2}$$

As usual, the psychometric curve, *P*(Δ*c*), corresponds to the probability of saying that the second interval had the higher contrast. Thus, a positive bias indicates an increased probability of reporting that the second interval had higher contrast (and thus corresponds to a negative mean for the underlying Gaussian distribution).

Given the above definitions for *P*(Δ*c*), we see that variance is related to the maximum slope of the psychometric curve, denote *s*, via

$$s = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}}. \qquad \text{Eq 8.3}$$

A large slope indicates small variance and thus highly sensitive performance. We derive functions for each individual and for the dyad, providing a measure of sensitivity for each as S_{indiv} and $S_{collective}$ respectively. The sensitivity of cooperative decision-making hinges on participants appropriately weighting their own and the other's opinions. For each participant we measure this weighting by the ratio of times they agreed with themselves (egocentric decisions) to agreement with the other's opinion (allocentric decisions). All statistical tests were two-tailed.

8.3 Results

As expected, our hormonal manipulation engendered a large increase in serum testosterone when comparing the time of behavioural testing (mean 9.3 ± s.d. 9.0 nmol/L) to either morning baseline (1.2 ± s.d. 0.5; $t_{(27)}$=4.7, P<0.0001) or placebo (1.1 ± s.d. 0.6; $t_{(19)}$=4.2, P<0.001). Crucially, testosterone had no effect on individual decision making. Individual sensitivity (S_{indiv}) under testosterone was no different to placebo when all 34 participants were considered (S_{indiv}; Placebo 3.11 ± s.d. 1.68; Testosterone 2.99 ± s.d. 1.76; $t_{(33)}$=0.5, P>0.6). This was also the case when considering either the better (S_{max} Plac. 3.80 ± s.d. 1.70; S_{max} Test. 3.69 ± s.d. 1.88; $t_{(16)}$=0.2, P>0.8) or worse performing member of each dyad (S_{min} Plac. 2.41 ± s.d. 1.38; S_{min} Test. 2.28 ± s.d. 1.33; $t_{(16)}$=0.5, P>0.6). The proportion of trials where the dyad's initial decisions diverged also remained unaffected by testosterone (Plac. 0.37 ± s.d. 0.10; Test. 0.39 ± s.d. 0.08; $t_{(16)}$=0.9, P>0.4).

Having shown testosterone did not compromise individual decisions we could then ask if it had a selective impact on cooperation. The logic of effective cooperation is that, if achievable, it benefits the individual more than acting alone (Axelrod, 1984; Dugatkin, 1997; Gintis et al., 2005). We tested this by asking if testosterone affected the performance benefit each individual accrued from cooperation, measured by $S_{collective}$-S_{indiv} (Fig. 8.2). We found testosterone caused a marked decrease in the individual performance benefit arising out of cooperation ($S_{collective}$-S_{indiv} Plac. 1.13 ± s.d. 1.33, Test. 0.54 ± s.d. 1.02; $t_{(33)}$=3.3, P<0.005). Furthermore, testosterone disrupted the benefit of cooperation for the better participant ($S_{collective}$-S_{max} Plac. 0.44 ± s.d. 1.14, Test. -0.17 ± s.d. 0.59; $t_{(16)}$=2.2, P<0.05) as well as for the worse participant in each dyad ($S_{collective}$-S_{min} Plac. 1.82 ± s.d. 1.15, Test. 1.24 ± s.d. 0.86; $t_{(16)}$=2.4, P<0.05). Thus, even from a purely self-interested point of view both dyad members were handicapped when testosterone disrupted the performance benefits from cooperation.

In an evolutionary framework (Brosnan and de Waal, 2002; Gintis et al., 2005) our data implicates testosterone as a proximate, mechanistic modulator of cooperation, and specifically one that reduces the propensity to cooperate. On this basis we would expect it to disrupt cooperation via a consistent bias in cooperative decision-making. To test this prediction we focused on participants' responses as they announced cooperative decisions, where they must appropriately weight each dyad member's opinion. Two considerations might explain how testosterone interferes with this weighting. First, testosterone could lead to a consistent overweighting of the other's opinion, engendering allocentric (other-centred) decision-making, in line with its effect of increasing offers when given in a bargaining game (Eisenegger et al., 2010). Second, it could cause consistent overweighting of participants' own opinions, where such egocentricity parallels its effects on trade-offs in animals, for example to eschew parental responsibilities and increase courtship (Wingfield et al., 1990; Ketterson and Nolan, 1994).

To arbitrate between these competing hypotheses, we computed an egocentric-allocentric (E-A) ratio of the number of trials where the announcer agreed with themselves to the number they agreed with the other. Each hypothesis makes a clear prediction: an allocentricity bias decreases the E-A ratio; and an egocentricity bias increases the E-A ratio. Our data fitted predictions from the second hypothesis, namely that testosterone consistently causes an egocentricity bias (Fig. 8.3). The E-A ratio increased under testosterone (1.61 ± s.d. 1.17) relative to placebo (1.26 ± s.d. 0.83; $t_{(33)}$=2.4, P<0.05). This increased E-A ratio was consistent across both the best and worst performing dyad members, as shown in a 2 decision-maker (S_{min}, S_{max}) by 2 drug (placebo, testosterone) analysis of variance in which there was a main effect of drug ($F_{(1,16)}$=5.8, P<0.05) but not decision maker ($F_{(1,16)}$=0.1, P>0.7) and no interaction ($F_{(1,16)}$=0.6, P>0.4). We also note this egocentricity bias was not accompanied by altered deliberation time for cooperative decisions (Plac. 7.56secs ± s.d. 3.25; Test. 7.44 ± s.d. 2.89; $t_{(33)}$=0.5, P>0.6); which in light of the broader choice literature suggests the effect was not related to decision uncertainty that is usually accompanied by reaction time changes (Slamecka, 1963).

8.3.1 Additional analyses

Neither E-A ratio nor sensitivity measures were related to blood testosterone levels. Biochemical data is available from 14 of the 17 dyads, with hormonal data from the remaining 3 dyads incomplete due to administrative errors in the University College London Hospitals biochemistry laboratory in which they were processed. There were no significant correlations between testosterone (individual or mean dyadic, at baseline or time of testing) and either sensitivity (S_{indiv} or $S_{collective}$) or Egocentric-Allocentric ratio.

Finally, a recent study suggesting participants' beliefs about which drug had been administered might affect choice (Eisenegger et al., 2010). Thus, on each day after completing the behavioural testing participants completed a questionnaire asking if they believed they had received testosterone or placebo. 2 of 34 subjects did not respond. When receiving testosterone 9 of 32 subjects believed they received testosterone, and when receiving placebo 11 of 32 subjects believed they received testosterone. There was no difference in E-A ratio when participants believed they had received placebo (mean=1.58 ± s.d. 1.18, n=44) compared to when they believed they had received testosterone (1.21 ± s.d. 0.62, n=20; independent samples ttest $t_{(62)}$=1.3, P>0.1).

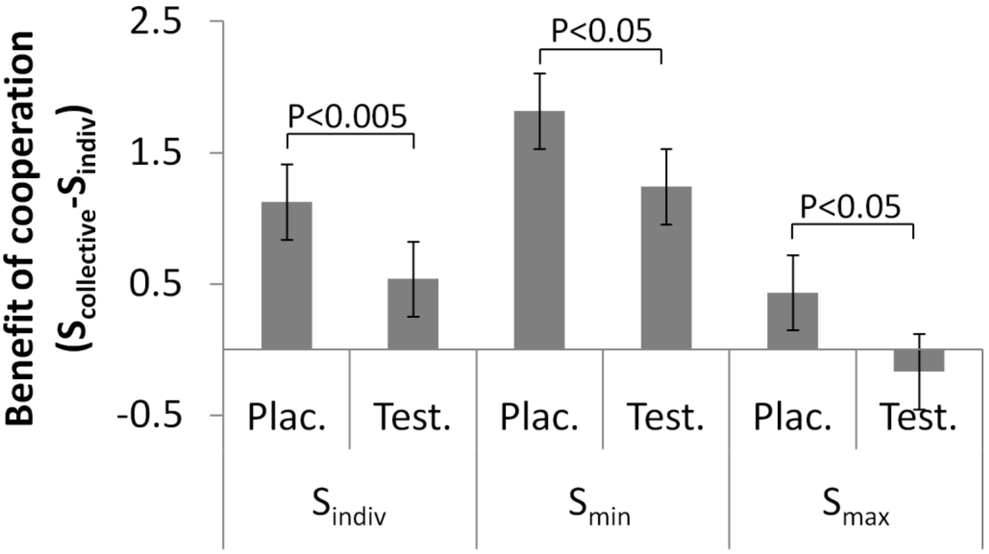

***Figure 8.2 Individuals derive a performance benefit from cooperation.** The dyad's cooperative decisions were more sensitive ($S_{collective}$) than the individuals' decisions alone (S_{indiv}). Our metric for this performance benefit on the vertical axis is the difference between an individual's sensitivity and the cooperative sensitivity achieved by their dyad (Benefit of cooperation = $S_{collective}$ - S_{indiv}). This benefit is attenuated by testosterone when collapsed across all 34 participants (S_{indiv}) and also when only the better (S_{max}) or worse (S_{min}) members of each dyad are included. Error bars indicate s.e.m..*

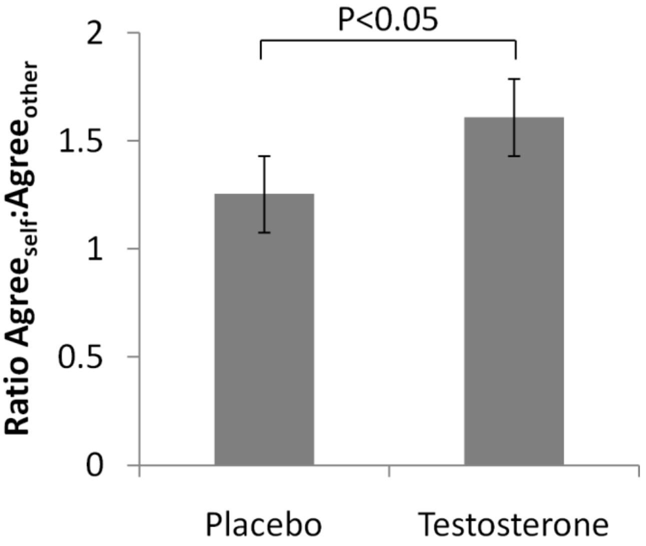

Figure 8.3 Testosterone disrupts cooperation by increasing the egocentricity of decision-making. *Each member of the dyad announced the dyad's joint decision in half the trials where such a cooperative decision was required. The sensitivity of cooperative decision-making hinges on the distribution in weighting attributed to one's own and the other's opinions. For each participant we measured this weighting by the ratio of times they agreed with themselves (egocentric decisions) to agreement with the other's opinion (allocentric decisions). An Egocentric-Allocentric ratio of 1 means that participants weight their own and the other's original judgement equally. On placebo there is trend towards egocentricity bias (one sample, $t_{(33)}$=1.8, P<0.1), an egocentricity bias that becomes marked on testosterone (one sample, $t_{(33)}$=3.0, P=0.005). Error bars indicate s.e.m..*

8.4 Discussion

Our data indicate testosterone selectively and causally disrupts cooperation by increasing egocentricity in decision-making, operationalised as an enhanced weighting of one’s own relative to another’s evidence. We note that such increased self-orientation is a consistent theme across a wide range of testosterone related behaviours, including sexual and reproductive (Wingfield et al., 1990; Ketterson and Nolan, 1994; Clark and Galef, 1999; Alvergne et al., 2009), status-associated (Mazur and Booth, 1998), competitive (Wobber et al., 2010) and aggressive (Wingfield et al., 1990; Dabbs and Hargrove, 1997; Archer, 2006) behaviours. For example high status, where the self is dominant over others, is associated with high testosterone in humans (Mazur and Booth, 1998; Archer, 2006), chimpanzees (Muller and Wrangham, 2004) and other mammals (Sachser and Pröve, 1986). Before competitive interactions between the self and others, anticipatory testosterone rises are seen in chimpanzees but not in more cooperative and egalitarian bonobos (Wobber et al., 2010). Furthermore, in human individuals who worked alone on an analytical reasoning task, higher testosterone correlated with higher scores in subjects told they were competing individually, but lower performance in subjects told their score would subsequently be pooled with someone else’s (Mehta et al., 2009). A natural consequence of testosterone causing increased self-orientation would be to down-rate perceptions of others or empathy, which are seen by using facial trustworthiness ratings (Bos et al., 2010) and empathising related to photographs of eyes respectively (van Honk et al., 2011). However, whilst our interpretation accords well with the wider literature, here both dyad members received testosterone on the same day and thus future work could usefully examine the possibility that testosterone may disrupt cooperation by reducing an individual’s ability to signal their confidence. Interestingly given a recent study suggesting an effect of participants’ beliefs about which drug they received (Eisenegger et al., 2010), which we did not find here, future work might also directly manipulate such an effect. More broadly, the idea that testosterone increases self-belief is now testable within a new framework by assaying meta-cognition (Fleming et al., 2010).

The success of social animals, particularly humans, depends on how well individuals manage a critical day-to-day trade-off between cooperative and more self-motivated behaviours. Biological mechanisms controlling this trade-off must tune behaviour to the social environment. Whilst a previous focus has been on factors promoting cooperation (Rilling et al., 2002; De Dreu et al., 2010; Phan et al., 2010), it is clear that without opposing factors, such as that we show for testosterone, this form of control mechanism would be lopsided. Indeed, diminished cooperation should not necessarily be seen in a negative light as it is likely to be critical in preventing exploitation and, more speculatively, might promote artistic and creative endeavours that are often characterised by a degree of iconoclasm.

What our data shows is that the humoral agent testosterone, known to be dynamically tuned by ecological contingencies, particularly with respect to conspecifics (Mazur and Booth, 1998; Wobber et al., 2010), serves a crucial role in modulating this delicate trade-off between cooperation and a more egocentric disposition.

Chapter 9. General discussion

9.1 Overview

In this thesis I aimed to use a biologically-based perspective to understand how people make choices. Specifically, I investigated two paradigmatic influences on individual choice, namely risk and the possibility of loss (Chapters 4 and 5); and two paradigmatic influences on social choice, namely fairness and cooperation (Chapters 6, 7 and 8). I used concepts from the quantitative social sciences, behavioural model comparison, functional Magnetic Resonance Imaging (fMRI) and employed causal manipulations of hormones and physiological state.

In the following sections I will discuss the contributions, limitations and future work arising from the studies included in this thesis: first from my examination of individual choice in Chapters 4 and 5; and then from my investigations of social choices in Chapters 6, 7 and 8.

9.2 Individual choice

In both Chapters 4 and 5 I described greater gambling for gains than losses, a finding inconsistent with a tied relationship between risk and valence that specifies a valence-induced bias in the opposite direction. Instead, we found behavioural and neural dissociations between the effects of risk and valence, consistent with an hypothesis that risk and valence exert independent influences on choice. I show that a simple manipulation of task structure dissociated the impacts of risk and valence, by selectively reversing the effect of valence while leaving a risk-induced bias unaffected; that individual preferences for each were also independent; and further that risk and valence were encoded by distinct neural systems. These dissociations are not predicted by existing behavioural economic theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), but can be accommodated in a biologically-based account of choice in which risk and loss bias approach towards economic stimuli. This mechanistic explanation provided testable hypotheses, which we confirmed using reaction time data in Chapter 5.

One set of limitations of this work arises from our operationalisation of risk: first, we limit choices to those involving risky prospects where the decision-maker knows all probabilities; and second we manipulated risk as the variance in outcome distributions. Firstly, risk in which all probabilities are known is an important case, and this is also the type of risk addressed by EUT and Prospect Theory. However, these “known unknowns” reflect only one type of risk, and it is unclear if our findings would extend to the “unknown unknowns”, or ambiguity, that have also been shown to impact on behaviour (Ellsberg, 1961) and that are reflected neurally (Bach et al., 2009). It is ill understood how ambiguity-related preferences might interact with valence (i.e. gain and loss outcomes), and this is an interesting future line of enquiry. Secondly, given our operationalisation of risk as variance, an aspect of risk that we do not address is other summary statistics describing outcome distributions. For example, skewness has been shown to impact on choice and to be reflected neurally (Symmonds et al., 2011; Wu et al., 2011). Again, future work could usefully examine how skewness preferences interact with valence. Further, our task does not directly address a broader meaning of risky behaviours often used by clinicians or lay people, where these are defined as those that may lead to harm such as in mountain climbing (Schonberg et al., 2011). This carries implications for the ecological validity of our findings, as it is known for example that individual differences in risky choice appear to be domain specific across different types of risk (Slovic, 1964; Weber et al., 2002).

Other limitations also arise from our operationalisation of valence. We use financial rewards and punishments. However, whilst this is an extremely well learnt conditioned reinforcer (Dayan and Seymour, 2008), it would be interesting to ask if our findings extend to cases outside the financial domain such as primary rewards (O'Doherty, 2004), pain (Seymour et al., 2004) or a mixture of pain and money (Talmi et al., 2009).

A further limitation to the generalisability of our findings derives from the way we gave our subjects the explicit probabilities. As is typical in economic studies, we used choices in which the key decision variables are explicitly stated. However, many learning theorists study choices in which options are evaluated on the basis of experience (Sutton and Barto, 1998, Stephens and Krebs, 1986; Mackintosh and Dickinson, 1979). Growing behavioural evidence suggests that valuation based on these different classes of information involve separable mechanisms (Hertwig et al., 2004; Jessup et al., 2008; Ungemach et al., 2009; Wu et al., 2009). Indeed a recent study has shown differential sensitivity to learned and described value and risk in brain regions commonly associated with reward processing (Fitzgerald et al., 2010). Again, how this aspect of risk evaluation or learning might interact with valence is a potentially fruitful avenue for future work.

The stable and independent inter-individual differences for risk and valence that we demonstrate also raises further questions. In particular, stability in the aversive impact of loss on choice over time has not to our knowledge been previously demonstrated, and is interesting in light of work suggesting framing effects may be genetically mediated (Roiser et al., 2009). However, it would be interesting to ask if this stability with valence lasted over even longer periods of time, and to ask if there are multiple domains within this valence sensitivity as is the case for risk (Slovic, 1964; Weber et al., 2002). This would speak to the external validity of our valence-related measures.

Neurally, our data revealed activity encoding the degree of stimulus risk in parietal cortex, which concurs with single unit and fMRI data showing enhanced activity during risky decision-making (Platt and Glimcher, 1999; Huettel et al., 2005; Mohr et al., 2010). This provides evidence in support of “summary statistic models”. However, it is important to ask how this risk evaluation is related to action selection, although we note that in our study this same region shows activity for choice and that this correlates with risk preference. Our neural data also link approach/avoidance mechanisms (Kim and Jung, 2006; Seymour et al., 2007; Dayan and Seymour, 2009) to both risk and valence. However, the degree of similarity between such risk and valence related systems is a matter for further study – and indeed more generally it is poorly understood if there exist one or many such mechanisms (Rangel et al., 2008).

Within an account of choice proceeding from option evaluation through to action selection (Corrado et al., 2009), we suggest that the risk and valence of an economic stimulus are processed by separable neural systems, and influence action-selection partly through reflexive systems that bias approach responses. This process model yields testable predictions, for example concerning the approach/avoidance mechanism using reaction times. However, critical in this process model are issues of temporal order, which are very difficult to address with fMRI. It is also possible that we could ask mechanistic questions of our RT data using diffusion modelling (Ratcliff, 2000), although this typically uses tasks in which the RT is much shorter (in the order of 1 to 1.5 seconds, not the 3.5 seconds we see in our free response experiments in Chapter 5). Further, our account implies causality that we could perhaps most usefully test using neuropsychological techniques, which could also ask which regions we identify are necessary as well as sufficient.

9.3 Social choice

With respect to the social choices examined in Chapters 6, 7 and 8, I examined the biological systems enabling social behaviour to respond flexibly to environmental contingencies. In the neural investigation of fairness presented in Chapter 6, the principal aim was a behavioural and neural characterisation of objective and contextual aspects of fairness. We defined the contextual component of fairness as a shift in choices in response to otherwise identical offers. Our finding of a marked context-dependence provides a perspective on fairness as a relative rather than absolute quantity, echoing findings in relation to other high-level quantities such as valuation (Ariely et al., 2006; Seymour and McClure, 2008; Vlaev et al., 2009). However, our neural data also highlight a fundamental role for objective social inequality that accords with effects seen in the UG across diverse cultures (Henrich, 2004) and in human infants (Fehr et al., 2008). Our data highlights how these objective and contextual aspects interact to construct a fairness motivation with sufficient flexibility to enable appropriate responses to the social environment. Our neuroimaging data strongly support inequality aversion models: first, we find a main effect of inequality in posterior insula; second, between subjects the envy parameter correlates with activity in the precuneus, left TPJ and frontopolar cortex; third, inequality modulated posterior and mid-insula activity more strongly when inequality is psychologically more aversive. These findings extend current inequality aversion models, demonstrating the behavioural and neural flexibility to avoid knee-jerk aversion to inequality.

One limitation of our design is that although we assume choice is the outcome of processes whose neural implementation may involve social computations such as prediction errors (Behrens et al., 2008; Hampton et al., 2008), we did not employ a non-social control. Another potential issue with the social content of the task was that although we led participants to believe they were playing with real others in the game, they actually responded to a predetermined set of offers.

Neurally, perhaps our most important findings relate to the role of insula cortex in fairness motivation. We propose functional segregation in this extensive (over 5cm long) and cytoarchitectonically diverse cortical region (Flynn, 1999; Varnavas and Grand, 1999), with posterior insula negatively correlating with inequality and anterior insula positively correlating with inequality. This would explain previously divergent findings in insula (Sanfey et al., 2003; Hsu et al., 2008). Since our study, a causal study in non-human primates using stimulation in insula has shown results highly consistent with this hypothesis(Caruana et al., 2011), where stimulation of more posterior regions led to affiliative behaviours, whilst stimulation more anteriorly led to more disgust-related behaviours.

In Chapter 7, we show that humans remain interested in fairness even with primary rewards and in a deprived state. In contrast, humans' closest relatives, chimpanzees, appear to be rational self-interested maximisers who do not reject unfair offers in the canonical fairness task, the UG (Jensen et al., 2007). Such behaviour suggests the motivation to reject unfair treatment may be uniquely human. However, whilst our human participants were not solely self-interested, neither were they solely motivated by fairness, and instead they exhibited a trade-off between these motivations. Importantly, our data speak to previously discrepant behavioural economic results as to whether individuals' choices are influenced by the size of the stakes relative to their wealth (Cameron, 1999; Camerer, 2003): here we show that raising the stakes by increasing osmolarity (analogous to reducing wealth) does matter, but only when these stakes impact upon the individual's subjective motivational state.

Finally, in Chapter 8 we addressed the hormonal regulation of cooperation. Our data indicate testosterone selectively and causally disrupts cooperation by increasing egocentricity in decision-making. This is important as whilst previous work has focussed on factors promoting cooperation (Rilling et al., 2002; De Dreu et al., 2010; Phan et al., 2010), it is clear that without opposing factors, such as that we show for testosterone, this form of control mechanism would be lopsided. Indeed, diminished cooperation should not necessarily be seen in a negative light as it is likely to be critical in preventing exploitation. However, whilst our interpretation accords well with the wider literature, one limitation is that both dyad members received testosterone on the same day. Future work could thus usefully examine the possibility that testosterone may disrupt cooperation by reducing an individual's ability to signal their confidence. One other possibility is that testosterone increases self-belief more generally and not just in social interactions, which could usefully be tested using a framework that assays meta-cognition (Fleming et al., 2010). Further, it will be important in future to extend this work such that physiological rather than above physiological doses are examined, and that men as well as women are investigated.

Together, our three social studies point to the importance of control mechanisms to manage a critical day-to-day trade-off between social and more self-interested motivations. Biological mechanisms controlling such trade-offs must tune behaviour to the social environment, and will be critical to the success of social animals such as humans.

9.4 Conclusions

This thesis began with two observations regarding new inter-disciplinary approaches that combine biological and economic perspectives on choice. Firstly, that these new approaches could permit the reintroduction of an earlier richness and complexity into models of human behaviour, but within a mathematically specifiable and empirically grounded framework. Second, that these inter-disciplinary approaches may provide better descriptive models of choice. It is my hope that the studies described here make modest contributions to both these aims. Certainly, they raised questions that I greatly enjoyed trying to answer.

References

d' Acremont M, Bossaerts P (2008) Neurobiological studies of risk assessment: A comparison of expected utility and mean-variance approaches. Cognitive, Affective, & Behavioral Neuroscience 8:363–374.

Akerlof GA (1979) The case against conservative macroeconomics: an inaugural lecture. Economica 46:219–237.

Akerlof GA (1982) Labor contracts as partial gift exchange. The Quarterly Journal of Economics 97:543.

Akerlof GA, Shiller RJ (2009) Animal spirits. Princeton, NJ: Princeton University Press.

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.

Alvergne A, Faurie C, Raymond M (2009) Variation in testosterone levels and male reproductive effort: Insight from a polygynous human population. Hormones and Behavior 56:491–497.

Andersen S, Harrison GW, Lau MI, Rutström EE (2008) Lost in State Space: are preferences stable? International Economic Review 49:1091–1112.

Andersson JL, Hutton C, Ashburner J, Turner R, Friston K (2001) Modeling geometric deformations in EPI time series. Neuroimage 13:903–919.

Archer J (2006) Testosterone and human aggression: an evaluation of the challenge hypothesis. Neuroscience & Biobehavioral Reviews 30:319–345.

Ariely D, Loewenstein G, Prelec D (2006) Tom Sawyer and the construction of value. Journal of Economic Behavior & Organization 60:1–10.

Axelrod R (1984) The Evolution of Cooperation. New York: Basic Books.

Bach DR, Seymour BJ, Dolan RJ (2009) Neural Activity Associated with the Passive Prediction of Ambiguity and Risk for Aversive Events. The Journal of Neuroscience 29:1648–1656.

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD (2010) Optimally Interacting Minds. Science 329:1081.

Barnard CJ, Brown CAJ (1985) Risk-sensitive foraging in common shrews (Sorex araneus L.). Behavioral Ecology and Sociobiology 16:161–164.

Battalio RC, Kagel JH, Jiranyakul K (1990) Testing between alternative models of choice under uncertainty: Some initial results. Journal of Risk and Uncertainty 3:25–50.

Behrens TE., Hunt LT, Woolrich MW, Rushworth MF. (2008) Associative learning of social value. Nature 456:245.

Blanchard DC, Blanchard RJ (1988) Ethoexperimental approaches to the biology of emotion. Annual Review of Psychology 39:43–68.

Boesch C, Boesch H (1989) Hunting behavior of wild chimpanzees in the Tai National Park. American Journal of Physical Anthropology 78:547–573.

Bolton GE, Ockenfels A (2000) ERC: A theory of equity, reciprocity, and competition. American Economic Review:166–193.

Bos PA, Terburg D, van Honk J (2010) Testosterone decreases trust in socially naïve humans. Proceedings of the National Academy of Sciences 107:9991.

Bossaerts P (2010) Risk and risk prediction error signals in anterior insula. Brain Struct Funct 214:645–653.

Brosnan SF, De Waal FB. (2003) Monkeys reject unequal pay. Nature 425:297–299.

Brosnan SF, de Waal FBM (2002) A proximate perspective on reciprocal altruism. Hum Nat 13:129–152.

Burnham TC (2007) High-testosterone men reject low ultimatum game offers. Proceedings of the Royal Society B 274:2327.

Bushong B, King LM, Camerer CF, Rangel A (2010) Pavlovian Processes in Consumer Choice: The Physical Presence of a Good Increases Willingness-to-Pay. The American Economic Review 100:1556–1571.

Calder AJ, Lawrence AD, Young AW (2001) Neuropsychology of fear and loathing. Nature Reviews Neuroscience 2:352–363.

Camerer C, Loewenstein G, Prelec D (2005) Neuroeconomics: How neuroscience can inform economics. Journal of economic Literature 43:9–64.

Camerer CF (1989) An experimental test of several generalized utility theories. J Risk Uncertainty 2:61–104.

Camerer CF (1998) Prospect theory in the wild: Evidence from the field. Social Science Working Paper-California Institute of Technology Division of the Humanities and Social Sciences.

Camerer CF (2003) Behavioral game theory: Experiments in strategic interaction. Princeton University Press Princeton, NJ.

Camerer CF, Loewenstein G (2004) Behavioral economics: Past, present, future. In: Advances in behavioral economics (Camerer CF, Loewenstein G, Rabin M, eds), pp.3–51. Princeton: Princeton University Press.

Cameron LA (1999) Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. Economic Inquiry 37:47–59.

Cardinal R, Howes N (2005) Effects of lesions of the nucleus accumbens core on choice between small certain rewards and large uncertain rewards in rats. Bmc Neuroscience 6:37.

Caruana F, Jezzini A, Sbriscia-Fioretti B, Rizzolatti G, Gallese V (2011) Emotional and Social Behaviors Elicited by Electrical Stimulation of the Insula in the Macaque Monkey. Current Biology 21:195–199.

Charness G, Rabin M (2002) Understanding Social Preferences with Simple Tests*. Quarterly journal of Economics 117:817–869.

Chen MK, Lakshminarayanan V, Santos LR (2006) How basic are behavioral biases? Evidence from capuchin monkey trading behavior. Journal of Political Economy 114:517–537.

Christie R, Geis F (1970) Studies in Machiavellianism. New York: Academic Press.

Christopoulos GI, Tobler PN, Bossaerts P, Dolan RJ, Schultz W (2009) Neural Correlates of Value, Risk, and Risk Aversion Contributing to Decision Making under Risk. J Neurosci 29:12574–12583.

Clark MM, Galef BG (1999) A Testosterone-Mediated Trade-Off Between Parental and Sexual Effort in Male Mongolian Gerbils (Meriones unguiculatus),. Journal of Comparative Psychology 113:388–395.

Coates JM, Herbert J (2008) Endogenous steroids and financial risk taking on a London trading floor. Proceedings of the National Academy of Sciences 105:6167.

Coombs CH, Huang L (1970) Tests of a portfolio theory of risk preference. Journal of Experimental Psychology 85:23–29.

Coombs CH, Pruitt DG (1960) Components of risk in decision making: Probability and variance preferences. Journal of Experimental Psychology 60:265.

Corrado GS, Doya K (2007) Understanding Neural Coding through the Model-Based Analysis of Decision Making. The Journal of Neuroscience 27:8178–8180.

Corrado GS, Sugrue LP, Brown JR, Newsome WT (2008) The trouble with choice: studying decision variables in the brain. In: Neuroeconomics: decision making and the brain (Glimcher PW, Camerer CF, Fehr E, Poldrack RA, eds). Academic Press.

Corrado GS, Sugrue LP, Brown JR, Newsome WT (2009) The trouble with choice: studying decision variables in the brain. In: Neuroeconomics: Decision making and the brain, pp.463.

Craig AD (2002) How do you feel? Interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience 3:655–666.

Craig ADB (2009) How do you feel--now? The anterior insula and human awareness. Nat Rev Neurosci 10:59–70.

Critchley HD, Mathias CJ, Dolan RJ (2001) Neural Activity in the Human Brain Relating to Uncertainty and Arousal during Anticipation. Neuron 29:537–545.

Critchley HD, Wiens S, Rotshtein P, Ohman A, Dolan RJ (2004) Neural systems supporting interoceptive awareness. Nat Neurosci 7:189–195.

Croy MI, Hughes RN (1991) Effects of food supply, hunger, danger and competition on choice of foraging location by the fifteen-spined stickleback, Spinachia spinachia L. Animal Behaviour 42:131–139.

Dabbs JM, Hargrove MF (1997) Age, testosterone, and behavior among female prison inmates. Psychosom Med 59:477–480.

Damasio AR, Everitt BJ, Bishop D (1996) The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex [and Discussion]. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 351:1413–1420.

Dayan P (2008) The role of value systems in decision making. Better than conscious:51–70.

Dayan P, Huys QJM (2009) Serotonin in affective control. Annu Rev Neurosci 32:95–126.

Dayan P, Seymour BJ (2008) Values and actions in aversion. In: Neuroeconomics: Decision making and the brain. London: Academic Press.

Denton D, Shade R, Zamarippa F, Egan G, Blair-West J, McKinley M, Fox P (1999) Correlation of regional cerebral blood flow and change of plasma sodium concentration during genesis and satiation of thirst. Proc Natl Acad Sci U S A 96:2532–2537.

Dickhaut J, McCabe K, Nagode JC, Rustichini A, Smith K, Pardo JV (2003) The impact of the certainty context on the process of choice. Proceedings of the National Academy of Sciences 100:3536–3541.

Dreher J-C, Kohn P, Berman KF (2006) Neural coding of distinct statistical properties of reward information in humans. Cereb Cortex 16:561–573.

De Dreu CKW, Greer LL, Handgraaf MJJ, Shalvi S, Van Kleef GA, Baas M, Ten Velden FS, Van Dijk E, Feith SWW (2010) The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans. Science 328:1408–1411.

Driver-Dunckley E, Samanta J, Stacy M (2003) Pathological gambling associated with dopamine agonist therapy in Parkinson’s disease. Neurology 61:422–423.

Dugatkin LA (1997) Cooperation among animals: an evolutionary perspective. Oxford University Press.

Duong TQ, Kim D-S, Uğurbil K, Kim S-G (2001) Localized cerebral blood flow response at submillimeter columnar resolution. Proceedings of the National Academy of Sciences 98:10904–10909.

Eisenegger C, Naef M, Snozzi R, Heinrichs M, Fehr E (2010) Prejudice and truth about the effect of testosterone on human bargaining behaviour. Nature 463:356–359.

Ellsberg D (1961) Risk, Ambiguity, and the Savage Axioms. The Quarterly Journal of Economics 75:643–669.

Engelmann JB, Tamir D (2009) Individual differences in risk preference predict neural responses during financial decision-making. Brain Research 1290:28–51.

Ert E, Erev I (2008) The rejection of attractive gambles, loss aversion, and the lemon avoidance heuristic. Journal of Economic Psychology 29:715–723.

Farrer C, Franck N, Georgieff N, Frith CD, Decety J, Jeannerod M (2003) Modulating the experience of agency: a positron emission tomography study. Neuroimage 18:324–333.

Fehr E, Bernhard H, Rockenbach B (2008) Egalitarianism in young children. Nature 454:1079–1083.

Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci (Regul Ed) 11:419–427.

Fehr E, Schmidt KM (1999) A Theory Of Fairness, Competition, and Cooperation*. Quarterly Journal of Economics 114:817–868.

Fiorillo CD, Tobler PN, Schultz W (2003) Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. Science 299:1898–1902.

Fitzgerald THB, Seymour BJ, Bach DR, Dolan RJ (2010) Differentiable neural substrates for learned and described value and risk. Curr Biol 20:1823–1829.

Flandin G, Friston K (2008) Statistical parametric mapping (SPM). Scholarpedia 3:6232.

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating Introspective Accuracy to Individual Differences in Brain Structure. Science 329:1541–1543.

Fliessbach K, Weber B, Trautner P, Dohmen T, Sunde U, Elger CE, Falk A (2007) Social comparison affects reward-related brain activity in the human ventral striatum. Science 318:1305.

Flynn FG (1999) Anatomy of the insula functional and clinical correlates. Aphasiology 13:55–78.

Fontani G, Lodi L, Felici A, Corradeschi F, Lupo C (2004) Attentional, emotional and hormonal data in subjects of different ages. Eur J Appl Physiol 92:452–461.

Frackowiak RSJ, Ashburner JT, Penny WD, Zeki S, Friston KJ, Frith CD, Dolan RJ, Price CJ eds. (2004) Human Brain Function, Second Edition, 2nd ed. Academic Press.

Friston KJ (2004) Experimental design and statistical parametric mapping. In: Human brain function (Frackowiak RSJ, ed). Academic Press.

Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RS. (1995a) Spatial registration and normalization of images. Human brain mapping 3:165–189.

Friston KJ, Fletcher P, Josephs O, Holmes A, Rugg MD, Turner R (1998) Event-Related fMRI: Characterizing Differential Responses. NeuroImage 7:30–40.

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RS. (1995b) Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2:189–210.

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RS., others (1995c) Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2:189–210.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC (1993) Assessing the significance of focal activations using their spatial extent. Human Brain Mapping 1:210–220.

Frith CD, Frith U (2006) How we predict what other people are going to do. Brain Res 1079:36–46.

Geere G, Jones J, Atherden SM, Grant DB (1980) Plasma androgens after a single oral dose of testosterone undecanoate. Archives of Disease in Childhood 55:218–220.

Gilbert SJ, Burgess PW (2008) Executive function. Curr Biol 18:R110–R114.

Gintis H, Bowles S, Boyd RT, Fehr E (2005) Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life. The MIT Press.

Glaser D, Friston KJ (2004) Variance components. In: Human brain function (Frackowiak RSJ, ed). Academic Press.

Glimcher PW (2003) Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics, 1st ed. The MIT Press.

Glimcher PW, Rustichini A (2004) Neuroeconomics: The Consilience of Brain and Decision. Science 306:447–452.

Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV (2007) Two takes on the social brain: a comparison of theory of mind tasks. J Cogn Neurosci 19:1803–1814.

Grèzes J, Frith CD, Passingham RE (2004) Inferring false beliefs from the actions of oneself and others: an fMRI study. Neuroimage 21:744–750.

Güroğlu B, van den Bos W, Rombouts SARB, Crone EA (2010) Unfair? It depends: neural correlates of fairness in social context. Soc Cogn Affect Neurosci 5:414–423.

Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization 3:367–388.

Guitart-Masip M, Fuentemilla L, Bach DR, Huys QJM, Dayan P, Dolan RJ, Duzel E (2011) Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. J Neurosci 31:7867–7875.

Guitart-Masip M, Talmi D, Dolan R (2010) Conditioned associations and economic decision biases. Neuroimage 53:206–214.

Halko M-L, Hlushchuk Y, Hari R, Schürmann M (2009) Competing with peers: mentalizing-related brain activity reflects what is at stake. Neuroimage 46:542–548.

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences 105:6741.

Handwerker DA, Ollinger JM, D'Esposito M (2004) Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage 21:1639–1651.

Harrison GW, Rutström EE (2008) Risk aversion in the laboratory. In: Risk aversion in experiments (Research in Experimental Economics, Volume 12), pp.41–196.

Heeger DJ, Ress D (2002) What does fMRI tell us about neuronal activity? Nat Rev Neurosci 3:142–151.

Henrich JP (2004) Foundations of human sociality: economic experiments and ethnographic evidence from fifteen small-scale societies. Oxford University Press.

Hertwig R, Barron G, Weber EU, Erev I (2004) Decisions from Experience and the Effect of Rare Events in Risky Choice. Psychological Science 15:534–539.

Hey J (2003) Intermediate Microeconomics. McGraw-Hill Higher Education.

Ho TH, Lim N, Camerer CF (2006) Modeling the psychology of consumer and firm behavior with behavioral economics. Journal of Marketing Research 43:307–331.

Homans GC (1961) Social behavior: its elementary forms. Harcourt, Brace & World.

van Honk J, Schutter DJ, Bos PA, Kruijt A-W, Lentjes EG, Baron-Cohen S (2011) Testosterone administration impairs cognitive empathy in women depending on second-to-fourth digit ratio. Proceedings of the National Academy of Sciences.

Hsu M, Anen C, Quartz SR (2008) The right and the good: distributive justice and neural encoding of equity and efficiency. Science 320:1092–1095.

Hubbard EM, Piazza M, Pinel P, Dehaene S (2005) Interactions between number and space in parietal cortex. Nature Reviews Neuroscience 6:435–448.

Huettel SA, Song AW, McCarthy G (2005) Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. J Neurosci 25:3304–3311.

Huettel SA, Song AW, McCarthy G (2008) Functional magnetic resonance imaging, Second. Sunderland, MA: Sinauer Associates.

Jensen K, Call J, Tomasello M (2007) Chimpanzees are rational maximizers in an ultimatum game. Science 318:107–109.

Jessup RK, Bishara AJ, Busemeyer JR (2008) Feedback Produces Divergence From Prospect Theory in Descriptive Choice. Psychological Science 19:1015–1022.

Jezzard P, Matthews PM, Smith SM (2003) Functional MRI: An Introduction to Methods, 1st ed. Oxford University Press, USA.

Jones CL, Ward J, Critchley HD (2010) The neuropsychological impact of insular cortex lesions. J Neurol Neurosurg Psychiatr 81:611–618.

Kacelnik A, Bateson M (1996) Risky theories—the effects of variance on foraging decisions. American Zoologist 36:402.

Kagel JH, Roth AE (1995) The handbook of experimental economics. Princeton, NJ.

Kahneman D, Knetsch JL, Thaler R (1986) Fairness as a Constraint on Profit Seeking: Entitlements in the Market. The American Economic Review 76:728–741.

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

Katz M, De Sanctis V, Vullo C, Wonke B, McGarrigle HH, Bagni B (1993) Pharmacokinetics of sex steroids in patients with beta thalassaemia major. British Medical Journal 46:660.

Ketterson ED, Nolan V (1994) Male Parental Behavior in Birds. Annual Review of Ecology and Systematics 25:601–628.

Kim JJ, Jung MW (2006) Neural circuits and mechanisms involved in Pavlovian fear conditioning: a critical review. Neuroscience & Biobehavioral Reviews 30:188–202.

Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E (2005) Oxytocin increases trust in humans. Nature 435:673–676.

Kőszegi B, Rabin M (2006) A Model of Reference-Dependent Preferences. The Quarterly Journal of Economics 121:1133–1165.

Krajbich I, Adolphs R, Tranel D, Denburg NL, Camerer CF (2009) Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. J Neurosci 29:2188–2192.

Kreps (1990) Course Microeconomic Theory. Pearson Higher Education.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540.

Kuhnen CM, Knutson B (2005) The neural basis of financial risk taking. Neuron 47:763–770.

Van Lange PA, Otten W, De Bruin EM, Joireman JA (1997) Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. J Pers Soc Psychol 73:733–746.

Laury SK, Holt CA (2005) Further reflections on prospect theory. Andrew Young School of Policy Studies Research Paper Series No 06-11.

Leland DS, Paulus MP (2005) Increased risk-taking decision-making but not altered response to punishment in stimulant-using young adults. Drug Alcohol Depend 78:83–90.

Levy JS (2003) Applications of prospect theory to political science. Synthese 135:215–241.

Lin P, Hasson U, Jovicich J, Robinson S (2011) A neuronal basis for task-negative responses in the human brain. Cereb Cortex 21:821–830.

Logothetis NK (2008) What we can do and what we cannot do with fMRI. Nature 453:869–878.

Mackintosh NJ (1983) Conditioning and associative learning. New York: Clarendon Press.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19:1233–1239.

Markowitz H (1952) Portofolio selection. The Journal of Finance 7:77–91.

De Martino B, Kumaran D, Seymour BJ, Dolan RJ (2006) Frames, biases, and rational decision-making in the human brain. Science 313:684.

Mazur A, Booth A (1998) Testosterone and dominance in men. Behavioral and Brain Sciences 21:353–363.

Mehta PH, Wuehrmann EV, Josephs RA (2009) When are low testosterone levels advantageous? The moderating role of individual versus intergroup competition. Hormones and Behavior 56:158–162.

Menon RS, Ogawa S, Hu X, Strupp JP, Anderson P, Uğurbil K (1995) BOLD Based Functional MRI at 4 Tesla Includes a Capillary Bed Contribution: Echo-Planar Imaging Correlates with Previous Optical Imaging Using Intrinsic Signals. Magnetic Resonance in Medicine 33:453–459.

Messick DM, McClintock CG (1968) Motivational bases of choice in experimental games. Journal of Experimental Social Psychology 4:1–25.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Mohr PNC, Biele G, Heekeren HR (2010) Neural Processing of Risk. J Neurosci 30:6613–6619.

Muller MN, Wrangham RW (2004) Dominance, aggression and testosterone in wild chimpanzees: a test of the "challenge hypothesis." Animal Behaviour 67:113–123.

Nachev P, Kennard C, Husain M (2008) Functional role of the supplementary and pre-supplementary motor areas. Nat Rev Neurosci 9:856–869.

von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton: Princeton University Press.

O’Doherty JP (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Current Opinion in Neurobiology 14:769–776.

O’Doherty JP, Hampton A, Kim H (2007) Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci 1104:35–53.

Ogawa S, Lee TM, Kay AR, Tank DW (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci U S A 87:9868–9872.

Ogawa S, Tank DW, Menon R, Ellermann JM, Kim SG, Merkle H, Ugurbil K (1992) Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proc Natl Acad Sci USA 89:5951–5955.

Van Overwalle F, Baetens K (2009) Understanding others’ actions and goals by mirror and mentalizing systems: A meta-analysis. NeuroImage 48:564–584.

Paulus MP, Hozack N, Zauscher B, McDowell JE, Frank L, Brown GG, Braff DL (2001) Prefrontal, parietal, and temporal cortex networks underlie decision-making in the presence of uncertainty. Neuroimage 13:91–100.

Paulus MP, Rogalsky C, Simmons A, Feinstein JS, Stein MB (2003) Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. NeuroImage 19:1439–1448.

Paulus MP, Stein MB (2006) An insular view of anxiety. Biol Psychiatry 60:383–387.

Phan KL, Sripada CS, Angstadt M, McCabe K (2010) Reputation for reciprocity engages the brain reward center. Proceedings of the National Academy of Sciences 107:13099–13104.

Plato (2005) Phaedrus. Penguin Classics.

Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. Nature 400:233–238.

Prelec D (1998) The probability weighting function. Econometrica 66:497–527.

Premack D, Woodruff G, others (1978) Does the chimpanzee have a theory of mind. Behavioral and Brain sciences 1:515–526.

Preuschoff K, Bossaerts P (2007) Adding prediction risk to the theory of reward learning. Ann N Y Acad Sci 1104:135–146.

Preuschoff K, Bossaerts P, Quartz S (2006) Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. Neuron 51:381–390.

Preuschoff K, Quartz SR, Bossaerts P (2008) Human insula activation reflects risk prediction errors as well as risk. J Neurosci 28:2745–2752.

Pronin E (2007) Perception and misperception of bias in human judgment. Trends Cogn Sci (Regul Ed) 11:37–43.

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

Real L, Ott J, Silverfine E (1982) On the tradeoff between the mean and the variance in foraging: effect of spatial distribution and color preference. Ecology:1617–1623.

Redelmeier DA, Tversky A (1992) On the Framing of Multiple Prospects. Psychological Science 3:191–193.

Rilling JK, Gutman DA, Zeh TR, Pagnoni G, Berns GS, Kilts CD (2002) A neural basis for social cooperation. Neuron 35:395–405.

Roiser JP, de Martino B, Tan GCY, Kumaran D, Seymour BJ, Wood NW, Dolan RJ (2009) A Genetically Mediated Bias in Decision Making Driven by Failure of Amygdala Control. J Neurosci 29:5985–5991.

Rolls ET, McCabe C, Redoute J (2008) Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. Cereb Cortex 18:652–663.

Rousseau JJ (1754) A discourse on inequality. London: Penguin Books Ltd.

Sachser N, Pröve E (1986) Social Status and Plasma-Testosterone-Titers in Male Guinea Pigs (Cavia aperes f. porcellus). Ethology 71:103–114.

Samuelson PA (1938) A Note on the Pure Theory of Consumer’s Behaviour. Economica 5:61–71.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755–1758.

Sapienza P, Zingales L, Maestripieri D (2009) Gender differences in financial risk aversion and career choices are affected by testosterone. Proceedings of the National Academy of Sciences 106:15268.

Schonberg T, Fox CR, Poldrack RA (2011) Mind the gap: bridging economic and naturalistic risk-taking with cognitive neuroscience. Trends Cogn Sci (Regul Ed) 15:11–19.

Schwarz G (1978) Estimating the dimension of a model. The annals of statistics 6:461–464.

Scott RC, Horvath PA (1980) On the Direction of Preference for Moments of Higher Order than the Variance. The Journal of Finance 35:915–919.

Seymour BJ, McClure SM (2008) Anchors, scales and the relative coding of value in the brain. Curr Opin Neurobiol 18:173–178.

Seymour BJ, Singer T, Dolan R (2007) The neurobiology of punishment. Nature Reviews Neuroscience 8:300–311.

Simon HA (1972) Theories of bounded rationality. Decision and organization 1:161–176.

Singer T, Critchley HD, Preuschoff K (2009a) A common role of insula in feelings, empathy and uncertainty. Trends in cognitive sciences 13:334–340.

Singer T, Critchley HD, Preuschoff K (2009b) A common role of insula in feelings, empathy and uncertainty. Trends Cogn Sci (Regul Ed) 13:334–340.

Slamecka NJ (1963) Choice Reaction-Time as a Function of Meaningful Similarity. The American Journal of Psychology 76:274–280.

Slovic P (1964) ASSESSMENT OF RISK TAKING BEHAVIOR. Psychol Bull 61:220–233.

Smith A (1759) The Theory of Moral Sentiments.

Smith A (1776) The Wealth of Nations.

Stephens D (1981) The logic of risk-sensitive foraging preferences. Animal Behaviour 29:628–629.

Stephens DW, Krebs JR (1987) Foraging Theory, 1st ed. Princeton University Press.

Symmonds M, Bossaerts P, Dolan RJ (2010) A Behavioral and Neural Evaluation of Prospective Decision-Making under Risk. The Journal of Neuroscience 30:14380–14389.

Symmonds M, Wright ND, Bach DR, Dolan RJ (2011) Deconstructing risk: Separable encoding of variance and skewness in the brain. Neuroimage 58:1139–1149.

Tabibnia G, Satpute AB, Lieberman MD (2008) The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol Sci 19:339–347.

Talmi D, Dayan P, Kiebel SJ, Frith CD, Dolan RJ (2009) How Humans Integrate the Prospects of Pain and Reward during Choice. The Journal of Neuroscience 29:14617–14626.

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neurosci 7:887–893.

The Economist (2007) In the money. The Economist.

Thorndike EL (1911) Animal intelligence: Experimental studies. New York: Macmillan.

Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307:1642–1645.

Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2007) Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J Neurophysiol 97:1621–1632.

Toga A, Mazziotta J (2002) Brain Mapping: The Methods, Second Edition (Toga, Brain Mapping). Academic Press. Available at: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0126930198 [Accessed July 26, 2011].

Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The neural basis of loss aversion in decision-making under risk. Science 315:515.

Tricomi E, Rangel A, Camerer CF, O'Doherty JP (2010) Neural evidence for inequality-averse social preferences. Nature 463:1089–1091.

Tsakiris M, Hesse MD, Boy C, Haggard P, Fink GR (2007) Neural signatures of body ownership: a sensory network for bodily self-consciousness. Cereb Cortex 17:2235–2244.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211:453–458.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty 5:297–323.

Ungemach C, Chater N, Stewart N (2009) Are Probabilities Overweighted or Underweighted When Rare Outcomes Are Experienced (Rarely)? Psychological Science 20:473–479.

Varnavas GG, Grand W (1999) The insular cortex: morphological and vascular anatomic characteristics. Neurosurgery 44:127–136; discussion 136–138.

Vlaev I, Seymour BJ, Dolan RJ, Chater N (2009) The price of pain and the value of suffering. Psychol Sci 20:309–317.

Vul E, Kanwisher N (2010) Begging the question: The non-independence error in fMRI data analysis. In: Foundational Issues in Human Brain Mapping (Hanson SJ, Bunzl M, eds). MIT Press.

Weber EU, Blais A, Betz NE (2002) A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. Journal of Behavioral Decision Making 15:263–290.

Weiskopf N, Hutton C, Josephs O, Deichmann R (2006a) Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage 33:493–504.

Weiskopf N, Hutton C, Josephs O, Deichmann R (2006b) Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage 33:493–504.

Weller JA, Levin IP, Shiv B, Bechara A (2007) Neural correlates of adaptive decision making for risky gains and losses. Psychol Sci 18:958–964.

Wingfield JC, Hegner RE, Dufty AM, Ball GF (1990) The “Challenge Hypothesis”: Theoretical Implications for Patterns of Testosterone Secretion, Mating Systems, and Breeding Strategies. The American Naturalist 136:829–846.

Wittmann M, Leland DS, Paulus MP (2007) Time and decision making: differential contribution of the posterior insular cortex and the striatum during a delay discounting task. Exp Brain Res 179:643–653.

Wobber V, Hare B, Maboto J, Lipson S, Wrangham R, Ellison PT (2010) Differential changes in steroid hormones before competition in bonobos and chimpanzees. Proceedings of the National Academy of Sciences 107:12457.

Wright ND, Mechelli A, Noppeney U, Veltman DJ, Rombouts SARB, Glensman J, Haynes J-D, Price CJ (2008) Selective activation around the left occipito-temporal sulcus for words relative to pictures: individual variability or false positives? Hum Brain Mapp 29:986–1000.

Wright ND, Symmonds M, Fleming SM, Dolan RJ (2011) Neural segregation of objective and contextual aspects of fairness. J Neurosci 31:5244–5252.

Wu CC, Bossaerts P, Knutson B (2011) The Affective Impact of Financial Skewness on Neural Activity and Choice. PLoS ONE 6:e16838.

Wu S-W, Delgado MR, Maloney LT (2009) Economic decision-making compared with an equivalent motor task. Proceedings of the National Academy of Sciences 106:6088–6093.

Wunderle JM, Castro MS, Fetcher N (1987) Risk-averse foraging by bananaquits on negative energy budgets. Behavioral Ecology and Sociobiology 21:249–255.

Xue G, Lu Z, Levin IP, Bechara A (2010) The impact of prior risk experiences on subsequent risky decision-making: The role of the insula. NeuroImage 50:709–716.

Zethraeus N, Kocoska-Maras L, Ellingsen T, Von Schoultz B, Hirschberg AL, Johannesson M (2009) A randomized trial of the effect of estrogen and testosterone on economic behavior. Proceedings of the National Academy of Sciences 106:6535.