Individual decision making, reinforcement learning and myopic behaviour

The University Of Sheffield.

Alvin Pastore

Department of Computer Science
Faculty of Engineering

This dissertation is submitted for the degree of
Doctor of Philosophy

Supervisors:
Prof. Eleni Vasilaki
Dr. Tom Stafford
Prof. James Marshall

January 2019

Per Michele, Isa and Alessandra

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and are the result of my own work and research during the years of my PhD. None of the contents has been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Alvin Pastore
January 2019

Acknowledgements

I would like to thank and acknowledge my supervisory team, composed of Eleni Vasilaki, Tom Stafford, and James Marshall. Thank you for the guidance and assistance during my PhD, especially during the hard times.

I would also like to thank Prof. Ido Erev, who kindly provided the experimental data used for the first part of the work presented in this thesis.

I will always humbly thank my parents, who have been incredibly important to me, for their unconditional moral support, the faith and trust they put in me, and their wise advice.

Thank you to all my friends in Sheffield, the PhDs and the students, Ali and Rich and all the musicians, the photographers and the artists, the cafes and the restaurants, the pubs and the clubs. This beautiful city has given me so much in these years, I hope I contributed to its vibrant community.

Thanks to my Brother, who I knew to be an open-minded, loving and kind young man, but also proved to be exceptionally selfless and supportive.

Thanks to my girlfriend Heather, who never ceased to believe in me, who taught me so much in the past and keeps teaching me every day, about love, respect, open-mindedness and resilience.

Abstract

Individuals use their cognitive abilities to make decisions, with the ultimate goal of improving their status. Decisions outcomes are used to learn the association between the decisions which lead to good results and those resulting in punishing outcomes. These associations might not be easily inferable because of environmental complexity or noisy feedback. Tasks in which outcomes probabilities are known are termed "decisions under risk". Researchers have consistently showed that people are risk averse when choosing among options featuring gains, while they are risk seeking when making decisions about options featuring losses. When the probabilities of the options are not clearly stated the task is known as "decisions under ambiguity". In this type of task individuals face an exploration-exploitation trade off: to maximise their profit they need to choose the best option but at the same time they need to discover which option leads to the best outcome by trial-and-error. The process of knowledge acquisition by interaction with the environment is called adaptive learning.

Evidence from literature points in the direction of unskilled investors behaviour being consistent with naive reinforcement learning, simply adjusting their preference for which option to choose based on its recent outcomes. Experimental data from a binary choice task and a quasi-field scenario is used to test a combination of Reinforcement Learning and Prospect Theory. Both the investigations include reinforcement learning models featuring specific parameters which can be tuned to describe individual learning decision-making strategies. The first part is focused on integrating the two computational models, the second on testing it on a more realistic scenario. The results indicate that the combination of Reinforcement Learning and Prospect Theory could be a descriptive account of decision-making in binary decision tasks. A two-state space configuration, together with a non-saturating reward function appears to be the best setup to capture behaviour in said task. Moreover, analysing the parameters of the models it becomes evident that payoff variability has an impact on speed of learning and randomness of choice. The same modelling approach fails to capture behaviour in a more complex task, indicating that more complex models might be needed to provide a computational account of decisions from experience in non-trivial tasks.

Table of contents

List of figures

List of tables

Nomenclature

Expected Value Theory

$E[X]$	The expected value of a random variable X (same as $E\{X\}$)
p_n	The probability of the n-th outcome of a random variable X
x_n	The n-th outcome of a random variable X

Prospect Theory

α_{PT}	The coefficient of risk-avoidance behaviour in Prospect Theory
β_{PT}	The coefficient of risk-seeking behaviour in Prospect Theory
λ_{PT}	The coefficient of loss-aversion in Prospect Theory
$v_{PT}(x)$	The value function for the outcome x in Prospect Theory
$\pi_{PT}(p)$	The decision weight associated with probability p in Prospect Theory
V_{PT}	The value of a prospect according to Prospect Theory

Rescorla Wagner

α_{RW}	The salience parameter in Rescorla-Wagner model
β_{RW}	The association value parameter in Rescorla-Wagner model
λ_{RW}	The maximum associative value in Rescorla-Wagner model
V_{RW}	The value of the conditioned stimulus in Rescorla-Wagner model

Reinforcement Learning

α	Learning rate, parameter of the RL model regulating the speed with which new information is acquired by the RL agent
β	Inverse temperature, parameter regulating the sensitivity of a RL agent's policy to pick the available actions based on their values
ε	The rate with which a RL agent picks the alternative to the best known action when using the Epsilon-greedy policy
γ	Discount factor, parameter regulating the amount of future information taken into account by the RL agent
R	The return, or expected cumulative reward of the RL agent
π	Action policy of a RL agent, associating the environment states to probability of picking any available action
$A(s)$	The set of possible actions the RL agent can take from state s
a_t	The action taken by the RL agent at time t
O	Observation space of a POMDPX
$Q^{\pi}(s,a)$	Q-value function, describing the value of a state-action pair (s,a)
r_t	The reward obtained by the RL agent at time t
S	The set of possible states of the RL environment
s_t	The state of the environment at time t in the RL setup
$V^{\pi}(s)$	Value function, describing the value of state s
$Z(\cdot)$	Observation function of a POMDPX

Probability theory

θ_M	The vector of free parameters associated with model M, $\hat{\theta}_M$ denotes the set of parameters for the maximum likelihood estimate
D	The data set to be analysed by Maximum Likelihood Estimate for model selection
M	The model proposed as generating the data analysed
$P(D \mid M, \theta_M)$	The likelihood that the data D is generated by model M with parameters θ_M

Model components

AT	Average-tracking update rule
FH	Full history state space
ID	Identity reward function
LO	Latest outcome state space
PT	Prospect theory's subjective value function
QT	Q-learning update rule
SS	Single-state environment configuration
TH	Hyperbolic tangent reward transformation function

Policy Gradient

α_{PG}	The step-size of the policy gradient update
∇_θ	The gradient (the vector of partial derivatives) for the parameters in θ
τ	The episode in the policy gradient update formalism
J	The expected return in the policy gradient update formalism
P_{PG}	The policy parameters to be perturbed in finite difference gradients

Payoff variability in Erev and Barron [2005]

σ_{subj}	The standard deviation of the outcomes observed by a subject
o_i	The outcome observed by a subject during the i-th trial
PV_{subj}	The payoff variability experienced by a subject as defined in Erev and Barron [2005]

Value assement model in Barron and Erev [2003]

α_{VA}	The weight of the value of the obtained payoff during an exploratory trial in the Value Assessment model in Barron and Erev [2003]
β_{VA}	The weight of the value of the obtained payoff during a non-exploratory trial in the Value Assessment model in Barron and Erev [2003]
κ	The parameter regulating the strength of the exploration in the Value Assessment model in Barron and Erev [2003]
$S(t)$	The value of the payoff variability in the Value Assessment model in Barron and Erev [2003]

Chapter 1

Introduction

A toddler struggles to stand on her own two feet, balancing the weight above the hips to avoid tumbling. A few months later this process is refined, the challenge is now coordinating her steps to avoid tripping. A few years later the same child, on her first bike ride, will learn to adjust her equilibrium to get rid of the stabilisers. All these tasks are completely new to the child but she manages nonetheless to master each task.

A Formula 1 driver is testing the car on a new track, the lap-time is quite poor and she almost loses control of the vehicle at a chicane. After a two-days practice session, her lap-time has improved by several seconds, her driving has become more precise, allowing higher speeds through the chicane.

A hungry monkey found some palm nuts but these have a hard shell which cannot be opened by hand. By exploring the environment, the monkey learns that it is possible to crush the shell and recover the nut and that a big rock is required to do so. Instruments and tools are often used by primates for provisioning. Also, they learn to assess if a nut is ready to be opened or not by tapping on its shell and listening to the sound it makes, leaving it to dry in the sun in case it is not dry enough. This behaviour shows highly complex use of information previously acquired.

The previous examples are all instances of a learning process, that combines newly acquired information with existing knowledge to improve the outcome of ones behaviour. The ability to learn is one of the defining features of animals (including humans) and more recently, machines as well. From an ecological point of view, both humans and animals are capable of adapting to their environment by changing their behaviour, with the goal of increasing their chance of survival and reproduction. Quite interestingly, this process is commonly carried out under uncertainty, in environments that can only partially be observed. The toddler doesn't know what happens when shifting her weight off balance and will learn that toppling over on her head is associated with undesirable pain, a form of punishment.

The Formula 1 driver is rewarded with faster lap times after practising driving to improve the trajectory of the curve, but if the car is crashed in the process, the driver will learn a very expensive lesson on how not to approach a curve.

Learning is a complex process involving many areas of the brain, bringing together memory, reasoning, uncertainty evaluation and eventually judgement. Researchers from different fields of research are trying to understand learning and decision-making, by analysing these at different levels of detail. Trying to describe, replicate or predict human behaviour has been historically the ultimate goal of many disciplines. Economics for example, has a long tradition of trying to encapsulate behaviours into formal descriptive or normative models. For example, Expected Utility hypothesis has been employed for several years (von Neumann and Morgenstern [1947]) but has been shown to be inadequate, giving rise to new trends like behavioural economics (Kahneman and Tversky [1979]; Starmer [2000]). Psychology is intrinsically concerned with decision-making. Behavioural psychologists for example, have been using sequential decision problems to evaluate people's risk attitude (Frey et al. [2015]; Kahneman and Tversky [1979]; Pleskac [2008]; Wallsten et al. [2005]), in order to predict actual proneness to risk in real-life scenarios (Hoffrage et al. [2003]; Lejuez et al. [2003a,b, 2002]; Wallsten et al. [2005]). While economics and psychology are focused on the emerging manifestations of decision-making and its implications, neuroscience is a field of research which aims to understand the biological structure and the mechanisms underpinning human and animal decision-making (Barraclough et al. [2004]; Britten et al. [1996, 1992]; Gold and Shadlen [2001, 2002, 2007]; Shadlen et al. [1996]; Shadlen and Newsome [1996]).

Recently these fields of research have started to cooperate, contributing to the rise of a new multi-disciplinary field called Neuroeconomics (Glimcher and Rustichini [2004]; Loewenstein et al. [2008]; Sanfey et al. [2006]). These disciplines approach the problem from different perspectives and levels of abstraction but, under the new paradigm, they influence each other to achieve a unified account of how humans or animals make decisions. This emerging discipline of research has great potential for understanding and describing behaviour, but faces the challenge of satisfying different points of view in its investigations.

Understanding the decision-making process could help improve people's welfare. For example, knowing how people perceive health risks related to medical treatments and how they act accordingly, would allow policy makers to redesign the information campaigns to have a better impact on people's perception. People tend to dismiss warnings about global warming levels and the impact their behaviour has on it; this could be changed by re-framing the information with the objective of altering individuals behaviour and improving their environmental awareness.

The aim of this thesis is to improve the understanding of human decision-making. To do so, the first part of this work attempts to model a binary decision task involving repeated non-trivial choices, with computational models of learning and decision-making. Components of prospect theory are integrated within the models developed to investigate whether this could help better describe human behaviour. Moreover, a series of phenomena highlighted in literature, linked to the perception of variability and how this appears to impair learning and performance is investigated.

In the second part of this thesis, the work focuses on describing the choices made by a set of players from an online financial trading simulation game. Using indications provided by literature and previous studies concerning unskilled investors as a starting point, this work attempts to model the transactions made by the players. The work presented in the second part has been published in the proceedings of the 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), with the title "Modelling Stock-market Investors as Reinforcement Learning Agents" (Pastore et al. [2015]).

The next chapters are organised as follows. Chapter 2 provides a detailed literature review of the field of neuroeconomics, dissecting it from the perspective of each of the three fields: Economics, Psychology and Neuroscience. A final section within this chapter is devoted to introducing the computational modelling framework of reinforcement learning and some of the models which will be adopted in this work and that are widely used in literature. Chapter 3 offers a deeper analysis of the experimental studies which this thesis uses as a starting point. Furthermore, the details about the experimental setup for the data analysed in this thesis will be presented in this chapter. Chapter 4 focuses on the proposed attempt to link prospect theory and reinforcement learning and the investigation of phenomena relative to perception and performance. Chapter 5 offers an attempt to generalise previous findings by testing the insights deriving from the experimental study to a quasi-field task. In Chapter 6 we conclude, discussing the limitations of the present work and the potential future extensions which can help improve this work.

Chapter 2

Literature Review

Human behaviour was thought to be rational for a long time but empirical evidence, in many instances, proved this assumption to be wrong. The causes of the irrational choices observed in literature have been identified in multiple documented biases and phenomena; in fact, distortions of information perception and representation can lead to incoherent behaviour. An example is the "recency bias", consisting in relying on partial information deriving from undersampling (Erev and Roth [2014]; Hau et al. [2010, 2008]; Hertwig et al. [2004]; Hertwig and Erev [2009]; Hertwig and Pleskac [2010]; Plonsky et al. [2015]; Tversky and Kahneman [1974]) Other causes of irrational behaviour are "framing", an instance of emotional modulation (Benartzi and Thaler [2007]; Kahneman and Tversky [1979]; Tversky and Kahneman [1981]), and "mental accounting", which involves treating money as if belonging to different categories (Barber and Odean [2013]; Benartzi and Thaler [1995, 2007]; Camerer [1999]; Hsu and Chow [2013]; Shefrin and Statman [1985]; Thaler [1980, 1985, 1999]; Thaler et al. [1997]). Human behaviour has been the main concern of disciplines such as Economics, Psychology and Neuroscience. These research areas share a substantial interest in decision-making and learning; as a result, they started to collaborate and influence each other. A summary of how each discipline proceeded to study the phenomena associated with behaviour is presented in the following sections, and how their collaboration has helped bridge the gaps in understanding learning and decision-making. In the last section of this chapter we present a powerful adaptive learning computational framework called Reinforcement Learning (RL), widely used both as a descriptive account of decision-making in its many instances, and as computational mechanism for control tasks in engineering and automation.

2.1 Neuroeconomics

Neuroeconomics is a recent field of research which combines the effort of three prominent research areas interested in behaviour: economics, psychology and neuroscience. These operate at different levels of abstraction and detail. Historically, economists have tried to formalise decision-making into theories to describe behaviour or to characterise the normative approach. Often these theories would show lack of complexity or would make unrealistic assumptions (Loewenstein et al. [2008]; Platt and Huettel [2008]) resulting in attempts to extend their explanatory capacity. The attitude of researchers in psychology instead, is rooted in empirical studies, delaying the formalisation of theories to when enough data is gathered. The two disciplines have clear overlap of interests as their main focus is to understand human behaviour. Thanks to the technological advancements in the late 20th century, Neuroscience, the study of the brain and the nervous system, became a prominent area of experimental research, deeply entangled with psychology. While neuroscience studies the neural circuitry and the relative computations, psychology examines the emerging effects of such computations. Even if these two research areas focus on different levels of detail, they both share a great deal of interest in the study on human behaviour.

Following this outline, an example can help clarify how these three fields are intertwined. When a traffic light turns yellow and it is about to turn red, a driver faces a choice whether to help the inertia of the car by accelerating or to slow the vehicle to a stop before the yellow light switches to red. The driver is facing a decision in which the outcomes are potentially crashing her car for entering the intersection too late or losing a certain amount of time at the stop. From the economics point of view, the driver's behaviour is a utility problem. It can be formalised to describe the observed behaviour or to develop a normative account, to determine which course of actions will maximise the driver's utility. If the driver is incredibly late, speeding through might be considered, therefore taking the risk of an accident to avoid a loss of time. Conversely, if the driver is in no rush, taking the risk of an accident to save a few seconds is not a good idea. A psychologist's point of view instead, would consider whether ambiguity has a negative effect on the driver's behaviour. For example, testing whether the information about the probability of crashing for speeding through an intersection would change the driver's behaviour, making her more risk-averse. Examining the same scenario at the lowest level of detail, it can be described as a stimulus (i.e. yellow/red traffic light) perceived through the visual apparatus, elaborated by the brain by combining previous experience and knowledge (e.g. past accidents and possible warning road signs) with current information (e.g. distance from the crossing, current speed), and resulting in the driver's decision to push either pedal, to accelerate or break. These three disciplines could benefit from each other, as they could inform each other, by sharing data,

models and theories. Neuroeconomics is the attempt to do this, with the ultimate goal of understanding and formalising behaviour, by making a wiser use of the resources and the knowledge base from each of these fields. We will now present these disciplines in more detail providing a summary of the research studies of interest for this work.

2.1.1 Economics Perspective

Expected Value Theory

A first attempt to formulate a decision theory to make optimal choices under uncertainty was made by Blaise Pascal and Pierre Fermat during the 17th century. Their epistolary exchange about the problem of points lead to both of them independently solving it (David [1962]). The problem involves two players competing in a game of chance. They agree that the first player to win a certain number of rounds wins the game, they both have the same probability of winning a round and contribute evenly to the prize pot. The problem arises when, stopping the game at any point before the end, the prize has to be divided between the players fairly. A few years later, Christiaan Huygens wrote the "De ratiociniis in ludo aleae" [1] in which he used Pascal's and Fermat's same principles to solve the problem of points, thus laying the foundations of probability theory. The expected value, or "mathematical expectation" of a random variable X ($E[X]$), which can take N values $x_1, \cdots, x_n$ with probabilities respectively $p_1, \cdots, p_n$ is:

$$E[X] = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n \tag{2.1}$$

As an example, a common six-faced die can be seen as a random process in which the potential outcomes are the values of the faces and the probability of each is $1/6$ (being it a fair die), then:

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5 \tag{2.2}$$

In this case the expected value is exactly the arithmetic mean but in a general case, when the probabilities of the possible events are not all the same, the expected value becomes the weighted average of the outcomes.

[1] tr. On Reasoning in Games of Chance

Expected Utility Theory

Expected value theory is flawed and this has been highlighted not long after its formulation. Less than a century after EV theory was first formulated, in 1738, Daniel Bernoulli used the St. Petersburg paradox to show how this theory was flawed from both a descriptive and a normative point of view (Jensen [1967]). The paradox was described by Daniel's cousin Nicolas, a quarter of a century earlier. The St. Petersburg paradox is about the estimation of the price to pay to enter a casino offering a gamble on a fair coin being tossed multiple times. The gamble is the following: the initial stake is 2 dollars and this gets doubled each time the coin toss lands on heads. The player wins the stake when the coin toss results in a tails. Therefore, the player wins 2 dollars if the coin lands tails at the first trial, 4 dollars if it lands tails on the second trial, 8 dollars if the first two coin tosses land on heads and the third lands on tails, and so on. Generalising, the player wins 2^t dollars where t represents the number of tosses after which tails appears. What can be considered a fair amount of money to pay the casino, m_{cas}, to be allowed to play this gamble? According to expected value theory, assuming the casino has infinite resources, the expected value of this gamble is:

$$\begin{aligned} E &= \frac{1}{2} \times (2 - m_{cas}) + \frac{1}{4} \times (4 - m_{cas}) + \frac{1}{8} \times (8 - m_{cas}) + \frac{1}{16} \times (16 - m_{cas}) + \cdots \\ &= 1 + 1 + 1 + 1 + \cdots - \frac{m_{cas}}{2} \times \frac{1}{1 - \frac{1}{2}} \\ &= 1 + 1 + 1 + 1 + \cdots - m_{cas} \\ &= \infty . \end{aligned} \tag{2.3}$$

This means that when offered to play this game, potential participants should be willing to play this game at any cost. In fact, no matter how high P is, the expected winning of this gamble is infinite. On the practical side though if P is high enough, there is very little chance of being able to recover the initial investment, let alone make a profit. As an example if the price to play this gamble P is 128 dollars, it will take at least 6 consecutive heads followed by tails to get at least the initial stake back. The probability of this happening is 2^{-t} with $t = 6$, that is 1 chance in 64. It is evident that this gamble is very unattractive, even if mathematically its expected value is infinite. Bernoulli shifted the view from the monetary value of the gamble to the utility it can potentially yield. He postulated that the price to pay for such a risky game should not be estimated from the monetary expected value but instead from the expected value of the utility deriving from the potential wins, denoting this as "moral expectation", in contrast with the EV mathematical expectation. This conclusion derives from the belief that a certain gain will yield different utility to different

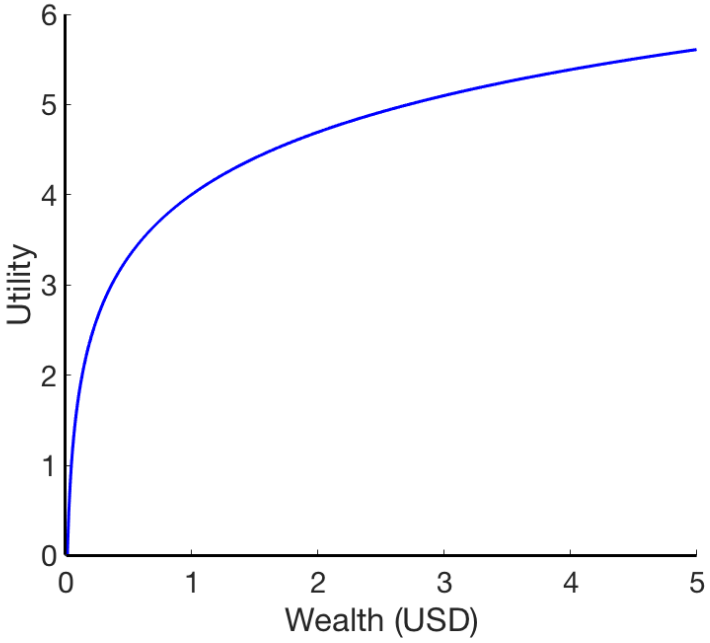

Fig. 2.1 A hypothetical Logarithmic Utility Model

individuals. Bernoulli even suggested adopting a logarithmic function as a utility model. This would signify that the utility of a win increases with the amount of the win itself but at a slower pace as shown in Fig. 2.1. By adopting this model, or other concave functions, it is possible to account for the risk aversion phenomena shown by people faced with this type of gamble. The expected utility theory (EUT) evolved during the years, eventually becoming the foundation for Von Neumann and Morgenstern game theory (Jensen [1967]; von Neumann and Morgenstern [1947]).

Prospect Theory

Despite expected utility theory models representing a simple, somewhat effective way of formalising decisions in uncertain environments, they are often unsuccessful in describing real-world behaviour (Loewenstein et al. [2008]; Starmer [2000]). As an example, in the famous franchise game "Deal or No Deal", which aired on television channels all over the globe, uncertainty perception and relative decisions by contestants were shown to be influenced by the history of their former choices (Platt and Huettel [2008]; Post et al. [2008]). According to expected utility these decisions should have been made solely on the basis of the prizes available at each point in time. In the real world, these violations are frequent and often appear to happen when the decision-maker (DM) is faced with risky or ambiguous probability distributions. The concept of ambiguity is distinct from risk and follows Ellsberg's

terminology (Ellsberg [1960]). Decisions under risk happen in environments in which probabilities about the outcome of the choices are known and represented symbolically (e.g. with percentages, frequencies or pictorial representations such as pie charts).

In 1979, Kahneman and Tversky criticised expected utility theory and presented a new model called prospect theory, which will become extremely influential in literature (Kahneman and Tversky [1979]). Specifically they identified the following phenomena which systematically violate EUT. The underweighting of uncertain outcomes as opposed to certain outcomes. This phenomenon leads to risk-aversion when individuals face choices involving certain wins and symmetrically, to risk-seeking behaviour when the same subjects face choices involving certain losses. They refer to this paradigm of choice as the reflection effect. Moreover, they identified a phenomenon that causes people to neglect common features of the available choices and focus on the characteristic traits. This inclination is called isolation effect and leads to inconsistencies in choice when the same alternatives are presented in different formats. As these effects exposed the flaws of EUT as a descriptive theory of decision-making, Kahneman and Tversky went on to propose their prospect theory. It is important to note that this theory was originally developed for decision tasks involving descriptive probabilities in which the subjects are informed from the beginning of the task about the options features (payoff and probability). Their account subdivides the decision process into two stages. The first is called editing while the second is named evaluation. The editing phase consists of a simplification of the available choices by means of the following operations. The DM makes adjustments about the reference point, according to which she judges the outcome of the choices. To back this claim, Kahneman and Tversky provide evidence of people's perception of the outcomes, not as ending states of wealth but as gains and losses relative to their personal reference point. The framing of the choices can affect the position of the reference point and the subsequent encoding of the potential gains and losses. Framing refers to the different possibilities of presentation of a problem, such as changes of perspective; these should not influence a rational decision-maker, in real life though framing can greatly sway people's decisions (Tversky and Kahneman [1981]).

The first to suggest that individuals use a personal reference system to judge the subjective utility of gains and losses, instead of using the final asset values, was Markowitz (Markowitz [1952]). In their account, Kahneman and Tversky equate judgement to human perception. Concrete examples of the way perception accommodates the assumption that previous experiences move the reference point for future assessment can be found in temperature or brightness perception. The same object can be perceived as hot or cold depending on an individual's previous state. Drinking a cup of hot chocolate, after spending a day in the snowy Alps in winter, is certainly more enjoyable than having the same drink during a hot

day of summer in a southern Italian city. Kahneman and Tversky also indicate how this assumption had already been widely accepted in experimental studies on utility assessment (Kahneman and Tversky [1979]). Combination is the operation by which some options can be conflated when they exhibit equivalent outcomes. As an example, if two options, with different probabilities $p_1 = 0.15$ and $p_2 = 0.25$, are available and both of them offer 100 dollars as a result, then they can be combined into a single prospect of 100 dollars potential win with probability $p_3 = p_1 + p_2 = 0.40$. When choices include a risk-free option this is separated from the risky counterpart, this process is called segregation. For example, from the choice between 500 dollars with probability $p_1 = 0.20$ and 100 dollars with probability $p_2 = 0.80$, an individual will extract a sure outcome of 100 dollars with a risky alternative consisting of the difference in the outcomes ($500 - 100 = 400$ with probability $p_3 = 0.20$). This happens in the same way in the losses domain. Cancellation happens when subjects ignore parts of available options shared among the various choices. For instance, if option A offers the following outcomes: (100, 0.20; 50, 0.30; 0, 0.50)[2] and option B offers: (100, 0.20; 100, 0.30; 25, 0.50), then the result of the cancellation operation will be a choice between A = (50, 0.30; 0, 0.50) and B = (100, 0.30; 25, 0.50). While the previous operations were applied to single prospects independently, cancellation is applied to a set of multiple choices. Another intuitive operation is simplification, which consists of rounding the outcome and/or probability stated in the prospect (e.g. 299, 0.51 becomes 300, 0.50). Despite its simplicity, this operation is quite powerful, since it could potentially lead to neglecting outcomes with very small probabilities. The last operation applied by subjects in the editing phase is the detection of dominance, consisting of the rejection of alternatives which are considered disadvantageous. For example, if A = (1000, 0.31; 201, 0.69) and B = (1000, 0.20; 199, 0.71), the subject will disregard option B altogether, after simplifying the second part of both options to (200, 0.70).

In the first phase the complexity of the choice is reduced while in the second phase, called evaluation, individuals compute the value of each option and then choose the highest one. This value is based on two quantities which in turn are a function of probabilities and outcomes. Each probability p is associated with a decision weight $\pi_{PT}(p)$ and each outcome x is associated with a number $v_{PT}(x)$. The scale π_{PT} is not a measure of probability in itself but represent the influence of the probability of an outcome on the general value of the choice. The scale v_{PT} instead, represents the subjective value of an outcome or how much an outcome deviates from the reference point in either the direction of gains or losses. The idea of using a weighting system instead of raw probabilities had already been proposed by

[2] for the sake of brevity the gambles notation assumes dollar as currency for the outcomes followed by the probability associated in the format (outcome, probability); e.g. (100, 0.2; 0, 0.8)

Edwards (Edwards [1962]; Kahneman and Tversky [1979]). Therefore, for several potential outcomes $x_1, \cdots, x_n$ with associated probabilities $p_1, \cdots, p_n$ the value V_{PT} will be calculated as:

$$V_{PT} = \sum_{i=1}^{n} \pi_{PT}(p_i) v_{PT}(x_i) \tag{2.4}$$

where the value function v_{PT} is formally defined as:

$$v_{PT}(x) = \begin{cases} x^{\alpha_{PT}} & \text{if } x \geq 0 \\ -\lambda_{PT}(-x)^{\beta_{PT}} & \text{if } x < 0 \end{cases} \tag{2.5}$$

with $\lambda_{PT} > 1$ being the coefficient of loss-aversion , and the parameters α_{PT} and β_{PT} being the coefficients of risk-avoidance or risk-seeking behaviour, respectively in the gain or loss domain, which range from 0 to 1 (Nilsson et al. [2011]). The values of these parameter was not originally estimated in Kahneman and Tversky [1979]. Subsequently in a refinement of their theory, the same authors provided parameter values based on the median estimates of the subjects in Tversky and Kahneman [1992] The parameters were estimated on a subject basis with a non-linear regression. Their median values were then presented: $\lambda_{PT} = 2.25$ and $\alpha_{PT} = \beta_{PT} = 0.88$.

Considering prospects of the type $(x, p; y, q)$, where outcome x has probability p of happening, y has probability q and outcome nothing has probability $1 - p - q$, the prospect will be considered regular if either $p + q < 1$ or $x \geq 0 \geq y$ or $x \leq 0 \leq y$. In this case:

$$V_{PT}(x, p; y, q) = \pi_{PT}(p) v_{PT}(x) \tag{2.6}$$

In the cases in which $p + q = 1$ and one of $x > y > 0$ or $x < y < 0$ is true (respectively strictly positive and strictly negative prospects), the segregation operation from the editing phase separates the available prospect into risk-free options which is the certain gain or loss and the risky component which is the probabilistic gain or loss.

$$V_{PT}(x, p; y, q) = v_{PT}(y) + \pi_{PT}(p)[v_{PT}(x) - v_{PT}(y)] \tag{2.7}$$

The value of such prospects is the value of the risk-free component summed to the difference of the values of the two outcomes, multiplied by the perceived influence of the probability of the more extreme outcome. The decision weight $\pi(p)$ is used on the risky part of the option (the value difference) but not on the risk-free component.

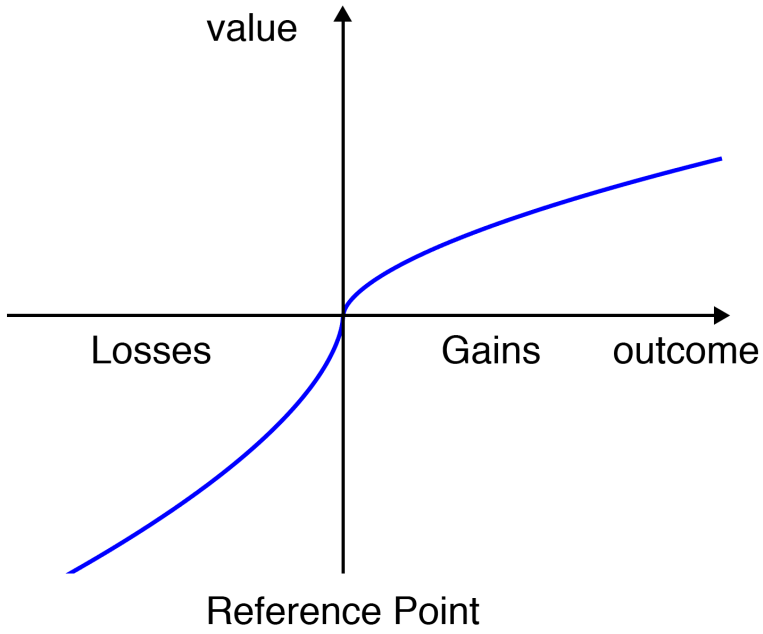

Fig. 2.2 Prospect theory hypothetical subjective value function, generated with $\alpha_{PT} = \beta_{PT} = 0.88$ and $\lambda_{PT} = 2.25$ estimated as median in Tversky and Kahneman [1992]. The kink in the origin characterises the principle of loss aversion for which losses loom larger than gains.

The prospects offered to individuals by Kahneman and Tversky are descriptions of monetary gains or potential holiday trips (Kahneman and Tversky [1979]). These decisions are assumed to be made under risk, the subject has a symbolic representation of outcomes probabilities and makes her choice based on this information. On the other end of the uncertainty spectrum are decisions under ambiguity which are characterised by lack of prior knowledge about outcomes and therefore require at least a certain degree of exploration for the individual to operate an informed decision. This exploration process consists, as an example, of sampling the available options. People violate expected utility again in this type of setup. A clarifying example is the Ellsberg paradox (Ellsberg [1960]), in which a subject is offered a choice between two urns (Fig. 2.3). The first contains 100 balls, of which 50 are white and 50 are black. The proportion of white and black balls in the second urn is not known. By drawing a ball of a specific colour the player wins 100 dollars. Participants have the tendency to choose indifferently between white or black when drawing from the first urn, indicating that their belief about probability distribution for such urn is uniform between white and black. Individuals show the same preference when separately betting on the second urn, which indicates they hold similar beliefs for the two urns. Nevertheless, subjects are more likely to bet on a specific colour if they are allowed to draw the ball from the first urn,

the one in which the proportion of balls is known. This behaviour is clearly inconsistent, the subjects who did not show any preference between colours in the second urn, now show a preference for the first urn, indicating that they believe the balls to be equally distributed (Ellsberg [1960]; Loewenstein et al. [2008]). Some possible sources of this phenomenon have been proposed in psychology works, which will be analysed more in depth in section 2.1.2.

Fig. 2.3 A graphical representation of the Ellsberg paradox: an example of the violation of subjective expected utility theory. The proportion of white and black balls in the first urn is known while it is not in the second. Subjects considering these two gambles separately, one risky and the other ambiguous, show the same preference. Indifferently betting on any colour being drawn. When offered the chance to bet on either of the two urns, subjects prefer the risky (known distribution) option to the ambiguous (unknown distribution).

A decision scenario can become even more complex when the payoff distributions are not stationary but change over time. Who can tell what will be the outcome of entering the stock market by purchasing stock A or stock B? Could the time investors enter the market affect the performance of their portfolio? The former question represents the "portfolio selection" task, the latter is known as the "market timing" problem. These are only two of the many problems investors face when dealing with financial markets. Unskilled investors - but also many professionals - tend to achieve a suboptimal return in these markets because of their flawed behaviour (Barber et al. [2007, 2014]; Benartzi and Thaler [1995]; Frey et al. [2015]; Lakonishok et al. [1992]; Odean [1998]; Shapira and Venezia [2001]; Strahilevitz et al. [2011]; Weber and Camerer [1998]; Weber and Welfens [2011]).

Table 2.1 **U.S Returns 1802-2000**

Period	Mean Real Return		Risk Premium
	Market Index	Relatively Riskless Security	
1802-1998	7.0%	2.9%	4.1 pps
1889-2000	7.9	1.0	6.9
1926-2000	8.7	0.7	8.0
1947-2000	8.4	0.6	7.8

Annual yield for U.S. Market and compared riskless security along with the risk premium (pps = percentage points). Data from Mehra [2003], originally from Siegel [1998], and Mehra and Prescott [1985].

Flawed Investors Decision-Making: the Equity Premium Puzzle

An interesting example of suboptimal financial behaviour is characterised by the systematic underinvestment in stocks as opposed to bonds. Investing in bonds, in fact, has been shown to underperform the market systematically and for many years (Mehra and Prescott [1985]). Nevertheless, investors keep showing this behaviour. The difference in return between equities (i.e. stocks) and risk-free options (e.g. bonds) is called the equity risk premium and is considered the premium compensating an investor for the risk taken by choosing a risky asset over a risk-less one. The intrinsic risk associated with equities alone, is not enough to explain the reluctance of investors to choose stocks over bonds. This inadequate pattern of behaviour is known as the Equity Premium Puzzle (EPP) and was first brought to attention by Mehra and Prescott [1985]. In their work the annual returns for an investment in the Standard and Poor 500 index are shown to have outperformed short-term debt average returns. Specifically in the 90 years considered, between 1889 and 1978, the average return of the S & P 500 index had been 7% while the average return of Treasury Bills (T-bills) had been less than 1%. This data is extended in Mehra [2003], including 110 years of average annual real returns providing evidence of the relevance of the EPP in modern times. Different time periods are presented in Fig. 2.1, showing the real returns (inflation-adjusted returns) and how the risk premium is unrealistically high for each of them.

In their work, Mehra and Prescott [1985], find that the level of risk aversion which characterise the EPP is implausibly explained by the combination of low risk-free rate and high equity premium together. According to their calculations, the coefficient of risk aversion needed to explain such effect would be one order of magnitude higher than the ones commonly used in literature. This phenomenon contributed to emphasise the failure

Table 2.2 **Returns for Selected Countries, 1947-1999**

Country	Period	Mean Real Return		Risk Premium
		Market Index	Relatively Riskless Security	
United Kingdom	1947-99	5.7%	1.1%	4.6 pps
Japan	1970-99	4.7	1.4	3.3
Germany	1978-97	9.8	3.2	6.6
France	1973-98	9.0	2.7	6.3

Annual yield for United Kingdom, Japan, Germany and France. Data from Mehra [2003], originally from Siegel [1998], and Campbell [2003].

of standard neoclassical financial economic theories at explaining financial behaviour in uncertain scenarios. In Benartzi and Thaler [1995] it is reported that annual real return of stocks has been about 7% from 1926 to 1995, as opposed to the meagre less than 1% yielded by T-bills. In their work, Benartzi and Thaler [1995] provide a theoretical account of how the EPP could be explained by using two concepts from prospect theory: loss-aversion and narrow-framing. This theoretical account has been tested experimentally by Thaler et al. [1997] with good results. Their findings indicate that a myopic framing of outcomes could help explain why investors are more willing to settle for lower returns from risk-free (or low-risk) options instead of chasing high returns from high-risk prospects. Thaler et al. [1997] work will be analysed in greater detail in Chapter 3 because of its crucial role in this thesis. Another potential explanation of the EPP lies in the direction of extrapolative investor behaviour (Choi and Mertens [2006]). According to this proposal, investors hold the belief that previous performance has predictive power over future performance. There is evidence of correlation between recent performance and subjective predictions (Andreassen and Kraus [1990]; De Bondt [1991]; Fisher and Statman [2000]). This belief is commonly considered irrational as the market returns have been shown to possess no correlation (Fama and French [1988]). An effort to understand investors behaviour can be made by acknowledging their adaptive learning skills. Adaptive behaviour, linked to the "Law of Effect" (Thorndike [1898] presented in section 2.1.2), could be a potential explanation for apparently irrational behaviour. The suboptimal choices made by investors which give rise to the equity premium puzzle (Barron and Erev [2003]; Zion et al. [2010]) are an instance of such irrational behaviour. The indication that adaptive behaviour could help explain the EPP is reflected by the findings in Choi et al. [2009], who studied 401(k) retirement savings behaviour. Their results indicate that those investors who experienced a positive outcome from previous savings accounts,

increase their commitment to such assets more than those who experienced lower outcomes. The documented behaviour is compatible with a naive reinforcement learning approach (Choi et al. [2009]), which is generally a reasonable approach in other tasks. For a large part of human history, decision-makers operated in scenarios where they did not have access to high level information about their options; for example, they did not know the long term implications of the actions the took. In the wild it is often the case that behaviour which has led to good outcomes in the past will produce similarly valuable outcomes in the future. This is not the case in financial markets (Choi et al. [2009]; Fama and French [1988]).

An analogous pattern of behaviour is found in Huang [2012]. By analysing data spanning 5 years and starting from 1991, the author provides evidence that positive trading experience in a particular industry increases the likelihood of engaging in trading similar securities within that industry as compared to other industries. The results in Huang's article are robust to external effects such as industry momentum or wealth effects. Two other important insights from Huang's research are that time dampens the effect of purchase from the same industry and that higher levels of investor sophistication attenuate the effect.

Similar overweighting of personal experience has been documented in Barber et al. [2009]; De et al. [2010]; Kaustia and Knüpfer [2008]; Malmendier and Nagel [2011]; Strahilevitz et al. [2011], and later reported in Barber and Odean [2013]. This type of financial behaviour is assimilable to the well documented chasing of past returns (Chevalier and Ellison [1997]; Ippolito [1992]; Sirri and Tufano [1998]; Zeckhauser [1993]) and is related to phenomena of market overreaction (Chopra et al. [1992]; De Bondt and Thaler [1990, 1985]; Offerman and Sonnemans [2004]). As previously noted, in financial scenarios it is not guaranteed that repeating behaviour which resulted in rewarding outcomes will yield good payoffs again. Such naive behaviour could potentially lead to catastrophic events, as in the case of the sub-prime mortgage crisis in 2008 (Zion et al. [2010]). An explanation of the causes for this large scale event is identified in a pattern of behaviour exhibited by the mortgage and trading agents. Their risky behaviour resulted in rewarding outcomes for these financial agents, leading to more risk being taken in their future choices, fuelling a vicious cycle which resulted in disastrous defaults and repossessions.

2.1.2 Psychology Perspective

Classical and Instrumental Conditioning

In learning, for both humans and animals, there are two distinct scenarios which the learner can face. The first of the two is known as classical conditioning[3] while the second type is instrumental conditioning. In the classical conditioning kind of learning procedure, a subject is presented with a neutral stimulus, for example a light, a buzzer sound or a bell; this is also called a conditioned stimulus (CS). Some time after this, the subject receives a valuable stimulus, which can be palatable food or some other attractive or desirable stimulus. This is called the unconditioned stimulus (UCS) and it usually results in an observable response. Classical conditioning refers to the process by which the subject learns to associate the two stimuli after experiencing them repeatedly in succession. In the famous Pavlovian dog experiment the conditioned stimulus (bell ring) is associated with an unconditioned stimulus (food) to the extent that, after removing the UCS, the dog's response (salivation) is elicited solely by the CS. The response, initially appearing at the time when the UCS was presented and named unconditioned response (UR), appears after learning the association of the stimuli at the time of the CS and takes the name of conditioned response (CR). In this type of scenario the subject does not operate a choice deliberately but simply shows a shift in response timing, from the time when the reward is received to the time when the CS is presented. A quantitative psychological model which captures this type of learning is the Rescorla-Wagner (RW) model (Rescorla and Wagner [1972]). This model is based on a discretisation of the conditioning process in trials, during which the subject learns to associate the CS to the UCS and exhibits a conditioned response. In this model there are two scalar quantities which encode the values of UCS and CS: the value of the unconditioned stimulus u_{RW} and the strength of the conditioned response v_{RW} which by extension captures the value of the CS. In the RW model, learning happens when, via repeated presentation of the CS and UCS, the value v_{RW} converges to u_{RW}. This convergence happens through updating the value of v_{RW} with difference between the two values, also known as prediction error δ_{RW}. This prediction error can be thought of as the "surprise" a subject experiences when the US appears. This model has two parameters which represent the salience of the CS and the strength of the US and they are respectively denoted with α_{RW} and β_{RW}. The Rescorla-Wagner model is formally defined as:

$$V_{t+1} = V_t + \alpha_{RW}\beta_{RW}(\lambda_{RW} - V) \tag{2.8}$$

[3] also respondent conditioning or Pavlovian conditioning, after Ivan Pavlov (Pavlov [2010]), the first to study this type of interaction with dogs experiments

where V_{RW} is the associative value of the CS with V_{t+1} and V_t respectively the new and previous values of V_{RW}. The salience is the parameter $\alpha_{RW} \in (0,1)$ and the association value (or strength) of the US is the parameter $\beta_{RW} \in (0,1)$. Salience can be tweaked by varying the stimulus intensity (for example the frequency of a buzzer) with $\alpha_{RW} \to 0$ if the CS does not draw any attention and $\alpha_{RW} \to 1$ if the CS draws the maximum attention of the subject. The strength of the US is encoded by β_{RW} and represents the learning-rate parameter in the RW model, with $\beta_{RW} \to 1$ meaning that the subject exhibits a great rate of conditioning. Finally, λ_{RW} is the maximum associative value that can be attributed to the CS, achievable within the experimental setup. It represents the strength of the UCR as a result of the US and captures the potential impact of time delays between stimuli. Classical conditioning allows for the study of reward prediction error signals in the brain.

Instrumental conditioning (also known as operant conditioning) instead, is a type of conditioning based on the active interaction of the subject with the environment. The consequences of the subject's behaviour influence the learning process. As opposed to classical conditioning, in instrumental conditioning what is learnt is the action (or course of actions) resulting in the best rewarding outcome. On the other hand, if the outcomes are negative, what is learnt is avoidance behaviour. Edward L. Thorndike has been the first to study this type of learning, investigating cats behaviour in an escape task (Thorndike [1898]). Cats trying to escape from puzzle boxes, initially took a long time to find the action that would set them free (e.g. pulling a cord or pushing a button). With more experience, this association of action and outcome became more clear and eventually cats would take the right action straight-away and escape. The observation that behaviours resulting in good outcomes are repeated while actions leading to poor consequences are avoided became Thorndike's "Law of Effect". As a result of this law, the quality of consequences, either satisfying or dissatisfying, leads respectively to the strengthening or weakening of behaviour. The positive outcomes can be considered "incentives" and represent the ultimate objective of behaviour stemming from the actions-outcomes associations (Dickinson and Balleine [1994]). These considerations allow us to draw a parallel with natural selection and evolutionary theories (Darwin and De Beer [1956]) on two levels. In a narrow view, only the stimuli and actions leading to good outcomes emerge within the action-selection process in animals, leading to improved fitness and life-expectancy maximisation. As an example, the association of an eye-catching colourful fruit with the tasty reward leads to an improvement of health for an animal. In the opposite direction, an unripe green colour will not elicit the same approaching behaviour for the animal. It follows that ripe fruits will be preferred while green fruits will be left hanging by their branches. A famous instrumental learning example is the hot stove

effect, for which a cat jumping on a heated cooker gets a punishment so bad the behaviour is never replicated, thus improving future fitness. In a more broad view, those animals which exhibited a good learning ability over the course of their life are more likely to reproduce, passing on their abilities to their offspring. Those unfortunate individuals who were deficient in these learning mechanisms and did not learn, or took too long to learn the best behaviour perished. As an example, the zebras which never learnt to run away from the water hole when the crocodiles got too close will have a lower probability of passing on their learning machinery.

Work in non-human primates involving both types of conditioning tasks provided some evidence of this. Mirenowicz and Schultz [1994] recorded single neuron responses of two Macaca fascicularis monkeys in behavioural tasks and showed that 75% of dopamine neurons (ventroanterior midbrain, substantia nigra pars compacta and dorsally adjoining groups) produced a short-latency phasic response to an unexpected reward (apple juice liquid drops) and that the same neurons stopped responding after the learning was completed. These neurons responded then, at the time when the conditioned stimulus was presented. Similar results have been fond in following works, including Schultz [1998] and Hollerman and Schultz [1998]. These findings are not well captured by a Rescorla-Wagner model. An extension of the RW rule that takes into account the timing of the interactions and that takes the name of temporal-difference learning is more representative of the results (Montague et al. [1996]; Schultz et al. [1997]). The characteristic neuronal response pattern in classical conditioning, shifting of the firing backwards from the UCS to the CS after learning, is well approximated by this model (O'Doherty et al. [2003, 2007]). These types of behaviour are also studied in contexts in which outcomes are not deterministic but involve a certain degree of uncertainty (Niv et al. [2002]; O'Doherty et al. [2004]). Non-deterministic tasks are more representative of the class of decision-making scenarios humans face in everyday life. In this regard it is convenient to clarify the terminology used and the findings of previous works in this area of psychology.

Risk-Ambiguity Difference, the Description-Experience Gap and Payoff Variability Effect

As briefly introduced in the previous sections, the literature on experimental studies features two types of configurations in which decision-makers operate their choices: risk and uncertainty. This distinction follows Knight's concepts (Knight [1921]) and Ellsberg's terminology (Ellsberg [1960]). Choice behaviour is said to be made under risk in those scenarios in which the probabilities of the options are known. On the other hand, scenarios regarding options without defined outcomes description are labelled as ambiguous choices. Individuals who are

normally risk-averse show even stronger aversion in ambiguous situations (Platt and Huettel [2008]). Generally, when faced with the decision between a risky and an ambiguous option people tend to avoid the ambiguous one. For example, between two medical treatments of which one has known probability of success (50%) and the other provides no information, people prefer the first (Curley et al. [1986]). The same type of ambiguity-avoidance has been shown in money based studies (Becker and Brownson [1964]; Curley and Yates [1985]; Slovic and Tversky [1974]). Many hypotheses have been proposed to explain such ambiguity-avoidance; among these there is the "hostile nature" hypothesis. According to this, subjects favour the options providing more information, because they believe the ambiguous options would produce adverse outcomes, according to an intrinsic non-randomness (Curley et al. [1986]; Ellsberg [1963]; Loewenstein et al. [2008]). Experimental work by Curley et al. [1986] provided no support for this hypothesis. In the same work another hypothesis was tested and supported by experimental data. This was the "other-evaluation" hypothesis, which assumes that subjects operate decisions anticipating the moment in which their actions will be assessed by someone else, leaning towards choices that are more justifiable (Curley et al. [1986]; Ellsberg [1963]; Knight [1921]; Loewenstein et al. [2008]).

Real life decisions are comprised of both types of scenarios but their observed distribution in every day life is highly skewed towards ambiguous decisions. It is very often impractical, although not impossible, to calculate the probability of events such as a house being struck by lightning or winning the national lottery. At the time of writing a national lottery radio advertisement claims that "playing makes it possible" but does not state the actual probability (roughly one chance in 14 million). In Kahneman and Tversky [1979] study, decisions were made by informed individuals who were told the probabilities of the prospects before deciding which one to pick; because of their structure, these tasks have been denominated "decisions from description". On the other end of the spectrum of information provided before a choice there are tasks in which no prior information is available and decision-makers have to build an idea of which option is best based on experience. This task structure is further divided in two types of information paradigms: the information obtained from subjects' interactions can be partial or complete. If only the outcome of the selected prospect is presented the task is referred to as "partial-feedback" (also known as "minimal information" in some works); if both the selected and forgone payoffs are shown to the decision-maker the task assumes the name of "full-feedback". Within this subset of choice tasks with no prior information there are two possible configurations that can be envisioned. In the first, decision-makers are allowed to experiment with the available options. They can sample them to construct a belief about which option is best to pick but only the final selection actually counts as their decision. These tasks represent the sampling-paradigm (Hertwig and Erev [2009]). These

tasks are similar to decisions from description in that they are one-shot problems but the information provided to the decision-makers have a different source, experience instead of description. Differently, tasks in which each decisions counts towards the final outcome are similar to the real-life scenarios in which each decision has a tangible, yet sometimes not immediate, outcome. A representation of these task categories is provided in Fig. 2.4

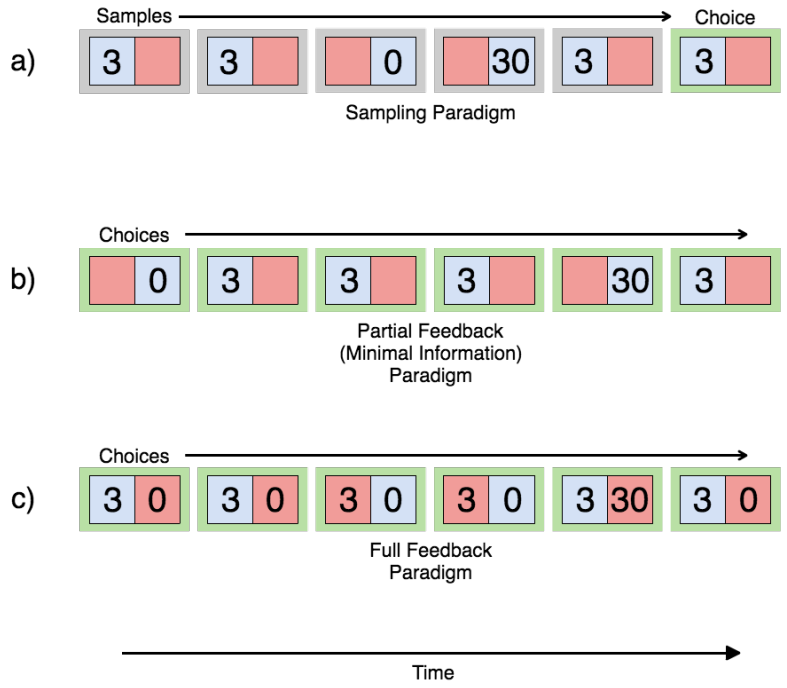

Fig. 2.4 Three different paradigms of choice task information. The outcome of the chosen option is in blue, the forgone in red. Panel a) shows the sampling paradigm. The selections are of sampling nature until the very last which counts as choice. In panel b) the partial feedback paradigm (also called minimal information). Each choice results in an outcome valid in the task. The forgone payoffs are not shown. Panel c) represents the full feedback paradigm, in which every choice is meaningful and all information is provided: the outcomes of both options after each choice.

In order to study human behaviour for tasks resembling the structure of the type of markets in which the equity premium puzzle arises, many works employed small decisions from experience tasks (also called small feedback-based decisions) (e.g. Barron and Erev [2003]; Erev and Barron [2005]; Gneezy and Potters [1997]; Thaler et al. [1997]). One of the main features of these tasks is that they involve repeatedly operating a choice in very similar situations, not much changes from one decision to the other. Another characteristic of these tasks is that the decisions are small and, considered singularly, cannot lead to bankruptcy or losing irreparably in the task. Furthermore, as it is the case for decisions from experience tasks, these decisions are operated based on direct experience and no prior knowledge is

available to the subjects. It follows that to make good decisions the individuals face an exploration-exploitation trade-off for which they need to decide whether to acquire more knowledge about the options or to dedicate their choices to capitalising on their current information. Finally, the decision-makers receive feedback after each choice, either partially in case only the chosen option outcome is provided, or fully in case all options outcomes are presented. The structure of the task and its features make this type of paradigm quite distinct from the one-shot decisions the subjects were asked to do in studies such as Kahneman and Tversky [1979], allowing these studies to be considered a close simulation of real-life decisions.

Some observations can be made about the emerging behaviour in these tasks, in comparison to one-shot decisions from description. Part of the work in Barron and Erev [2003] focuses on some of these. A potential difference is that direct experience and real-time feedback could elicit adaptive behaviour. This adaptive learning process should lead to expected value maximisation behaviour, in accordance with the law of effect (Barron and Erev [2003]; Thorndike [1898]). In contrast, work focused on decisions from description showed consistent deviations from maximisation (Kahneman and Tversky [1979]). This prediction is also supported by findings that subjects were more likely to play a high-expected value gamble if they were allowed to play it repeatedly, instead of only once. It is important to note that there are substantial differences in the number of times a decision-maker is asked to make a choice is much lower for some conditions in these studies (Keren [1991]; Keren and Wagenaar [1987]; Lopes [1981]; Wedell and Böckenholt [1990]).

Another prediction regarding this type of interaction paradigm is that DMs will use the obtained feedback to estimate the underlying payoff distribution for the available choices. This prediction is the basis for all sorts of differences in behaviour deriving from sampling problems, such as undersampling of rare events or large variability in the payoff distributions. Findings from other studies suggest a potential difference in behaviour based on the recency bias, that is a propensity to weigh recent outcomes more heavily than past ones (Camerer and Ho [1999]; Erev and Roth [1998]). The collection of evidence regarding the diverse behaviours exhibited by decision-makers when interacting with the two paradigms is termed description-experience gap (Hau et al. [2008]; Hertwig and Erev [2009]). One of the main differences reported is that, while people tend to overestimate the probability of rare events in decision from description, as shown by the various studies on prospect theory (Kahneman and Tversky [1979]; Tversky and Kahneman [1992]), when DMs are involved in decisions from experience they tend to underestimate the probability of rare events (Barron and Erev [2003]; Benartzi and Thaler [1995]; Erev and Barron [2005]; Hau et al. [2010]; Hertwig et al. [2004]; Hertwig and Erev [2009]; Hertwig and Pleskac [2010]; Jessup et al. [2008]; Newell and

Rakow [2007]; Thaler et al. [1997]). This type of behaviour is also shown in animal studies, such as Real [1991]. One proposed explanation for the gap is that, while in description-based decisions there is no learning, the repeated nature of decisions from experience together with the availability of feedback involves learning, specifically reinforcement learning (Erev and Barron [2005]; Jessup et al. [2008]). Another potential explanation that has been tested in Hertwig et al. [2004] and Hau et al. [2008] is undersampling, which refers to the lack of knowledge of a particular outcome due to its rarity in conjunction with the scarce number of times the associated option was selected. This type of behaviour is related to the observation that people exhibit more sensitivity regarding the frequency of the outcomes instead of their average (Erev et al. [2003]; Estes [1976a,b]; Yechiam and Busemeyer [2006]).

In Hertwig et al. [2004] and Hau et al. [2008], participants were allowed to sample as much as they wanted and only their final preference was registered as the actual choice and then compared to the description-based decision group. Again the gap was found, even if attenuated by larger sample size (Hau et al. [2008]), proving that it cannot be due to the sampling experience alone. The DE gap was also shown empirically in an experimental study concerning choices between risky and ambiguous prospects (Dutt et al. [2013]). Recency has been also identified and studied as potential cause of this behavioural discrepancy. Limited memory capacity could be the reason why even large sampling does not eliminate the gap as it would substantially reintroduce the problem of a biased sample of experiences being considered when making decisions.

Findings in this direction are contrasting, sometimes indicating that the recency bias could be a potential explanation and in other instances showing no supporting evidence of this. In Hertwig et al. [2004] the second half of samples was shown to have predictive power over choices compared to the first half. The same could not be said for the choice data in the study by Hau et al. [2008]. Interestingly enough, in Rakow et al. [2008], recency effects were significant only when the samples were actively collected by the subject and not when these were deriving from passively observing the outcomes. A further common discrepancy that has been highlighted by this research branch is the reversal of risk-related preferences. Specifically, the results for one of the experiments in Barron and Erev [2003] shows a more prominent risk-aversion in the loss domain, while studies on decisions from description indicate a stronger risk-aversion in the gain domain (Kahneman and Tversky [1979]). The loss-aversion phenomenon is another crucial aspect of prospect theory which has been researched and found to be consistently present in both paradigms, as shown by the replication of some conditions from Thaler et al. [1997] in Barron and Erev [2003].

As these conditions are important for the work and analysis offered in this thesis, a more accurate perspective of these learning experiments is provided in Chapter 3. While literature

consistently documented the phenomenon of loss-aversion, it also produced some instances of inconclusive results or even showing evidence in the opposite direction (loss-seeking behaviour). In an elegant experiment by Katz [1964], subjects were asked to make a decision about which of two light bulbs would light up. These were equally likely to do so but the subjects were ignorant to this. One option was deemed safe as it returned either $+1$ or -1 depending on the activation of the bulb. The alternative option was considered risky as the returns, depending on activation, were $+4$ of -4. A loss-averse individual would shy away from the risky light bulb option as it could potentially lead to higher losses compared to the safe option and, according to loss-aversion, these losses would be perceived as more painful. The results showed that participants were indifferent to the choice offering no supporting evidence for the loss-aversion principle. As noted in Erev et al. [2008] the results obtained by Katz, as well as the ones provided by Thaler et al. [1997], could be explained by the alternative hypothesis of diminishing sensitivity. Another possible way to capture Katz's results is with a refined loss-aversion hypothesis involving loss-possibility avoidance (Erev and Barron [2005]).

In an experimental study by Erev et al. [2008], similar conditions were tested providing results contradicting the loss aversion hypothesis and supporting the diminishing sensitivity hypothesis. Specifically, in the conditions involving the prospects "safe" paying off 0 and "risky" being a binary equally likely outcome of $+1000$ or -1000, subjects showed some preference for the risky option. These results are in contrast with the original loss-aversion definition by Kahneman and Tversky [1979] as well as Erev and Barron [2005] loss-probability minimisation theory. The findings in Erev et al. [2008] show that a low sensitivity to the spread of the outcomes can explain the subjects' preference for those options which guarantee a positive return as opposed to riskier prospects which lead to higher average return but with a greater spread between the payoffs. The results also point in the direction of a nominal payoff magnitude effect, according to which subjects exhibit a more marked diminishing in sensitivity for decisions with payoffs in the range of hundreds of points, while this effect was not found when the nominal payoffs were lower.

The dispersion of observed outcomes has also been shown to affect optimal decision-making, this was named "payoff variability effect" (Busemeyer and Townsend [1993]) and was previously researched by Myers and Sadler [1960] with a "card flipping" paradigm. In binary choice tasks, if the option with the highest expected reward has a very variable outcome distribution, decision-makers tend to reduce the proportion of choices for such option. This has been proved to be a robust phenomenon in Haruvy et al. [2001], Erev and Barron [2005] and Erev et al. [2012], but these studies all focus on aggregated population choice data while it would be interesting to investigate different individual experiences in

more detail. In Myers and Sadler [1960] subjects were presented with two cards. The first card would be flipped and reveal a certain outcome (representing a safe option), then the subject could reject this payoff by deciding to reveal the second card's outcome (the risky option). This task was carried out within a mixed description-experience paradigm because the subjects were presented with information about the outcomes obtainable from the first option but not for the second. This experimental setup was associated with partial feedback for the second option: in case a subject accepted the first card's payoff, the forgone payoff for the second card would not be revealed. The safe deck of cards outcomes would be +1 or -1 with even chances for each outcome. The risky deck of cards consisted of outcomes within different ranges depending on the group the subjects were assigned to.

Three groups were denoted with the range of the absolute values and defined according to the following ranges:

- **R(4):** 100 cards evenly divided for each of the integer values between -6 and +6, with exception for -1, 0 and +1;
- **R(9):** 100 cards evenly divided for each of the integer values between -11 and +11, no card for values -1, 0 and 1;
- **R(14):**90 cards evenly divided for each of the integer values between -16 and +16, again with values -1, 0 and 1 not represented.

The procedure consisted of 100 trials where the subject decided whether to accept the safe option or to risk flipping the second deck and being forced to accept the resulting outcome. Generally, subjects exhibited behaviour in the direction of the optimal strategy: to accept the first deck if it presented a +1 and gamble with the second deck in case of a -1. Two sources were found to be significant in explaining the variance of proportion of risk taken: the value (comparison of +1 and -1) and the value combined with range. An increase in deviations from maximisation options was observed for groups with more sparse ranges of outcomes from the risky deck. This was explained with the assumption that subject choices tend to be affected less by the mean of the outcomes and more by the magnitude of the observed outcomes.

Results from Haruvy and Erev [2002] are presented and discussed in Erev and Barron [2005]. Specifically, the following three problems are analysed. These all followed the partial feedback paradigm, and included 14 subjects making 200 decisions.

The options presented were the following:

- **Options in Problem 1**
  - **H**: 11 points for sure
  - **L**: 10 points for sure
- **Options in Problem 2**
  - **H**: 11 points for sure
  - **L**: 19 points or 1 point with even probabilities
- **Options in Problem 3**
  - **H**: 21 points or 1 points with even probabilities
  - **L**: 10 points for sure

The proportion of choices H in the last 100 trials, denoted with Pmax2, has been estimated and used for the comparison. The values are 90%, 71% and 57% respectively in each problem. The comparison between problem 1 and 3 shows a reduced propensity to select option H when this is more variable. At the same time, comparing problem 1 and 2 shows a similar and specular pattern, with the option L being selected more when more variable. An increased exploratory behaviour is proposed as explanation for these result. In those cases where the variability associated with the outcomes is large, exploratory behaviour is not the best course of action. In Erev and Barron [2005] it is suggested that choice behaviour becomes more random as a result of the payoff variability effect, especially if the increased variability is linked to the option yielding higher expected value. More results from similar problems, but with negative outcomes, show that the payoff variability effect emerges in both gain and losses domain. Moreover, the effect is present also in problems within the full feedback paradigm, where subjects could observe the outcome of both chosen and forgone option. The payoff variability effect is present in a single decision task (Busemeyer and Townsend [1993]), but can also be observed in repeated tasks, as shown in Erev and Barron [2005]. These indications are at the base of the study in chapter 4.

Findings by Zion et al. [2010] point in the direction of investors decision-making deriving from adaptive behaviour, confirming the indications in Choi et al. [2009]. Specifically in their experimental setup, subjects are asked to allocate 100 tokens in one of three available funds. Participants were split into groups with different feedback conditions, one in which they received limited information about the sole outcome of the fund they had invested

in, the other group received full feedback regarding all three funds outcomes. The pattern of behaviour resulting from this tasks indicates that people tend to follow experience and feedback in chasing returns, leading to risk-seeking behaviour and under-diversification. In scenarios such as financial markets, similar to the experimental conditions studied in fact, it is possible that adaptive behaviour leads to deviations from maximisation of expected return.

Shifting the reference point: the House-Money effect and Break-Even effect

Mental accounting indicates the mechanisms intrinsically adopted by decision-makers when evaluating choices with associated outcomes and when assessing their wealth (Benartzi and Thaler [1995, 2007]; Kahneman and Tversky [1984]; Thaler et al. [1997]). This process often results in fallacies and inconsistencies. An interesting mental accounting inconsistency phenomenon, in the context of financial decision-making, is the "house-money effect" (Hsu and Chow [2013]; Thaler and Johnson [1990]). This refers to the tendency of investors to seek risk when they obtained positive outcomes in the past. It can be explained by picturing a Las Vegas casino gambler betting a quarter of a dollar in a slot machine and winning $100. The win shifts the benchmark for assessing future gains and losses (as in prospect-theory reference point) and results in a higher propensity to treat the gained $100 as an independent reference point to the gambler's budget, for assessing future prospects and outcomes. The house-money effect produces a behaviour as if the casino gambler does not yet account the quantity gained as their own and considers future decisions to be made using the casino's money (the house-money). There is experimental evidence (Thaler and Johnson [1990]) as well as field study findings (Hsu and Chow [2013]) confirming the existence of this irrational behaviour. From a normative point of view, decision-makers should consider incremental outcomes only when assessing gains and losses (Thaler and Johnson [1990]). Another account on how previous outcomes lead to a change in behaviour is the "break-even effect" (Thaler and Johnson [1990]). This phenomenon, in contrast to the house-money effect, describes the tendency to seek risky options after having incurred previous losses, in the attempt to offset these with potential gains. These two effects have been used in Erev et al. [2008] to explain the potential shift of reference point and subsequently of risk-related behaviour. Respectively, the house-money effect for predicting risk-seeking behaviour following gains and the break-even effect to explain the same type of risk preference after losses. These effects are related to the effect of sunk cost on choice behaviour (Arkes and Blumer [1985]; Thaler [1980]). The sunk cost effect is another mental accounting fallacy affecting decision-makers. Normatively and rationally, past costs should not influence future choices, but it is often the case that DMs factor in these previous expenditures when making

future decisions. Interestingly, this type of fallacy is exhibited only by adult humans and is not exhibited by lower animals or children (Arkes and Ayton [1999]).

From this review it is clear that psychological and economical research are deeply intertwined and the potential for each to influence and inform the other is high and relevant. These pieces of information are combined in this work with the ultimate goal of improving the understanding of how adaptive learning behaviour and decision-making are influenced by each other and by other phenomena.

2.1.3 Neuroscience Perspective

Neuroscience is the field of science which studies the nervous system and the brain in particular. A crucial insight provided by this field of research is that the brain is not a single processor, but the result of the integration of multiple processes, which take place in disparate areas and are activated depending on the type of task and its salience (Brocas and Carrillo [2008]; Loewenstein et al. [2008]). In Bernheim and Rangel [2004], for example, it is proposed that the brain could be modelled as working in one of two modes: "cold" or "hot". Which mode is adopted at any time depends on the current situation and on previous behaviour in similar situations. Another dichotomous modelling of behaviour as a result of competing and collaborating brain functions is suggested in Loewenstein and Donoghue [2005]. Behaviour is the result of the interaction of a "deliberative" and an "affective" system. The affective system is believed to be in control of routine activities; the deliberative system is capable of influencing the affective system by exerting cognitive demanding work, referred to as "willpower" (Loewenstein and Donoghue [2005]; Loewenstein et al. [2008]). Similar considerations are expressed in Brocas and Carrillo [2008] and Benhabib and Bisin [2005]. The former study suggests that when emotional processes have partial information, controlled processes step in and constrain them. The latter study, similarly, proposes that automatic processes are constrained by executive ones, in case these produce suboptimal decisions. This view of the brain as a dual-system can be seen metaphorically as a ship with a naïve ferryman being in control, with the wise captain capable of taking over when needed. This view is also presented in Kahneman's best seller "Thinking, Fast and Slow" (Kahneman [2011]). In this book, which is intended for a general audience, brain function is explained in layman's terms as the result of the interaction and competition of "system 1" and "system 2". System 1 is believed to apply a series of heuristics and simplifying assumptions when making decisions, while system 2 is considered the logical and reasoning counterpart.

While these examples show some researchers are concerned with high-level decision-making other neuroscientists are focused on a lower level of abstraction, trying to understand the structure, the dynamics and the functionality of single brain areas; their aim is to connect

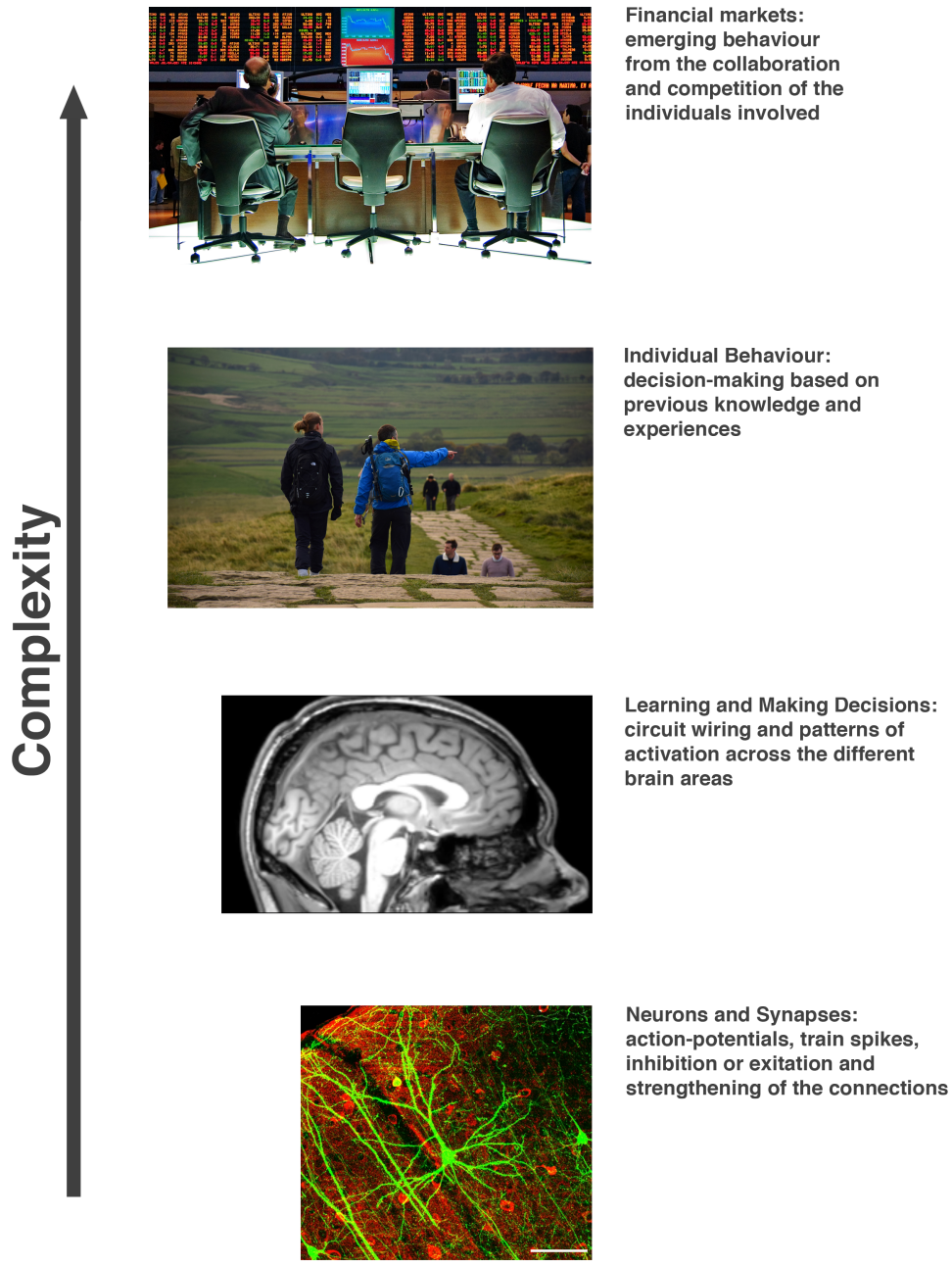

Fig. 2.5 Neuroeconomics abstraction layers. From bottom to top: neurons and synapses (Scale bar: 100 μm) representing the basic informational units; MRI (magnetic resonance imaging) sagittal section of the brain, comprising different structures and communicating areas; individual behaviour, learning and decision-making; market behaviour, the result of many actors interacting and competing. [4]

these pieces together to form a detailed map of how the brain processes information and ultimately makes decisions.

From a biological point of view, the general structure and composition of a brain is well known. The brain of an adult human is composed of roughly 10^{11} neurons, which can be of different types and can be connected to other neurons in different areas. Even if this quantity of neurons can seem impressive, it is definitely not the sheer number that matters. Larger animals have more than double the number of neurons of a human and much smaller animals have far less neurons but exhibit a great deal of competence in learning and decision-making. The way neurons wire together in the brain is believed to play a crucial role in the competence and complexity of behaviour. The map of connections between neurons is called a "connectome". It is believed to be storing the cognitive and computational properties of a brain (Lichtman et al. [2008]; Sporns et al. [2005]). Seung [2009] even suggests the connectome is what defines a person. The branch of neuroscience studying the wiring of neurons is called connectomics and is believed to be of central importance to the understanding of the brain's functioning. A series of discoveries in relevant fields such as molecular biology and electrophysiology together with new imaging methodologies allowed a more refined study of neuronal interactions and connectivity in the last couple of decades (Friston [2011]; Hai et al. [2010]; Lichtman and Sanes [2008]; Minderer et al. [2012]; Perin et al. [2011]; Song et al. [2005]; Wedeen et al. [2012]; Wickersham et al. [2007]; Zhang et al. [2007]). Moreover, increased computational power and accessibility have provided support for computer-based simulations in this branch. This timely combination of findings and enhanced machine capabilities created a vibrant research scenario which promoted the development of some ambitious projects. For example the mapping of the human connectome in a dataset of healthy adults (The Human Connectome Project: Toga et al. [2012]) or the simulation of human brain activity on high-performance computers (The Human Brain Project: Frackowiak and Markram [2015]).

The idea that neurons involved in a certain function cluster together and operate as a single processor comes from connectionism theory (Rumelhart et al. [1987]; Sejnowski and Rosenberg [1987]; Thorndike [1898]). These brain areas are characterised by strong connectivity patterns between the neurons and present highly specialised competencies. Neuronal interactions of different brain areas give rise to the observable behaviour. The formation of such connected structures is therefore not random but is created and strengthened

[4]First image: CC BY 2.0 Rafael Matsunaga [https://www.flickr.com/photos/78629042@N00/479370088] (modified: cropped), last image: CC BY 2.5 Wei-Chung Allen Lee, Hayden Huang, Guoping Feng, Joshua R. Sanes, Emery N. Brown, Peter T. So, Elly Nedivi [Dynamic Remodeling of Dendritic Arbors in GABAergic Interneurons of Adult Visual Cortex. Lee WCA, Huang H, Feng G, Sanes JR, Brown EN, et al. PLoS Biology Vol. 4, No. 2, e29. doi:10.1371/journal.pbio.0040029, Figure 6f, (modified: plus scalebar, minus letter "f".)]

over time. This process is one of the most important and peculiar a brain can achieve and is what characterises intelligent animals: learning. The retention of information is achieved in the brain by strengthening the connections among neurons, the synapses. The number of neurons in a human adult brain is estimated to be 86 ± 8 billions of cells (Azevedo et al. [2009]). The number of synapses is even higher, with estimates in the order of 10^{15} (a quadrillion) synapses in an adult brain; their number slowly declining with age (Drachman [2005]). These connections possess a property called synaptic plasticity which leads to an improvement or a degradation of the connection between two neurons depending on their activity (Hebb [1949]; Thomson [2000]). There are two types of synaptic plasticity which can happen between neurons. They are referred to as short-term and long-term plasticity and are different, not only for their duration as the names suggests, but also for the underlying biological and chemical processes happening (Fioravante and Regehr [2011]; Kullmann and Lamsa [2007]).

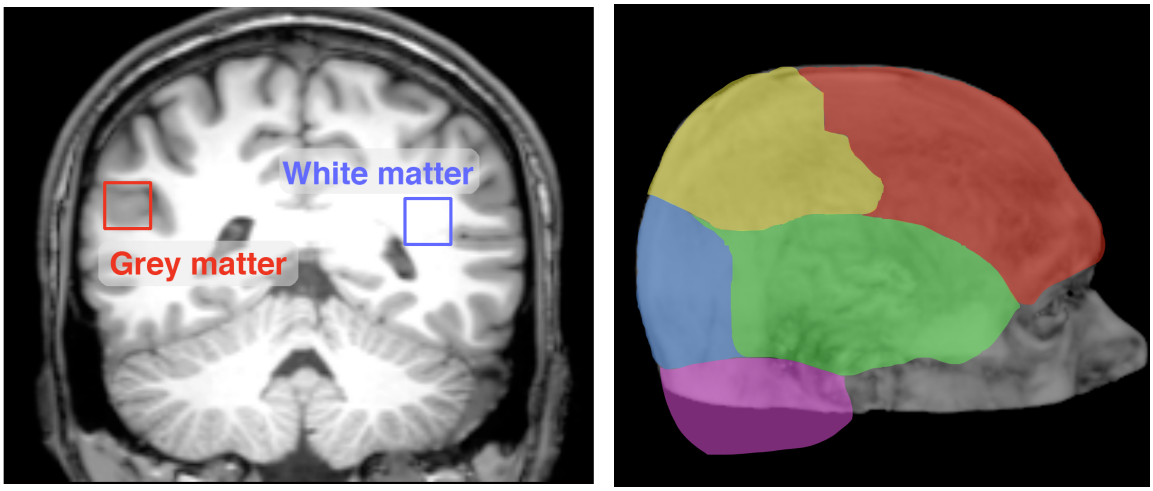

Fig. 2.6 Panel a) MRI coronal section of the human brain (of the author). Grey and white matter are indicated by red and blue squares respectively. Panel b) Human brain MRI sagittal external view. The colours represent areas of the cerebral cortex: frontal lobe in red, parietal lobe in yellow, occipital lobe in blue, temporal lobe in green and cerebellum in magenta.

Synaptic plasticity is at the base of memory formation and the learning process. The study of the microscopic changes at a neuronal level characterises the lower levels of knowledge in the neuroscience field. Modifications at this level result in observable adaptive behaviour at higher level, in both animals and humans. Animals demonstrate learning within simple tasks while humans can learn much more complex functions. Communication includes understanding and producing language; while humans mastered this task thousands of years ago, animals are only capable of low-complexity information transmission.

Neurons are clustered and organised within areas of the brain presenting detailed connections among them. The modular nature of the brain is considered to be a well-established fact in literature (Brocas and Carrillo [2008]). Neuroscientific research is interested in determining which areas of the brain are involved in specific functions and how these "talk" to each other to achieve more complex functions. Many of these structures are located in the cerebral cortex. This is the outer layer of the brain and consists of a sheet of neural tissue of about 2 millimetres of thickness which covers the cerebrum. The cortex is composed of grey matter, which contains neural cell bodies, synapses, capillaries. The grey matter also presents glial cells, which are non-neuronal cells supplying nutrients and oxygen to the neurons, also providing them structural support and insulation. The underlying white matter is composed of myelinated axons, also known as nerve tracts. White matter is devoid of neuronal cell bodies or dendrites.

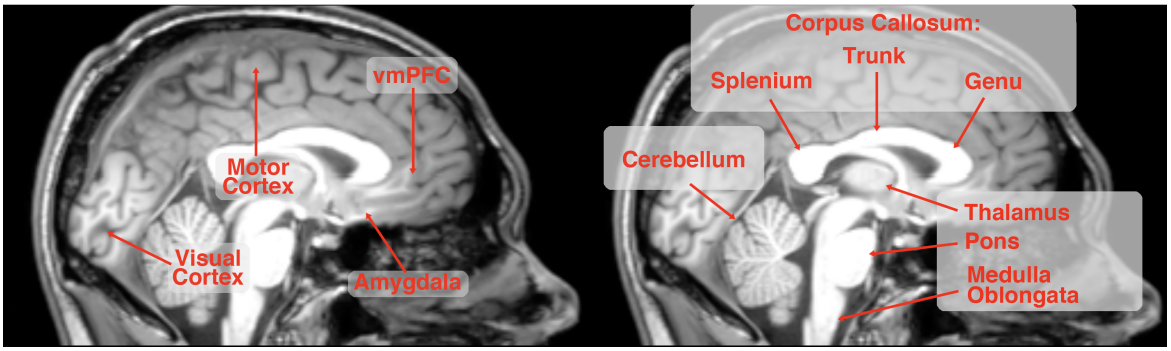

Fig. 2.7 Panel a) MRI sagittal section, medial view of the human brain (of the author), with the location of the Motor Cortex, Visual Cortex, Amygdala and ventro-medial Pre-Frontal Cortex. Panel b) Cerebellum, the components of the Corpus Callosum along with the Thalamus, the Pons and the Medulla Oblongata indicated by the arrows.

The cortex topology is subdivided into regions which are often associated with different functions. The main subdivision is portrayed in Fig. 2.6, panel b) and consists of frontal lobe, parietal lobe, temporal lobe, occipital lobe and cerebellum. The visual cortex is located in the occipital lobe of the brain, at the back of the head and above the cerebellum, which is located at the insertion of the neck with the skull. This area of the cortex receives the visual information from the eyes. Before getting to the visual cortex, this signal passes through the thalamus which is one of the innermost regions of the brain.

The thalamus is located near the centre of the brain and is composed of grey matter although not being part of the cortex. It is divided in two symmetrical halves and is thought to be tightly involved in forwarding sensory-motor signals to the cortex (Sherman and Guillery [2001]). The neuroconnective structures supporting the transmission of signals between brain regions are called neural pathways and consist of tracts, which are axon bundles (Moore et al.

[2013]). The thalamus is also involved in other systems, such as somatosensory and auditory. The thalamus is not only a relay for information propagation to the cortex but it also involved in a two-way pathway to the cortex (Sherman and Guillery [2006]; Sherman [2007]). Other regions of the cortex and the brain in general are indicated in Fig. 2.7.

Among the various brain regions, the basal ganglia are of particular interest for behavioural research. The basal ganglia are a group of subcortical nuclei located near the thalamus; they consist of the striatum, the subthalamic nucleus, the globus pallidus, and the substantia nigra. These nuclei are deeply interconnected with other brain regions and are part of the cortico-basal ganglia-thalamo-cortical loop (CBGTC loop). In dysfunctional basal ganglia, this pathway is thought to be involved in movement diseases such as Parkinson's disease or Huntington's disease (Cameron et al. [2010]; Mahlon and Thomas [2009]; Miller et al. [2008]; Reiner et al. [1988]; Stocco et al. [2011]). The CBGTC loop is also of interest in conditions such as schizophrenia, obsessive-compulsive disorder, attention-deficit disorder and different types of addiction (Inta et al. [2011]; Redgrave and Gurney [2006]; Redgrave et al. [1999]).

In healthy subjects, the basal ganglia are involved in motor-control, learning and decision-making (Bogacz and Gurney [2007]; Bogacz and Larsen [2011]; Redgrave et al. [1999]; Wickens et al. [2007]). The main input comes from the striatum (STR) and the subthalamic nucleus (STN), while the main output locations are the substantia nigra pars reticulata (SNr) and the internal portion of the globus pallidus (GPi). The external globus pallidus (GPe) instead, represents an internal signal hub, with connections to and from the subthalamic nucleus and the striatum. Modulatory signals in striatum and other regions of the basal ganglia are provided by the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA). These areas are composed of dopaminergic (DA) neurons, which are involved in the release of the neurotransmitter dopamine; these regions represent part of the mesolimbic (or reward) pathway (Sulzer [2005]). This pathway, similarly to the CBGTC loop, is involved in behavioural processes, such as learning and motivation. Dopaminergic responses have been associated with the reward prediction error: the discrepancy in the expected reward of an action and the actual outcome of that action (Montague et al. [1996, 2004]; Redgrave and Gurney [2006]). Highly addictive substances, such as cocaine, affect this pathway by hijacking the dopamine neurons (Bernheim and Rangel [2004]; Montague et al. [2004]). The substance pushes the DA neurons to signal a positive reward when taking the drug, even if the actual consequences of doing so are adverse to the individual (Bernheim and Rangel [2004]; Everitt and Wolf [2002]; Schultz et al. [1997]). Other areas that have been shown to be activated in relation to addiction are the amygdala, the orbito-frontal cortex (OFC), the anterior cingulate cortex (ACg) as well as the striatum and the dorsolateral PFC (Everitt

and Wolf [2002]; Grant et al. [1996]; Maas et al. [1998]; Schoenbaum and Setlow [2005]; Wang et al. [1999]). Multiple fMRI studies of decision-making tasks in uncertain scenarios, show that insular cortex (INS) and the ventrolateral pre-frontal (vlPFC) are activated when decision-makers face adverse stimuli, which can be either decisions resulting in punishment or involving increased risk (Huettel et al. [2005]; Paulus et al. [2003]; Platt and Huettel [2008]; Sanfey et al. [2003]).

DA neurons functions are of central importance in cognition and motor control; they are also known to be involved in psychiatric disorders as well as neurodegenerative diseases (Montague et al. [2004]; Van den Heuvel and Pasterkamp [2008]). Dopaminergic neurons also play a role in the value individuals attribute to money or personal relationships (Montague et al. [2004]). Work by Schultz provided further evidence that these neurons are involved in the encoding of outcome values: they are activated in case the observed outcome is better than predicted, they are depressed in case the outcome is worse than anticipated and remain inactivated in case of the event being as good as previously thought (Schultz [1998]). This is confirmed by the activation of ventral striatum and medial pre-frontal cortex (mPFC) as shown in several fMRI studies (Delgado et al. [2000]; Kable and Glimcher [2007]; Knutson et al. [2005]; Kuhnen and Knutson [2005]; Platt and Huettel [2008]; Tom et al. [2007]).

Because the systems presented so far are involved in decision-making as well as reward-based learning, they have been commonly formalised with the help of the reinforcement learning (RL) framework (Montague et al. [2004]). This is a set of theories and rules used to model decision-making problems and to provide a solution based on a trial-and-error approach, therefore representing a logical approach to model the neuromodulatory systems and their functioning within the brain. Reinforcement learning is adopted when modelling behavioural scenarios, instead of a simpler methods such as Rescorla-Wagner (RW) models, because RL provides a more sophisticated approach. Animals tackle decision-making tasks following procedures which depend on the state they consider themselves to be in and RL allows to capture this belief in a handy way, while a RW would simplistically focus on the associations between stimulus and response, as portrayed in eq: 2.8 (Montague et al. [2004]). Because of its powerful descriptiveness and since the RL framework is widely adopted in literature, the work of this thesis will adopt it as a descriptive model for the study of behavioural data. Therefore the reinforcement learning framework will be presented in the next section in greater detail.

This brief introduction to neuroscience research provides some examples of how the discoveries made in this field along with the models adopted by the researchers are of great interest for psychology, economics and the study of learning and decision-making as a unifying field of research.

2.2 Reinforcement Learning

In machine learning and generally in control theory the final objective is to train an automatic agent to produce some kind of behaviour. Often this goal is achieved by encoding an algorithm which uses a set of examples to generalise some rule which is then applied to new, previously unknown instances. A classic example is the email spam filter, a type of supervised learning mechanism. This algorithm is trained on a binary classification task by the user. The user provides the algorithm with positive and negative examples leading, eventually, to the correct classification of new emails as spam or legitimate. Reinforcement learning (RL) instead, is not based on this type of training by examples. The environment gives a numerical signal and the agent processes this information to improve future behaviour. Generalising this idea, any agent which is cast into an unknown environment and needs to learn by trial and error the mapping of what to do in which situation, is a RL agent. This definition immediately exposes the connection of these ideas to Thorndike [1898] experiments, in which a cat was placed into a shutterbox and had to learn to escape by trying different actions. It is evident how such a general approach to learning can be used to describe animal or human behaviour. As previously noted, modelling behaviour or neural mechanisms with models based on temporal difference (TD) error has been proved to be of great value (Daw et al. [2005, 2006]; Doya [2000, 2007]; Hollerman and Schultz [1998]; Houk and Wise [1995]; Joel et al. [2002]; Montague et al. [1996, 2004]; Schultz [1998]; Schultz et al. [1997]). There are several indications that TD learning is implemented in particular brain areas. A TD error signal is, in fact, consistent with the activation of dopaminergic neurons in the striatum and orbitofrontal cortex (Cohen et al. [2007]; Joel et al. [2002]; Lohrenz et al. [2007]; O'Doherty et al. [2003]). This powerful framework has been used extensively to model and comprehend behaviour and the neural processes involved, therefore becoming of pivotal importance in the decision-making research (Dayan and Daw [2008]).

In order to fully understand the learning algorithm mechanisms and how they capture neural computation or emergent behaviour, it is useful to provide some context and formally define the foundations on which this powerful technique is grounded. The following sections are organised so that the general concepts are linked to the actual techniques used to implement them and the mathematical background needed by the framework. Much of this material follows the perspectives from Sutton and Barto [1998], reorganising the information to best fit this thesis.

Reinforcement learning is ultimately a method of learning a mapping of which actions to take in specific situations, or goal-directed learning. The final goal is to maximise the numerical reward obtained from the environment. The learning agent learns which actions are good or bad, by inferring this from its interactions with the environment. The quality of

an action is encoded in the signal the environment gives back to the agent after an action. In some tasks actions not only elicit rewards but also affect the future state of the agent within the environment, in turn modifying the future attainable reward. As the RL agent is not informed or instructed about the correct actions to take, it is required to explore and evaluate the outcome of the exploration. These two functions represent the key components of the learning procedure: action selection and belief update. One of the simplest settings in which the agent can find itself to be is a two choice problem, also known as two-armed bandit. An armed bandit is a slot machine, a n-armed bandit problem is a task in which a decision-maker needs to decide which slot machine to play (which lever/arm to pull). In a canonical scenario, a gambler enters a casino and faces two slot machines. The task is to decide which of the two bandits offers the best return. This setting includes two possible actions in a stateless environment. More complex versions of the armed bandit problem have been developed and studied. In Vernade et al. [2017] for example, a stochastic version of the multi-armed bandit (MAB) problem is employed in the context of internet advertisement delayed conversion. MABs are tackled using algorithms based on the Upper Confidence Boundary (UCB) and Kullback-Leibler UCB (KL-UCB) frameworks. The UCB algorithm has robust performance within the class of limited stochastic rewards problems (Cappè et al. [2013]). These algorithms are based on minimising a quantity called the "regret", defined as the difference between the sum of the rewards for the chosen options and the sum of rewards had the agent known the best option (Cappè et al. [2013]; Garivier and Cappè [2011]). The exploration-exploitation dilemma arises in this setup too. The regret can increase due to either greedily choosing the current best action known, possibly missing better options, or spending too much time collecting information about the reward distributions associated to the actions. Treatment selection in medicine is an application of such problem formulations and proposed solutions (Cappè et al. [2013]). For example, it is not known beforehand which of the clinical trials of new drugs will lead to the best outcome for the patients. The patients are allocated sequentially to each treatment and observations about the performance of such treatment are taken, with the final goal of identifying the course that achieves the most favourable outcomes.

In more complex problems, the environment is composed of more states and the outcome of an action might change the state in which the agent is. A commonly used example is the grid-world (Fig. 2.8).

Here the agent can move on a grid of adjacent states and each action can result in a reward signal and a deterministic (or stochastic) movement to another state. To better understand this learning procedure it is useful to discretise time and introduce the components of this system. The agent is the decision-maker in the task (also known as the learner). Everything

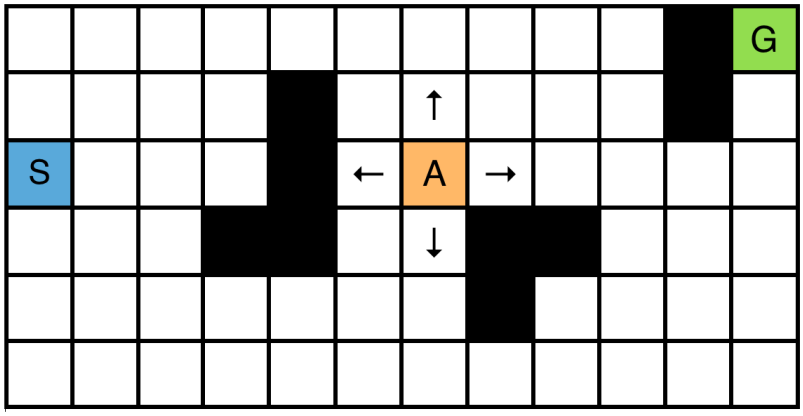

Fig. 2.8 An example of grid-world, A represent the agent, each square represents a state where S is the start state and G is the end goal state. At each point in time the available actions are the four cardinal directions (N,W,S,E).

which is not internal to the agent is the environment. The interaction between agent and environment can be described as a cycle (Fig. 2.9). At time t the environment informs the agent about which state s_t it is in, the agent takes an action a_t. In return the environment produces a reward r_{t+1} and updates the state s_{t+1} of the agent.

Markov property

Formally, for trials $t = 1, 2, \cdots T$, state $s_t \in S$ where S is the set of all possible states. The available action $a_t \in A(s_t)$ where $A(s_t)$ is the set of actions available in state s_t. After taking an action, the agent receives a real valued signal called the reward $r_{t+1} \in \mathbb{R}$ and moves to state s_{t+1}. This is a powerful abstraction that allows for many different tasks to be modelled and tackled by RL agents. States, specifically, are a compelling concept of this framework. They allow the agent to receive knowledge about its surroundings by encapsulating information which can be more than basic perceptions. At the same time, a state should not inform an agent about everything. It would be cheating, and ultimately counterproductive to provide an agent with information about the unobservable future information. This would make the agent an oracle and the task trivial and uninteresting. To make decisions incrementally in a task, a RL agent must possess information about the environment that is necessary and sufficient. What is required from the state is to represent the entire history of previous interactions and nothing more. A state satisfying this property is said to be markovian. The environment needs to possess the Markov property to allow this. This property is formally

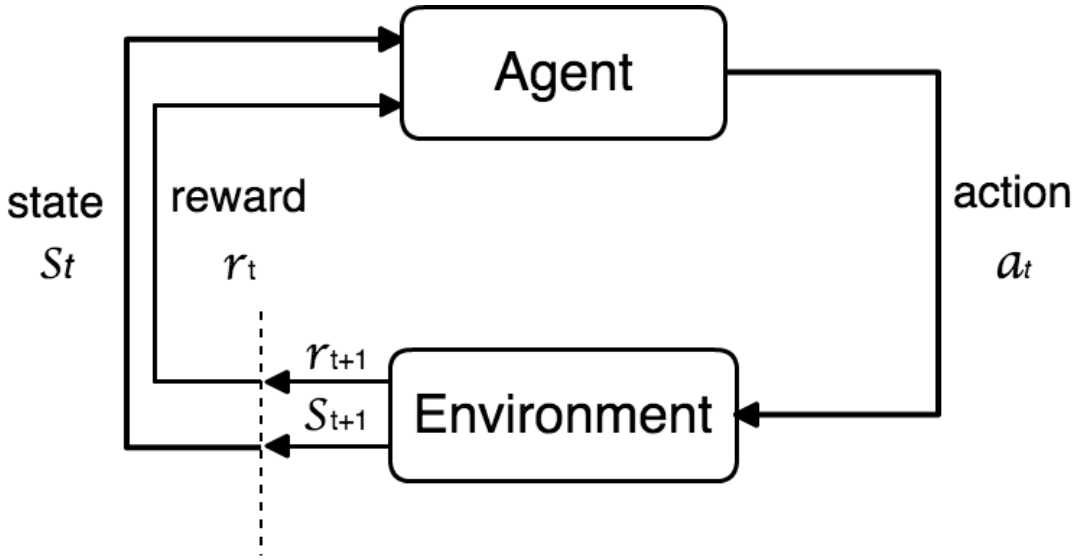

Fig. 2.9 The agent and environment interaction scheme

defined as follows. For a general environment, the state and reward at time $t+1$ depend on all previous history:

$$P(s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \cdots, r_1, s_0, a_0) \tag{2.9}$$

for s', r and all possible previous values of states, actions and rewards. A state is said to be markovian if:

$$P(s_{t+1} = s', r_{t+1} = r | s_t, a_t) \tag{2.10}$$

For all s',r,s_t and a_t. Meaning that the response of the environment at time $t+1$ depends only on the previous state and action. A state is markovian if the equations 2.9 and 2.10 are equal. If this holds true for all states $s_t \in S$ and actions $a_t \in A(s_t)$, the entire environment can be considered markovian. The Markov property allows for prediction of future state and expected reward based on the current state and action. It is possible indeed, to determine the best policy for picking an action in a Markov state in the same way that it would be possible to determine the best policy by knowing the entire history (Sutton and Barto [1998]). This setup for the states is readily representable with a Markov decision process (MDP). A MDP is generally defined as a stochastic control process in discrete-time satisfying the Markov property. The set of states, the set of actions and the one-step interactions with the environment define the MDP for a task. The interacting structure of a MDP is defined by two quantities. Considering a state-action pair (s,a), the probability of the next state s' is:

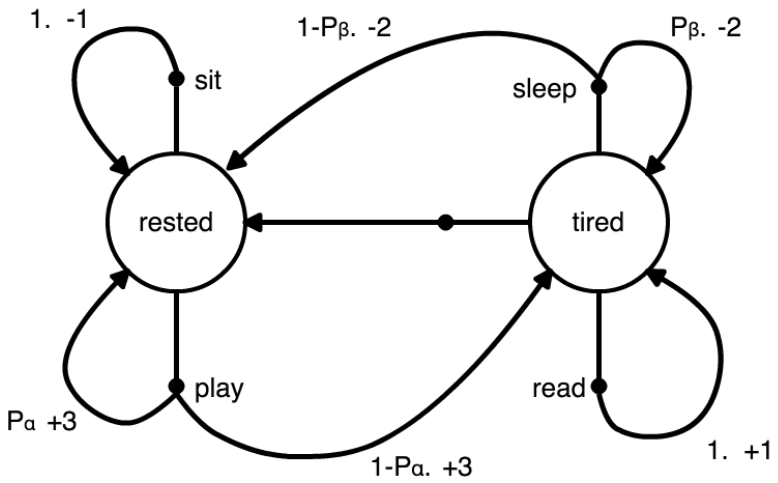

Fig. 2.10 An example of transition graph for a stochastic MDP. In this simple example an agent needs to learn which action maximises the reward in each state but in certain cases the action can lead to a change in state according to some probability (actions search have multiple output arrows).

$$P^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \tag{2.11}$$

This quantity represent the transition probability of going from state s to state s' via action a. Given the triplet defined by state s, action a and next state s', the expected value of the reward is:

$$R^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \tag{2.12}$$

A useful graphical representation of an MDP is a transition graph which pictures states as big hollow circles, and actions with small filled circles. These nodes are labelled respectively with the states names and actions names. The dynamics of the MDP are depicted by the arrows, which are labelled with the reward and can be labelled with the probabilities in case of a stochastic environment.

Discounting

A task can be episodic or continuing, meaning that the discrete time interactions the agent has with the environment can be finite or not. In a finite task the goal is to maximise the expected return, which is usually defined as the sum of the rewards or a function of it. For example:

$$R_t = r_{t+1} + r_{t+2} + \cdots + r_T \tag{2.13}$$

with T being the time step of the final interaction. For a continuing task instead, $T = \infty$ and a simple scenario in which the agent obtains a positive reward at each time step will lead to an unbound return which cannot be maximised. To obviate to this, the concept of discounting is introduced. This consists in multiplying each reward at each time step by a discount rate:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.14}$$

with γ being a parameter that can take values in the range $(0,1)$. For $\gamma = 0$ the agent is said to be myopic because at time t it will pick the action a_t trying to maximise the return $R_t = r_{t+1}$ (all terms after the first are multiplied by $\gamma = 0$). For such value of the discount factor the agent is trying to maximise immediate rewards only. As $\gamma \to 1$ the agent becomes more far-sighted and is concerned with maximising rewards in the long term. Discounting encapsulates a well documented human behaviour which violates the rationality principles of rational theories such as Expected Utility Theory (Frederick et al. [2002]; Sanfey et al. [2006]; Thaler [1981]). In many instances individuals show a preference for immediate or short term rewards as opposed to delayed rewards. For example, individuals would consider that a dollar today is worth more than a dollar in a year time, or in another formulation they would choose to receive 10 dollars today instead of 11 tomorrow but at the same time they would be happy to wait one more day when deciding between 10 dollars in a year or 11 in a year and a day. There are two main types of discounting that have been studied in literature: exponential and hyperbolic (Fig. 2.11). Hyperbolic discounting has been commonly presented as a more realistic alternative to exponential discounting (Azfar [1999]; Haith et al. [2012]; Kobayashi and Schultz [2008]) but it has also been found to fail in some cases (Luhmann [2013]). Exponential discounting has been preferred in temporal difference learning methods as it can be conveniently expressed with a recursive formulation (Alexander and Brown [2010]). Discounting is believed to play a key role in impulsive behaviour which can lead to addiction (Ahmed and Gutkin [2011]; Bickel and Marsch [2001]; Kurth-Nelson and Redish [2010]; Story et al. [2014]).

Mapping states to actions: the policy

The problem for the agent is then to learn the probability of picking each of the available actions, a probabilistic mapping of states to actions. This experience-based learning is

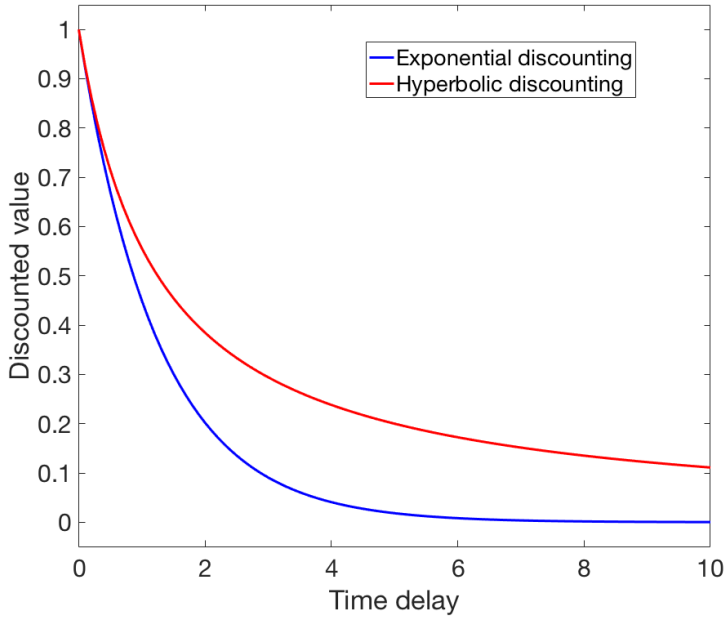

Fig. 2.11 Exponential vs hyperbolic discounting

captured by the policy, denoted with π_t. Formally π is the mapping of each state $s \in S$ and action $a \in A(s)$ to the probability of picking a when the agent is in s, denoted $\pi(s,a)$. The learning problem is for the agent to modify this policy so to maximise the profit in the long term. In order to assess how good a policy is, the RL algorithms first evaluate how good a particular situation is for the agent. This is done with a value function. A value function V^π represents the expected return obtainable following policy π:

$$V^\pi(s) = E_\pi\{R_t|s_t = s\} = E_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\middle|s_t = s\right\} \tag{2.15}$$

with E_π denoting the expected value following policy π. While the value function concerns states values, it is also possible to define an action-value function to evaluate state-action pairs under a policy:

$$Q^\pi(s,a) = E_\pi\{R_t|s_t = s, a_t = a\} = E_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\middle|s_t = s, a_t = a\right\} \tag{2.16}$$

The key feature of these value functions is that they possess a peculiar recursive property:

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\{R_t | s_t = s\} \\ &= E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s \right\} \\ &= E_{\pi}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s \right\} \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_{t+1} = s' \right\} \right] \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^{\pi}(s') \right] \end{aligned} \tag{2.17}$$

The last equation is the Bellman equation and it shows the recursive connection between a state and its successors. It can be used to evaluate a policy or extended to represent the optimal value function. If the expected return under policy π, $V^{\pi}(s)$ is greater or equal compared to another policy π' for all states $s \in S$, then policy π is better than or equal to policy π':

$$V^{\pi}(s) \geq V^{\pi'}(s) \implies \pi \geq \pi' \tag{2.18}$$

Having defined an order of policies based on the expected return, it is possible to state that there is at least one best policy, defined as π^*.

$$V^*(s) = \max_{\pi} V^{\pi}(s) \tag{2.19}$$

where V^* is the optimal state-value function. There is always one optimal policy which is better or equal to all other policies. An action-value function Q is associated to each policy π and, in the specific, each optimal policy π^* is associated to an optimal action-value function Q^* defined as:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) = E\{r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\} \tag{2.20}$$

for each state $s \in S$ and for each action $a \in A(s)$.

Now it is possible to reformulate the Bellman equation in terms of optimal value function V^*, also called the Bellman optimality equation:

$$\begin{aligned} V^*(s) &= \max_{a \in A(s)} Q^{\pi^*}(s,a) \\ &= \max_a E_{\pi^*} \{R_t | s_t = s, a_t = a\} \\ &= \max_a E_{\pi^*} \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, a_t = a \right\} \\ &= \max_a E_{\pi^*} \left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s, a_t = a \right\} \\ &= \max_a E \{r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\} \\ &= \max_{a \in A(s)} \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^*(s')] \end{aligned} \tag{2.21}$$

The corresponding Bellman optimality equation for the action-value function can be expressed as:

$$\begin{aligned} Q^*(s,a) &= E \left\{ r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a) \middle| s_t = s, a_t = a \right\} \\ &= \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma \max_{a'} Q^*(s', a')] \end{aligned} \tag{2.22}$$

It follows that, for a given MDP with known transition probabilities and rewards functions (respectively $P^a_{ss'}$ and $R^a_{ss'}$), it is possible to solve the Bellman optimality equation. Once the V^* is known, then it suffices to decide greedily with respect to such optimal value function to find an optimal policy π^*. In a similar fashion, having Q^* allows the agent to search over the available actions in each state, instead of having to do a look-ahead to find the next best state, using the optimal value function V^*. As state-action pairs encode information about the future attainable rewards from the next time-step, the optimal action-value function Q^* is more efficient in that it allows for a compact search of best action at each time-step. Of course, solving the Bellman optimality equation is quite demanding as it is a system of equations directly dependent on the number of states in the MDP. Moreover, to solve the equation, the dynamics of the environment ($P^a_{ss'}$ and $R^a_{ss'}$) need to be known and this is not always possible. Taking as an example the game of chess, each configuration of the pieces on the board represents a state and it is evident that the dimensionality of the state-space makes it computationally intractable to solve the Bellman optimality equation for this task.

2.2.1 Model Free

Considering those cases in which the MDP is not completely known - that is either or both the transition probability function and the reward function are unknown - it is impossible to directly solve the Bellman optimality equation. In this type of scenario, the agent can still achieve satisfactory learning in the environment via sampling. Incrementally updating its internal belief of which states are best to visit and which actions are better to take in each state. Model free solutions are characterised by the agent not trying to learn the environment and its transition probabilities but only which action is the best to take in specific states. One of the most important breakthroughs in the theory of reinforcement learning, which makes the whole framework feasibly applicable to interesting problems, is the concept of temporal difference (TD) learning.

TD Learning

The key concept at the basis of temporal-difference learning is the TD prediction error. The prediction error refers to the quantity calculated as the difference between the predicted reward and the actual observed reward for a specific state (or state-action pair). Instead of waiting until the end of an episode (which sometimes is not feasible, as in continuing problems), the agent updates its knowledge after each step. This class of learning models are well suited to describe human learning because of the similarity between the concept of prediction-error and the dopaminergic neuron activations in the brain (Daw [2003]; Daw et al. [2005, 2006]; Doya [2000, 2007]; Houk and Wise [1995]; Joel et al. [2002]; Montague et al. [1996]; O'Doherty et al. [2007]; Schultz et al. [1997]; Suri and Schultz [1998]). TD learning is often compared with the Rescorla-Wagner model with the the notable difference that the latter is concerned with single time-steps and not the entire course of interaction between agent and environment (O'Doherty et al. [2003]; Schultz [1998]). Moreover, TD learning allowed for studies involving both behavioural data and neurophysiological data like functional magnetic resonance imaging data (fMRI) (Niv et al. [2012]; O'Doherty et al. [2003, 2007]).

Using a small grid-world (Fig. 2.8) as example, a RL agent aiming at learning the action-value function Q^{π}, will store the state values in a look-up table and update their values after each time it moves. For instance, let's consider an agent randomly moving and with initial values for all state-action pairs equal to 0, $V(s) = 0 \forall s \in S$. Assuming each state yields reward 0 except for the top-right state being the goal state with reward 1 and the terminal state finalising the episode. Starting from bottom-left corner of the grid-world and taking

action right, leads to a reward of 0. At this point the agent will operate the following update rule:

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \tag{2.23}$$

where α is the step-size parameter which tunes the rate of learning for the agent, r_{t+1} is the reward obtained in the last movement, γ is the discount factor and the values $V(s_t)$ and $V(s_{t+1})$ are the value of the state from which the agent started and the value of the state in which it arrived. The quantity $r_{t+1} + \gamma V(s_{t+1})$ is called the target of the update and represents the knowledge acquired after the action the agent took. The difference between the target and the value of $V(s_t)$ is the TD prediction error and guides the learning towards the actual state value. When the agent is one step away from the terminal state and moves right, finally landing in the goal state, the update rule will move the value of the state just before the goal $V(s)$ slightly towards its actual value. Commonly for terminal states it is assumed that the future reward $V(s_{t+1}) = 0$. This method is said to be a bootstrapping method because it updates estimates of the value for each state based on previous estimates. In fact, during the next episode, the agent will at some point move to the state just before the goal and will learn that this state has a higher value leading to a back-propagation of the reward along the path which lead to it. It is important to note that the values used in the update, within the target quantity, are estimates of the expected value of a state $V(s_t)$ and not the actual value under the current policy $V^{\pi}(s_t)$. As this method does not require the agent to wait until the end of its interactions during an episode, it is considered an on-line method. While the value of a state is denoted $V(s)$, the value of a state-action pair is called Q-value and is denoted $Q(s,a)$. This last quantity expresses the value of taking action a when in state s. The TD learning update rule represent only the first half of the learning process. It is limited to improving the knowledge of the value of each state (or state-action pair) but in the previous example the agent was not allowed to make decisions about which action to take, as it was assumed to move randomly. Now that the agent has a way of determining which state is better than the others it should be allowed to decide in which direction to move, what was previously defined as the policy π.

Action-Selection Policies

In the previous example the agent was assumed to be using a random strategy to decide where to move at each time-step. Of course this is a poor strategy and other, more or less refined, strategies can be taken into account. Supposing the correct value-function is known, as mentioned at the end of section 2.2, the best strategy is to select what to do in a greedy

fashion, that is always picking the action with the highest expected return. When the agent does not posses full knowledge about the actual values of the states, it is best to include a certain degree of exploration in the action-selection in order to acquire better information. To explore, in this context, means to follow a course of actions which does not follow the known best outcome. This is known as the exploitation-exploration trade-off. One of the possible ways to address this problem is to act greedily most of the time but devote a certain amount of time (i.e. interactions) to exploring the environment. Epsilon-greedy is an example of this type of exploration: it consists in picking the best action with probability $P = 1 - \varepsilon$ and exploring in the complementary $P = \varepsilon$ cases. As an example, if $\varepsilon = 0.1$, the agent will follow the best known strategy 90% of the time and explore the remaining 10%. This is not very sophisticated as it does not allow for a sensible search through the available actions but randomly selects one of them. In the previous grid-world example, if in a particular state the best action is believed to be north, it might be worth exploring east instead of south or west. Soft-max action-selection allows for a ranking of the probability of picking each action based on their values. To achieve this more refined exploratory strategy it is possible to make use of the Boltzmann distribution:

$$P(east) = \frac{\exp(Q_t(east) \cdot \beta)}{\sum_a^4 \exp(Q_t(a) \cdot \beta)} \tag{2.24}$$

where $P(east)$ is the probability of picking action east at time t, $Q_t(east)$ is the value of action *east* in the current state at time t, the denominator represents the sum over all the possible actions and β is a positive free parameter referred to as inverse temperature and represents the greediness of the agent. High values of β (low temperature) lead to a wider spread between the action values resulting in higher probability for the best action to be selected. Low values of β (high temperature) indicate actions are equiprobable which leads to exploratory behaviour. In the limit $\beta \to 0$ the agent action selection becomes random. A visualisation of the impact of beta on the probabilities is shown in Fig. 2.12. In the figure, the best action is North (Q-value = 0.9), while the second best is East (Q = 0.8). For increasing values of β the best action becomes the most likely to be picked while the others decrease, the second best action is still more probable than the others to get selected. Exploration is implemented in this way and is regulated by the parameter β, when $\beta = 0$ the actions have the same probability of being taken ($\frac{1}{4}$) and the agent effectively follows a random policy.

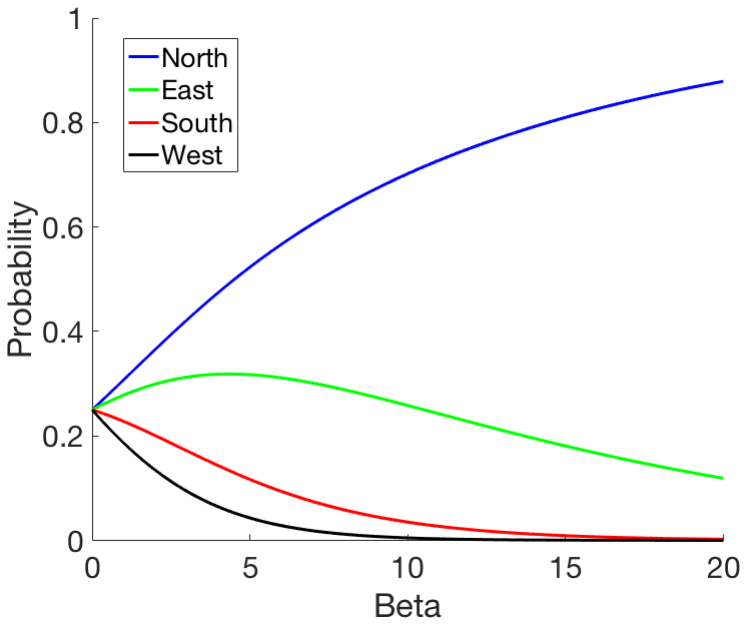

Fig. 2.12 Soft-Max graphical example: the value of the probability of an action being picked is a function of the value of the inverse temperature free parameter β. The higher the value of β becomes, the more greedy the policy becomes. Q-values of the actions: $North = 0.9, East = 0.8, South = 0.6, East = 0.4$.

Q-Learning

One of the most adopted models which implements TD-learning is Q-Learning (Watkins [1989], Watkins and Dayan [1992]). The learning model is based on the following update rule:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)] \tag{2.25}$$

where α is the learning rate and γ is the discount factor. The difference with TD-learning is that this update rule uses as target the best achievable reward ($\max_a Q(s_{t+1})$) from the next state instead of an arbitrary estimate. In other words this update rule implements a greedy action selection for the next step look-ahead. Because of this greedy update, which is likely to be different from the one currently used by the agent, Q-learning is considered an off-policy method. This is quite a powerful model as it has been proved to converge to the correct Q^* independently of the policy of the agent, with the only requirement being that the state-action pairs are continuously updated (Sutton and Barto [1998]; Watkins [1989]; Watkins and Dayan [1992]).

Sarsa

While Q-learning is an off-policy method, it is possible to use an update rule which is instead on-policy. To do so let us consider the tuple representing a trial of interaction of an agent with the environment. This is characterised by five values:

s_t a non-terminal state from which the interaction begins,

a_t the action taken for the interaction (at time t),

r_{t+1} the reward obtained as a result of the action taken,

s_{t+1} the landing state where the interaction has led the agent at the next time-step $t+1$,

a_{t+1} the action the agent will take at the next interaction, a look-ahead of one time-step into the future.

This quintuple $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ gives rise to the name of this on-policy algorithm: Sarsa. In case the next state s_{t+1} is the terminal state its Q-value is assumed to be 0, $Q(s_{t+1}, a_{t+1}) = 0$. The update rule for this method is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{2.26}$$

The difference between Q-learning and Sarsa is subtle in the calculation but meaningful in the resulting behaviour. The only term which differs in the two update rules is the expected return in the target: in Q-learning this is the estimate deriving from greedy behaviour while in Sarsa this is the estimate deriving from the current policy behaviour. For Sarsa to converge, it is required that the state-action pairs are updated continuously and that the action-selection policy converges to greedy (Sutton and Barto [1998]). This can be achieved by using a time dependent value for ε, such as $\varepsilon = \frac{1}{t}$ where t is the current time-step.

2.2.2 Policy search

When expert knowledge is available it can be integrated in the the modelling by adopting a different problem formulation, so that the policy of the agents becomes the focal point of the learning problem. The structure of the problem, as well as the initialisation of the policy itself, can be enriched to accommodate for the domain-specific information available. This type of approach is of particular interest to the field of robotics (Kober et al. [2013]). In addition to being able to integrate prior domain knowledge into the problem definition, another advantage of policy search is that generally optimal policies present many fewer

parameters as opposed to optimal value functions. Policy search is a valuable alternative to methods based on value-functions in the field of robotics even if the optimal solution cannot be directly derived from the Bellman optimality equations (Kober et al. [2013]). Formally, policy search methods attempt at optimising a policy π, with parameter vector θ_i by iteratively perturbing the parameters in the directions that leads to an increase of expected return, following the rule:

$$\theta_{i+1} = \theta_i + \Delta\theta_i \tag{2.27}$$

Multiple updates have been developed and tested, such as general stochastic optimisation methods (Bagnell and Schneider [2001]), cross-entropy (Rubinstein, Reuven Y and Kroese [2004]), population based methods (Goldberg, David E and Holland [1988]), pairwise comparison (Ng et al. [2006]; Strens, Malcolm JA and Moore [2001]) and gradient estimation with finite policy differences (Kohl, Nate and Stone [2004]; Roberts, John W and Moret, Lionel and Zhang, Jun and Tedrake [2010]; Sato, Masa-aki and Nakamura, Yutaka and Ishii [2002]).

Policy gradient

Policy gradient methods are based on the estimation of the likelihood-ratio (Sutton et al. [1999]). This approach is based on the hill-climbing updating methodology, implying that the updates are to be made, with a step-size α_{PG}, according to the gradient of the expected return, denoted with J:

$$\theta_{i+1} = \theta_i + \alpha_{PG}\nabla_\theta J \tag{2.28}$$

Several methods have been developed to estimate the gradient $\nabla_\theta J$, with many depending on the fine tuning of the step-size parameter α_{PG}. One of such methods is called finite difference gradients. This method requires tuning, not only of the step-size parameter, but also of the number, type and magnitude of the perturbations to be performed on the policy parameters. An estimate of the gradient is evaluated based on the policy parameters perturbed, P_{PG}:

$$\Delta\hat{J}_p \approx J(\theta_i + \Delta\theta_p) - J_{ref} \tag{2.29}$$

with $\Delta\hat{J}_p$ being the estimate of the influence that the perturbations have on the return, $p = [1, \cdots, P_{PG}]$ [5] being the perturbations of the parameters and J_{ref} being the reference

[5] PG subscript omitted for spacing convenience in the following equations

return when the parameters have not been perturbed. The gradient is estimated using linear regression as follows:

$$\nabla_\theta J \approx (\Delta\boldsymbol{\Theta}^T \Delta\boldsymbol{\Theta})^{-1} \Delta\boldsymbol{\Theta}^T \Delta\hat{\boldsymbol{J}} \tag{2.30}$$

with the matrix $\Delta\boldsymbol{\Theta}$ being the collection of the stacked perturbations samples $\Delta\theta_p$. Even if this straightforward approach can be applied to non differentiable policies, it is nonetheless considered inefficient and noisy.

Likelihood ratio methods are an alternative class of methods, where the episodes of a task are believed to be generated according to a particular distribution:

$$P^\theta(\tau) = P(\tau|\theta), \tag{2.31}$$

with the return of a particular episode given by:

$$J^\tau = \sum_{h=1}^{H} R_h, \tag{2.32}$$

where H is the number of steps in the episode. The expected return for a specific set of policy parameters can be expressed as:

$$J^\theta = \sum_t P^\theta(\tau) J^\tau. \tag{2.33}$$

While, the gradient of the episode distribution is expressed as follows:

$$\nabla_\theta P^\theta(\tau) = P^\theta(\tau) \nabla_\theta log P^\theta(\tau) \tag{2.34}$$

This update is known as the likelihood ratio or as the REINFORCE trick, which is an acronym standing for REward Increment = Nonnegative Factor times Offset Reinforcement times Characteristic Eligibility (Williams [1992]).

By combining the last two equations, the gradient of the expected return can be expressed as follows:

$$\nabla_\theta J^\theta = \sum_\tau \nabla_\theta P^\theta(\tau) J^\tau = \sum_\tau P^\theta(\tau) \nabla_\theta log P^\theta(\tau) J^\tau \tag{2.35}$$

$$= E\{\nabla_\theta log P^\theta(\tau) J^\tau\} \tag{2.36}$$

Considering a stochastic policy, denoted $\pi^{\theta}(s,a)$, generating episodes τ, it is not necessary to track the episodes' probabilities as it is possible to specify the gradient in terms of the policy:

$$\nabla_{\theta} log P^{\theta}(\tau) = \sum_{h=1}^{H} \nabla_{\theta} log \pi^{\theta}(s,a). \tag{2.37}$$

While the gradient of the expected return with respect to the parameters of the policy is calculated as follows:

$$\nabla_{\theta} J^{\theta} = E\left\{ \left(\sum_{h=1}^{H} \nabla_{\theta} log \pi^{\theta}(s_h,a_h) \right) J^{\tau} \right\}. \tag{2.38}$$

Since the rewards obtained at the beginning of an episodes are not due to the action taken at the end of the previous episode, the return of the episode can be replaced by the state-action value function in the following way (Peters and Schaal [2008]):

$$\nabla_{\theta} J^{\theta} = E\left\{ \sum_{h=1}^{H} \nabla_{\theta} log \pi^{\theta}(s_h,a_h) Q^{\pi}(s_h,a_h) \right\}, \tag{2.39}$$

this is equivalent to the policy gradient theorem from Sutton et al. [1999]. In this context, the exploration is achieved by the stochasticity intrinsic to the policy. Both REINFORCE and finite differences gradients are considered to be slow (Kober et al. [2013]).

By considering the reward as an improper probability distribution, it is possible to derive a different class of policy-search methods, inspired by Expectation Maximisation (Dayan, Peter and Hinton [1997]). Some examples of approaches from this class that were demonstrated to be effective in robotics research are reward-weighted regression by Peters and Schaal [2008], MonteCarlo EM by Vlassis, Nikos and Toussaint, Marc and Kontes, Georgios and Piperidis [2009], policy learning by weighting exploration with the returns and cost-regularised kernel regression by Kober and Peters [2009]. By combining policy search with the principle of optimality, Bagnell, J Andrew and Kakade, Sham M and Schneider, Jeff G and Ng [2004] proposed a policy search by a dynamic programming method, consisting in learning a non-stationary policy, without attempting to impose the Bellman equation. This approach is one of the most reliable within function approximation, as reported by Kollar, Thomas and Roy [2008], for trajectory selection in a robotic map exploration task.

Methods based on value-function search attempt to find an optimal value-function that can be used to find optimal solutions by simply following a greedy policy, picking the best action in each available state. This is often not achievable because of the high-dimensional state-action spaces of particular tasks (e.g. in the field of robotics). The number of state-action

pairs becomes quickly intractable in moderately complex scenarios. Policy-search methods instead are focused on the current policy and its surroundings. In a robotics context, this approach has the desirable property of being able to accommodate for continuous features, but at the same time it usually cannot find the global optima but is confined to a sub-optimal solution local to the current policy.

In Kober et al. [2013], a robotic arm is used to learn the task of paddling a ball with a ping-pong racket. The combinations of robot position, bearing, actuators' velocity and angles quickly leads to representational and computational intractability. For such tasks it is therefore necessary to resort to function approximation.

2.2.3 Function Approximation

Function approximation is a class of methods which allows the representation of a function which would otherwise be intractable, either computationally or from an information theory point of view (Rivlin [1969]). In Reinforcement Learning, function approximation is often necessary to be able to represent the problem in both continuous state-spaces and discrete, large scale ones (Kober et al. [2013]; Sutton and Barto [2018]). In fact, it is impractical and time consuming to visit all the available states and try all the possible action for each one of these. It is likely that many of the available state-action pairs lead to similar results, especially if they are neighbours. Therefore, function approximation can conveniently generalise these, condensing the amount of information which needs to be stored. Parametric function approximators attempt to fit the observed data as closely as possible by finding the set of parameters which allows this. Linear basis functions and neural networks are examples of this class of function approximators. The former is popular in literature mostly because of its simplicity in approximating the value-function, which represents a state with a scalar value. For example, a state-space for a grid-world can be represented with a grid of Gaussian basis function centered in each of the states. The approximated function value evaluated at each particular state point can be estimated as the weighted sum of the basis functions. This type of approximation can also be applied in cases where the state is not defined by coordinates of a discretised grid-world. For example, a robotic arm state can be approximated with a linear combination of the features capturing the position, velocity, angle of the actuators of the arm. The set of weights is chosen to minimise the distance between the observed data and the approximation developed. Linear regression can be used to estimate the weights when the mean squared error is used as distance measure. Tile coding is another type of function approximation (Sutton and Barto [1998, 2018]), which consists in subdividing the space in tiles which can be irregular to accommodate for portions of the

space with different salience. This method is flexible because it allows a multi-dimensional continuous space to be approximated with overlapping tilings at the desired level of detail.

Theoretically, any supervised learning technique can be used as a function approximation method (Sutton and Barto [2018]), but some of these have not been studied extensively. One common way to implement non-linear function approximation is with Artificial Neural Networks (ANNs). Multi-layer networks provide a hierarchical mapping of input features into output with a higher level of abstraction. This particular network topology does a particularly good job as function approximator because works as a feature extractor (Duan et al. [2007]; Gaskett et al. [2000]; Thrun [1995]). More complex features of the space can be represented, instead of adopting only the hand-crafted ones designed or identified by the human expert. In supervised learning ANNs are usually trained by changing the weights according to the gradient leading in the direction which minimises the distance between a set of labelled examples and predictions. In reinforcement learning, on the other hand, these networks can learn value functions by using TD errors (Sutton and Barto [2018]). ANNs are trained by backpropagation, an algorithm that combines forward and backward passes of the network. In the forward pass the networks calculates through weights and activation functions the prediction, while in the backward pass the partial derivatives for each weight is computed and the weights updated. Alternatively, ANNs can be trained using principles from reinforcement learning. This methodology is less efficient but could be closer to the real computations happening in the brain (Sutton and Barto [2018]). Function approximation can also benefit policy-search methods, by reducing the number of parameters and making the problem tractable (Kober et al. [2013]). Again, neural networks can be used for this purpose, for example in a robotic peg-in-hole insertion task and in a ball-balancing task (Gullapalli et al. [1994]), as well as in a navigation task (Hailu and Sommer [1998]).

2.2.4 Partial observability

An extension of the reinforcement learning setup is possible by considering Markov Decision Processes characterised by partial observability. Partially Observable Markov Decision Processes (POMDPs) are an extension of MDPs. POMDPs differ from MDPs because the partial observability implies that the agent does not have full information about the state. This greatly increases the complexity, leading to an exact solution virtually impossible to find (Braziunas [2003]). It is, in fact, not possible to include previous observations as stored memories within the state representation (Sutton and Barto [2018]).

An MDP is enhanced to become a POMDP by extending it with the observation space O and the observation function $Z(\cdot)$. The first is a set of observations which can be obtained by the agent. Formally, the POMPD is a tuple $< S, A, T, R, O, Z >$, comprising the state-space

S, the action space A, the transition function $T(\cdot)$, the reward function $R(\cdot)$, the observation space O and the observation function $Z(\cdot)$. To draw a parallel, in an MDP the agent possesses complete knowledge of the system state:

$$O \equiv S. \tag{2.40}$$

While in a POMDP, the observations depend on the underlying state in a probabilistic fashion. There arises the problem of determining the state in which the agent is, as multiple observations can be obtained in distinct states.

The observation function represents the relationship between the states of the environment and the observations and is defined as:

$$Z : S \times A \rightarrow \Pi(O). \tag{2.41}$$

The probability that observation o' can be seen after the agent takes action a and arrives in state s' is denoted with:

$$Z(s', a, o') = P(O^{t+1} = o' | S^{t+1} = s', A^t = a). \tag{2.42}$$

Within the formulation of a POMDP, the agent has to decide if it has enough confidence about the current state of the environment, so that it can take an action greedily. Alternatively, the uncertainty is such that more exploration is needed to be able to take a better action at a later time. Many applications can be identified for this stochastic-domain planning problem formalisation: process control within factories, raw resources location exploration, elevator control, marketing, logistics and transportation can all benefit from adopting POMDPs and RL (Cassandra [1998]; Kaelbling et al. [1998]). While for MDPs it is possible to get the optimal policy and act accordingly for each state, in POMDPs if the agent cannot have full certainty of the state it is in, the action cannot be chosen deterministically. A mapping from observations to probability distributions over actions is developed for this case (Kaelbling et al. [1998]). This stochastic approach allows the agent to explore with some probability in cases in which the state appears to be similar to other states. A formalisation of the tasks examined in this thesis could adopt a POMDP, but considering the state-spaces which will be proposed in the next chapters there is no necessity for this.

2.2.5 Inverse Reinforcement Learning

So far, the reinforcement learning techniques presented were all based on the assumption that the agent has access to the reward function, has information about the environment

and will learn a particular behaviour for the task specified. This type of learning is similar to the trial-and-error that can be observed in animals and humans. There is an alternative setup to this, called Inverse Reinforcement Learning (IRL). By providing an agent with the information about the environment and examples of behaviour from an external source, it is possible to teach the agent the reward function (Fig. 2.13, panel b)). This type of learning is similar to another learning method exhibited by animals and humans, learning by imitation. As an example, this is the type of learning mechanism people use initially when learning martial arts. The instructor shows the movements in detail and the students observe and repeat. They will then, by trial-and-error, perfect the movement and improve its efficiency. The combination of this two learning points of view is called apprenticeship learning (Abbeel and Ng [2004]; Kober et al. [2013]).

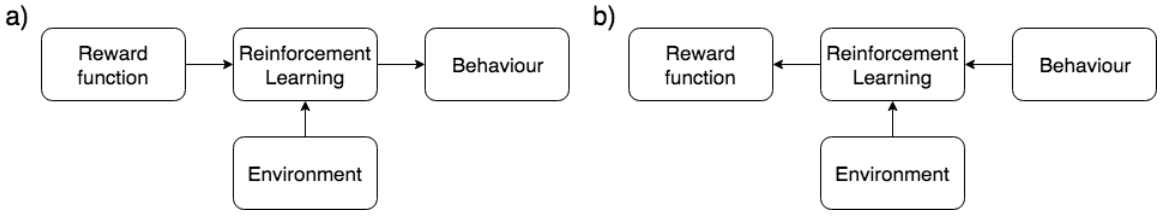

Fig. 2.13 The scheme of information flow in reinforcement learning in panel a), and inverse reinforcement learning in panel b).

The expert trials shown to the agent can be used as training data, with a supervised learning approach, to map states to actions (Kober et al. [2013]). Moreover, the available information can be used to restrict the boundaries of the search and avoid global exploration for both the state-space and the policy (Kober et al. [2013]). Naturally, if the expert knowledge provided is not in the neighbourhood of the optimal solution, only local optima can be achieved. In a robotics setup, the expert demonstrations can be performed by manipulating the robot itself or by showing the movement and letting the agent observe it. The latter case can sometimes be problematic due to limitations of degrees of freedom or simply incomparable properties of human demonstrator and robot. Nevertheless, in a ball-in-a-cup task by Kober et al. [2008] and in a pendulum swing-up task by Atkeson et al. [1997], motion-capture technology has been used to record and encode the movements of the demonstrators so that the robot could interpret them. Explicit expert demonstrations by direct manipulations can be achieved by, for example, guiding a robot in a navigation task to initialise the Q-function table in Conn and Peters [2007]. Alternatively, direct manipulation of the robot, also known as kinesthetic teaching, has been used in several works: a reaching task (Bitzer et al. [2010]; Guenter et al. [2007]), in a T-ball batting task (Peters et al. [2004]), in a ball-in-a-cup task (Kober and Peters [2009]) and in a door opening and object pick-up task (Kalakrishnan

et al. [2011]). State-of-the-art acrobatic helicopter flight learning was accomplished by Coates et al. [2009], by a combination of trajectories demonstration by human experts, extraction of approximate models via machine learning and locally optimal control methods (Kober et al. [2013]).

Reinforcement learning has also been used to create Financial Trading Systems (FTS), with promising results (Bertoluzzo and Corazza [2007, 2012]; Chen et al. [2007]; Lee [2001]; Moody and Saffell [2001]; O et al. [2006]), and to develop agent-based Stock Market Simulations (Rutkauskas and Ramanauskas [2009]). These works use RL as a predictive tool for future prices but do not attempt to use RL as a descriptive account for describing investors' behaviour. In an effort to provide a unified account of decision-making tasks involving learning from experience we link computational models from reinforcement learning with ideas from prospect theory. The former is a descriptive computational framework used to capture the mechanisms behind the formation of biases DMs exhibit when operating decisions in the tasks of interest. The latter is a descriptive theory of the psychological biases which lead to deviations from maximisation and irrational behaviour in such tasks.

Chapter 3

Experimental studies of decisions from experience

The first part of this study aims to clarify how decision-making behaviour deviates from rationality for subjects learning from experience and the role of myopia and payoff variability in this process. Relevant studies on decision tasks involving experience and feedback are presented in this chapter, along with the modelling efforts and their findings. This review will include details on the experimental data used in this thesis, which has been collected in Barron and Erev [2003], partly replicating the experimental setup in Thaler et al. [1997].

Both these studies present some challenges: firstly the behaviour shown by the subjects is quite varied, with some players performing opposite decisions. Moreover, the attempts to model these decisions were based on aggregate modelling, which leads to lack of granularity. The relationship between myopic behaviour and performance is studied with a methodological shift from these previous studies, by modelling participants separately. Another methodological limitation is the adoption of a measure of distance between observed and predicted choices as model estimation, which can be improved with a probabilistic approach. Results from these studies identify Reinforcement Learning as a potential component of the models studied, leading to the choice of testing subjective RL models in this thesis.

The previous studies do not try to model the influence of the history of the observed payoffs on the preference of the subjects. This thesis' first hypothesis is focused on this and intends to shed light on the relationship between information previously obtained by a subject and future choice preference. To do so, the RL models adopted will be augmented with a two-state space setup. Confirming this association between previous outcomes and decision-makers' risk preference will inform future experiment designers that payoff information is a crucial part of the environment and the modalities of presenting it should be carefully evaluated.

Loss aversion has been widely proven to affect choice behaviour. Even if some attempts have been made to encapsulate this into the models adopted in previous studies, these never tested alternatives. Multiple reward functions are adopted in this work to test the hypothesis that subjects are loss averse and show diminishing sensitivity to gains and losses. This test will help improve the understanding of the modalities with which decision-makers internalise information about potential losses and act accordingly.

Similarly to how previous studies have attempted to model loss-aversion tendencies, speed of learning has been included in the components of the models proposed. More variability in the observed payoffs leads to an increase in the time required for a subject to learn the correct action. This modelling approach is maintained in this thesis in order to capture this phenomenon and as an attempt to confirm these findings within the RL modelling framework.

According to these works, the variability in observed payoffs also influences the degree of exploration of the subjects. A fourth hypothesis is proposed in order to better understand this potential relationship within the proposed modelling scenario.

Finally, all the modelling attempts made in these studies decisions do not focus on time discounting, which could prove a valuable addition to understanding individual decision-making. Therefore, another hypothesis of this work is focused on the relationship between far-sightedness and task performance. In case this relationship is confirmed, this work will be evidence for a need of more complex models when attempting to describe choice behaviour.

An overview of the previous works is provided in the next sections. The modelling efforts attempted will be detailed in section 3.3. The results of these studies are presented in section 3.2 while in the last section 3.4, a full critique about these works is presented.

3.1 Previous studies on loss aversion and myopia

In their study, Thaler et al. [1997] structured the experimental conditions in a way to test two phenomena that affect the decision-making process in tasks involving experience and direct feedback: loss aversion and myopia. Loss aversion is a phenomenon which leads to the preference of decision-makers (DM) for safe options, even when these lead to suboptimal results in the long term. Out of two prospects, the option that maximises the long term return but also yields negative outcomes, is considered less worthy than the lower valued option which consistently grants low but positive payoffs. In Gneezy and Potters [1997], Thaler et al. [1997] and previously in Benartzi and Thaler [1995], the concept of myopia is defined according to features of mental accounting, which according to PT, defines the framing of the choices and the subsequent outcomes a DM faces. Individuals who frame decisions in a

narrow way will have the tendency to operate short-term decisions, while those subjects who have a wider framing will tend to follow long-term strategies. In the same fashion, individuals will evaluate their gains and losses more frequently when they frame past outcomes in a narrow way. According to Thaler et al. [1997], these two points of view combined - narrow framing of decisions and of outcomes - define a myopic investor. These concepts are related to the equity premium puzzle. Mehra and Prescott [1985] show that DMs prefer the less volatile option (T-bills, bonds) as opposed to the risky one (equities, stocks), even if the risky option has been proved to greatly outperform the risk-free counterpart. This can be explained as a result of both loss aversion and myopic accounting (Benartzi and Thaler [1995]; Thaler et al. [1997]). Investors are reluctant to take on assets which can lead to losses. At the same time investors who do not wait enough time to assess their decisions outcomes might never witness the higher average return of stocks achievable in the long-term.

The experimental design used to test these effects in Thaler et al. [1997] is the following: 80 students from University of California Berkeley were subjects in the task of portfolio management. They were asked to allocate a portfolio comprising 100 shares between 2 available investment options. Fund A payoff was a draw from a normal distribution with mean 0.25% and standard deviation 0.177%, truncated at 0 to avoid negative payoffs. This option describes the safe option. Fund B payoff, which instead characterised the volatile option, was a draw from a normal distribution with mean 1% and standard deviation 3.54%. These distributions were chosen to characterise the real-world returns of five-year bonds and value-weighted stock index over 6.5 weeks. The subject did not know the distributions beforehand and had to acquire this information, learning about the returns of the options through direct experience. The subjects were subdivided into four groups, each assigned to a different condition. In the first group, named "monthly" condition, the subject had to decide the portfolio allocation for 200 trials. In the second group, denominated "yearly" condition, subjects did the same for 25 times, each decision was binding for the subsequent 8 trials. For the third group, "five-yearly" condition, subjects made only 5 allocations, each one biding for 40 trials. The last group, "inflated monthly" followed the same structure as the "monthly" group but the payoffs were shifted by 10% in order to prevent the subject from experiencing negative returns in any of the investments available. At the end of the 200 trials, subjects were asked to make a final decision which would be binding for 400 trials and for which there would be no intermediate feedback. Summarising, the available portfolio options were:

A : $X \sim \mathcal{N}(0.25,\ 0.177)$ truncated at 0

B : $X \sim \mathcal{N}(1,\ 3.54)$

The conditions to which the students were assigned were:

1 monthly: 200 allocations during 200 trials;

2 yearly: 25 allocations, binding for the following 8 trials;

3 five-yearly: 5 allocations, binding for the following 40 trials;

4 inflated monthly: same as monthly but with a positive shift of 10% for the payoffs.

Similar experimental conditions were studied in Barron and Erev [2003], to better understand the deviations from maximisation that occur when DMs face a feedback-based decision task, which are often in the opposite direction of the ones that happen when DMs face decisions from description. Interestingly, two of the conditions examined are replications of the first and the last scenarios from Thaler et al. [1997], respectively "monthly" and "inflated monthly". A third condition not studied in the original work is added. The design for the replication study was similar but with substantial differences.

36 students participated in the experiment, most of them second or third year industrial engineering or economics students with some knowledge of probability and economics from their course. Subject were not told the experiment would last exactly 200 trials in order to avoid a change in their risk attitude, but they knew the duration of the experiment would be approximately 30 minutes to an hour. Subjects were randomly distributed in each of the three conditions.

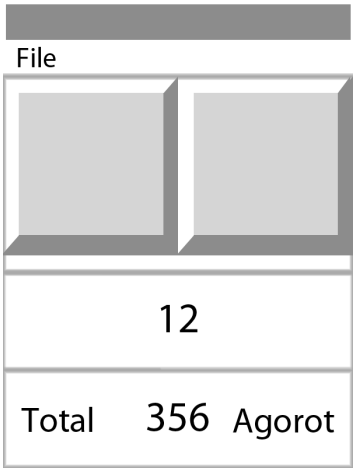

Fig. 3.1 A replication of the original experimental screen for the "gamble" tasks studied in Barron and Erev [2003]. The buttons represent the two available choices and the two numbers represent the outcome of the last action and the total accumulated reward.

The baseline condition, a replication of the "monthly" condition in Thaler et al. [1997], featured two unlabelled options (in this summary L for low risk and H for high risk):

L : $X \sim \mathcal{N}(25,\ 17.7)$ truncated at 0 to avoid negative payoffs (implied mean 25.63).

H : $X \sim \mathcal{N}(100,\ 354)$.

The second condition (replicating "inflated monthly" condition in Thaler et al. [1997]) displayed the same options, but their underlying payoff distributions were shifted of 1200 points, to avoid negative payoffs, while keeping the standard deviations the same:

L : $X \sim \mathcal{N}(1225,\ 17.7)$.

H : $X \sim \mathcal{N}(1300,\ 354)$.

In the third conditions, which was not tested in the original study, the options had the same mean as in condition 2 but both had a standard deviation of 17.7.

L : $X \sim \mathcal{N}(1225,\ 17.7)$.

H : $X \sim \mathcal{N}(1300,\ 17.7)$.

The "money-machine" game interface presented to the subjects is replicated in Fig. 3.1 and consisted of two buttons and two feedback outputs: the latest payoff which appeared after a choice was made and lasted for 1 second and the total accumulated payoff, which was permanently displayed. Subjects only made a selection between the two options and received only the payoff information about the selected option and an update of their history payoff balance. No information was given about the forgone payoff (the payoff of the option not selected). The accumulated points were then translated into real money at the end of the trials with a conversion rate of 100 points = 0.05 Shekels, which at the time of date collection corresponded to about 0.0125 USD. Subjects also received a flat reward of 5 Shekels for having participated in the experiment. The final payoffs ranged from 6.25 to 15 Shekels, corresponding to about 2 to 4 USD.

The proportion of maximisation (*Pmax*) performance measure in Barron and Erev [2003] is calculated as:

$$Pmax = \frac{1}{N}\sum_{i=1}^{N}[c_i = high] \tag{3.1}$$

where $[\cdots]$ are the Iverson brackets [1], c_i is the choice at the $i-th$ trial and $N = 200$ is the total number of trials. This measure cannot be directly evaluated by the subjects as they are

[1] defined on a logical proposition P: [P] = 1 if P is true, 0 otherwise.

not aware of which option is the one with the highest expected payoff. It is instead used as the measure of task performance which the experiment designer considers to assess whether individuals improve over the course of the trials (e.g. Barron and Erev [2003]; Erev and Barron [2005]; Erev et al. [2012]; Thaler et al. [1997]).

The proportion of maximisation choices (Pmax, section 4.1.16, eq. 3.1) offers a measure of the subjects performance in this task. It is used to track the overall change in behaviour over the course of the entire duration of the subjects' interaction with the money-machine. Each subject choice dataset comprises 200 trials. The plots in Fig. 3.2 are obtained by subdividing the choices in four blocks of 50 trials. Each panel of this figure represents a condition and each line represents a subject's Pmax across the four blocks. In the first two panels of Fig. 3.2, which present the Pmax learning curves for the first two conditions, it can be noted that the subjects produce quite diverse behaviour. This variability is what motivates the proposed methodology for this study, analysing each subject separately and characterising their strategy with the most appropriate model among the ones proposed. Previous analyses aggregated the choice-data for each subset of subjects within a condition, producing an "average subject". As shown in the figures, there is quite a lot of variability among subjects, therefore the average subject would not be representative of the experiences and decisions of each individual subject and could only reflect a distorted version of the strategies composing it. It is also interesting to note an overall shift of the learning curves towards a higher proportion of maximisation choices between condition 1 and 2. As suggested in Barron and Erev [2003], this is likely due to the loss aversion phenomenon. In fact, subjects in condition 1 experienced either positive but low outcomes or a wide range of payoffs which could also be negative. Even if the distance between the means and the payoff variance was held constant between condition 1 and 2, the mean payoffs were shifted so that in the second condition no negative payoff could be experienced. The third figure shows the Pmax learning curves for the third condition and presents the strongest learning pattern exhibited across the three conditions. It can be seen that the subjects learned to consistently choose the option which would maximise their profit early in the first blocks and increased this preference over the course of the remaining trials. These general patterns of learning behaviour are in line with other studies (Hertwig and Pleskac [2010]; Mehlhorn et al. [2013]), where the decision-makers showed an impaired ability to choose between two options when the observed difference between the payoffs of the two was smaller.

Subject 9 in condition 2 (the blue line at the bottom) exhibits a compelling instance of odd behaviour. It appears to briefly choose the best option in the first block before turning to the lower one and continuously selecting it for the rest of the interaction. It could appear that this subject represents an outlier. A more detailed inspection of the interactions pertaining to

this individual revealed that in the first block the maximisation option was selected only at a point, resulting in a much lower outcome compared to the payoffs obtained in the many selections of the lower option up to that point. In substance, this subject briefly explored but never witnessed an outcome which would appear to be truly worth switching choice. Because this strategy is still reasonable, yet suboptimal, subject 9 was not removed from the dataset analysed. Another interesting pattern is the one presented by subject 11 in the same condition, portrayed by the purple line peaking during block 2 and 3, at the 100th and 150th trial and then falling to a proportion of 10% of Pmax choices. This inversion of trend seems counter-intuitive but it is readily explained by the outcomes obtained which are plotted as average for each block in Fig. 3.3. Condition 2 subjects are all plotted with dashed lines except for subject 11, in order for the comparison to be more immediate and readable. It can be noted how for such subject the average payoff greatly reduced in block 3 (150 trials mark), leading to a change in strategy in the subsequent 50 choices. This peculiar example shows how the task in this condition, as well as in condition 1, was not trivial. On the contrary, the graphical indications from condition 3 figure point in the direction of straightforward learning for the majority of the individuals.

3.2 Findings and interpretation

The results in Thaler et al. [1997] support both myopic loss aversion predictions. The first hypothesis predicted that the commitment of individuals to the low-risk option would decrease as the evaluation time-frame increased. The percentage of final allocation is used to evaluate the effect of longer evaluation periods. The monthly condition is used as baseline for the comparison and all the values refer to the final allocations made by the participants. In the monthly condition 21 subjects committed on average $59.1\% \pm 35.4\%$ of their assets to bonds (low-risk option). In the yearly condition 22 individuals chose on average to assign to this option $30.4\% \pm 25.9\%$ of their assets. The five-yearly condition group (22 subjects) showed a similar response, with a mean of $33.8\% \pm 28.5\%$ their portfolio devoted to bonds. The two alternative time-frame conditions (yearly and five-yearly) are significantly different to the baseline condition (monthly) but are not significantly different from each other. The authors identify two potential explanations for the two mean values of bond-allocation being similar in the yearly and five-yearly conditions. Firstly, the frequency of feedback for the five-yearly condition is much less than the yearly condition, respectively 5 trials and 40 trials. Secondly, the negative feedback derived from stock oscillations is thought to be the most insightful experience. These events are much less frequent in the yearly condition as opposed to the monthly condition, respectively 14 % and 39 % of the trials yielded a negative

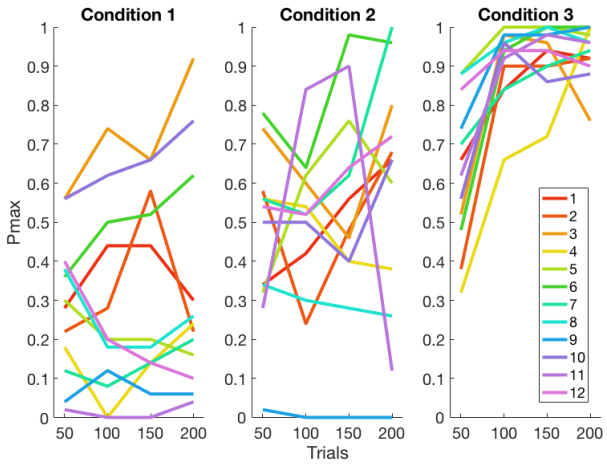

Fig. 3.2 Proportion of maximisation choices aggregated over four blocks (50, 100, 150 and 200 trials marks).

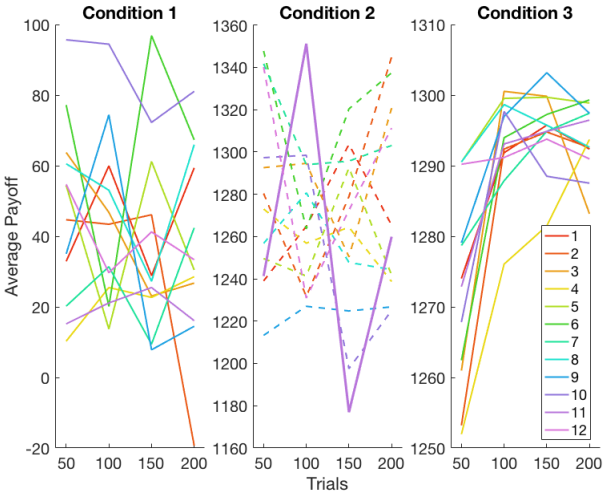

Fig. 3.3 Average payoff obtained by subjects in each of the four blocks of 50 trials each. The dashed lines are used to make subject 11 easier to visualise.

outcome. Negative feedback was completely absent in the five-yearly as no negative return was experienced by the subjects. Therefore the difference in frequency between the two wider time-range conditions is small enough not to produce a noticeable effect in choice behaviour. The second hypothesis predicted that the low-risk option preferences would decrease when all available options yield strictly positive outcomes. The 21 subjects in the inflated-monthly condition presented an even greater reduction from the baseline monthly condition allocation when compared to the longer time-frame conditions. The subjects chose bonds on average for only $27.6\% \pm 23.2\%$ of their allocation.

The replication of these conditions by Barron and Erev [2003] leads to similar results to those obtained in Thaler et al. [1997]. Conditions 1 and 2 replicated respectively the monthly and inflated-monthly conditions from Thaler et al. [1997]. The aggregated results of the experiment for the first two problems show similar choice behaviour. The overall preference emerging from subjects' aggregated proportion of choices for the risky option, which corresponds to stocks in the previous study, is 30%($\pm$22%) in condition 1 and 51% ($\pm$21%) in condition 2. The preference for the low-risk options (corresponding to bonds) decreases when the experienced outcomes are strictly positive in a similar way as in Thaler et al. [1997]. The results for condition 3, in which the two payoffs have the same mean as in condition 2 and both have the same variance, show a big increase in subjects' commitment to the risky option: $85\% \pm 8\%$. These results indicate that there are two possible explanations for the deviation from maximisation behaviour in condition 1. The first explanation, as in Thaler et al. [1997], is loss aversion while the second is payoff variability effect: subjects deviation from maximisation increases when the payoff variability increases.

3.3 Modelling

In Barron and Erev [2003] the authors use a value assessment (VA) model to provide an abstraction of the principles believed to cause the behaviour observed: loss-aversion and reliance on recent outcomes. These are captured in the model with the following assumptions. The adjusted value of a choice at each time is calculated as a weighted average of subjective value of the obtained payoffs at previous times.

$$A_j(t+1) = (1-w_t)A_j(t) + w_t v(x_t) \tag{3.2}$$

where $A_j(t+1)$ is the adjusted value of choice j at time $t+1$, $v(x_t)$ is the subjective value of payoff x_t and $w_t \in (0,1)$ is the weight of the value. The model also assumes that the weight w_t is dependent on the trial type:

$$w_t = \begin{cases} \alpha_{VA} & \text{if } t \text{ is an exploration trial} \\ \beta_{VA} & \text{otherwise} \end{cases} \tag{3.3}$$

where $0 \leq \beta \leq \alpha < 1$. The recency effect is captured by the fact that both α and β do not decrease with time. The adjustment speed, corresponding to the weight w_t, is constrained to be independent of time t in order to capture the recency effect.

To avoid the case in which repeated bad outcomes cancel the probability of choosing a prospect, the model is assumed to explore and exploit based on probability:

$$P(explore\ at\ time\ t) = \frac{\kappa}{t+\kappa} \tag{3.4}$$

where κ is a parameter regulating the strength of the exploration. The loss aversion principle is captured by the subjective value function:

$$v(x_i) = \begin{cases} x_i & \text{if } x_i \geq 0 \\ \lambda x_i & \text{if } x_i < 0 \end{cases} \tag{3.5}$$

where λ is the loss-aversion coefficient of prospect theory's subjective value function. In Barron and Erev [2003], the authors did not estimate this coefficient but simply adopted the value suggested from literature of $\lambda = 2.25$ (Tversky and Kahneman [1992]). For exploitative trials the model will pick the choice with the best associated adjusted value. The ability of this model to capture the results of the experiments in Barron and Erev [2003] is estimated by means of minimisation of mean squared deviations (MSD) between the observed maximisation rates and the model's predictions.

In a following work, the same authors proposed a more complex model which assumes REinforcement Learning Among Cognitive Strategies (RELACS). This more complex modelling effort was made, among other purposes, to capture the V-shaped curve of condition 1 (Erev and Barron [2005]), characterised by the choice between L ($X \sim \mathcal{N}(25,\ 17.7)$ truncated at 0) and H ($X \sim \mathcal{N}(100,\ 354)$). This model assumes that in each trial the DM will follow one of the three cognitive strategies implemented and that the probability of picking one of the three is based on previous reinforcements obtained in past experiences with each rule. This model is based on four assumptions, some of which are based on the observation that probability matching rules provide a good fit to the choice data. The first rule, named "Fast Best Reply" is a weighted adjustment rule, like the one adopted in the VA model (Eq. 3.2):

$$R_j(t+1) = R_j(t)[1-\beta] + v_j(t)\beta \tag{3.6}$$

where $v(t)$ is the observed payoff from option j in trial t and $\beta \in (0,1)$ is a recency parameter which tunes the relevance of recent outcomes. High values of β indicate overweighting of recent outcomes. As the task follow the partial feedback paradigm, the value of action j is updated only when the payoff resulting from action j is observed.

The second strategy considered in the RELACS model assumes a case-based reasoning along with a loss-avoidance and is structured in two stages. In the first stage the DM randomly

recalls one of the previous trials for each option and uses the outcomes from the selected trials as basis for the belief forming process. The loss-aversion check is enforced in the second stage, in which the DM checks if the action with higher payoff recalled from the first stage is associated with larger and more frequent losses. This is done by recalling a set of κ previous outcomes and computing the number of losses and the total losses from such outcomes. κ is a free parameter which is intended to represent the sensitivity to rare losses.

The third assumption, named "Slow Best Reply", assumes a stochastic strategy for action selection which includes a decreasing but continuous exploration. The probability of action j being picked is calculated with a soft-max rule (Boltzmann distribution):

$$p_j(t) = \frac{exp(W_j(t)\lambda/S(t))}{\sum_{k=1}^{2} exp(W_j(t)\lambda/S(t))} \tag{3.7}$$

where λ is the exploration-exploitation trade-off parameter; when λ assumes low values the action selection becomes more exploratory. This parameter does not have any relation to the homonym loss-aversion coefficient adopted in their previous work. $W_j(t)$ is the weighted average payoff for action j and $S(t)$ is a value which measures the payoff variability. The value of $W_j(t)$ is calculated like in Eq. 3.6 but with a parameter $0 < \alpha < \beta$ which implies slower updating:

$$W_j(t+1) = W_j(t)[1-\alpha] + v_j(t)\alpha \tag{3.8}$$

The value of the payoff variability $S(t)$ is calculated as:

$$S(t+1) = S(t)[1-\alpha] + |v(t) - max(Last_1, Last_2)|\alpha \tag{3.9}$$

where the second term is the absolute value of the difference between the observed payoff $v(t)$ and the maximum between the last received payoffs of the two available actions, multiplied by α. The last assumption is the rule for choosing among the three strategies, which follows the same rule in Eq. 3.7 with the constraint that the weighted average for each rule is updated only when such strategy was used.

3.4 Discussion and further modelling extensions

The analysis carried out in Thaler et al. [1997] is based on aggregate data but the distribution of final allocation shows a great degree of difference in allocation among subjects within each group as shown in Fig. 3.4.

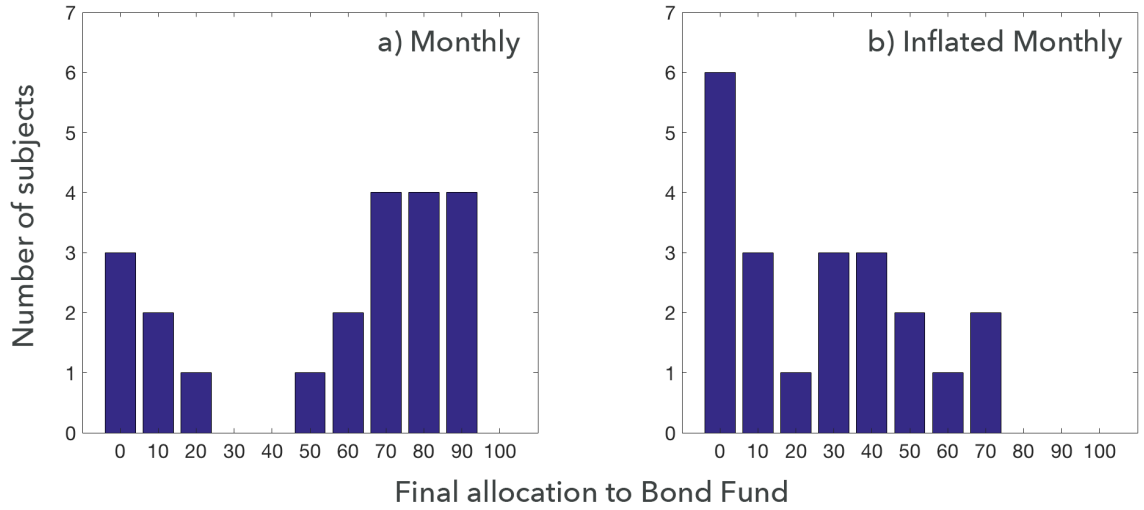

Fig. 3.4 Distributions of the final allocation to bond fund across subjects for each condition in Thaler et al. [1997]. The x-axis represents the percentage of allocation to Bond Fund in the final allocation (duration of 400 trials), the y-axis represents the number of subjects who allocated that specific quantity of their portfolio to Bond Fund.

This experimental work showed that exposure to less frequent intermediate outcomes elicits different responses from individuals as opposed to subjects receiving more frequent feedback. Moreover the results show an increase in the propensity to pick the risky option when there is no loss involved in doing so. These findings indicate that there is potential for a more in depth analysis. In fact, examining the results from Thaler et al. [1997] reproduced in Fig. 3.4, a large variability in choice is noticeable between subjects who learnt that the stock fund would reward them with higher outcomes in the long term, and subjects who kept acting sub-optimally, allocating consistent portions of their portfolio to bonds. This split is particularly strong in the monthly condition (Fig. 3.4 A), where the population distribution appears bimodal and U-shaped but an irregular distribution of preference among participants is also present in condition inflated-monthly (Fig. 3.4 B). A bimodal distribution could indicate that there are two approaches to the task followed by subjects sub-groups. A better understanding could be achieved by considering subjects individually and including the evolution of the subject's choices and their outcomes over the course of the trials into the modelling. Even if the results from the aggregate analysis are statistically significant and interesting in the context of this decision-making tasks, a deeper and quantitative examination of individual differences could help clarify the potential connections between myopia and performance. An alternative explanation for the non-normal distributions observed in Figs. 3.4, is the low number of subjects in the study. The number of subjects considered is

restricted to the data obtained and extending the subject pool is a good starting point for future improvements of the current analysis.

The approach in Barron and Erev [2003] provides an interesting graphical overall comparison for each condition but fails to explain inter-subject variability and does not attempt to capture the degree to which subjects exhibit myopic behaviour. High values of standard deviation scores are reported and the analysis of subjects singularly results, as in the previous work by Thaler et al. [1997], in bimodal U-shaped distributions. The VA model in Barron and Erev [2003] fails to address this as the analysis is based on the assumption that all subjects share the same parameters. This model also fails to capture the long-term learning trend. In condition 1 for example, the trend of observed choice data results in a V-shaped curve while the VA model predicts a linear decreasing proportion of maximisation choices over time. Moreover, the fitness of the model to the data is estimated with the mean squared deviations (MSD) between observed and predicted choices probabilities, but this method has an important shortcoming in that it aggregates choice data over a large set of experiments and conditions (4 experiments and a total of 11 conditions). This failure to model subjective preferences can be addressed with a personalised modelling approach based on maximum likelihood estimation. This technique, which will be presented and formalised in the next chapter, does not necessarily aggregate data over many subjects but can be used to maximise the probability of a model being representative of a series of choices individually, for each subject.

Finally, the RELACS model was fitted to the data by running 200 computer simulations of the 40 tasks deriving from the experimental conditions in analysis and searching for the optimal value for each parameter (Erev and Barron [2005]). The comparison to the observed data is carried out by means of MSD value and the graphical depiction of the learning curves for both observed and predicted choice data. The RELACS model is compared to simpler models, including a reinforcement learning one, which can be described as simplified versions of RELACS using only one of the strategies (Slow Best Reply). It is pointed out how the Slow Best Reply learning rule is important as it captures the payoff variability effect by means of the term $S(t)$ in Eq. 3.7. A further analysis shows that the overall assumption that the learning process switches among strategies is not very important as random strategy choice has a fitting performance not much worse than the full RELACS one. These findings indicates that the RL strategy with stochastic action selection (Slow Best Reply) and accounting for payoff variability effect, could be the most important component of RELACS, but the authors favour the full version of their model as it achieves better MSD score on all the data sets tested. Moreover, the authors point out that the results expose significant individual differences. The analysis of the individual choices gave rise to many

disparate distributions of subjective preferences, for many of the problems considered; in some cases resulting in bimodal U-shaped distributions (Erev and Barron [2005]). The distribution of proportion of maximisation choices, for three conditions replicating the three conditions in Thaler et al. [1997], is shown in Fig. 3.5.

After analysing the methods, findings and shortcomings of previous works and in order to give a meaningful answer to the research questions of this thesis, it is necessary to develop a different approach to analyse the available data. The proposed modelling methodology could help link together two powerful descriptive framework of decision-making. Moreover, the proposed approach is based on the individual fitting of subjects' choice data, which will allow for a more in-depth investigation of the differences of decision-makers and how their strategies correlate to their task performance. In this regard, the suggested models are based on the powerful reinforcement learning framework, which allows subjects characterisation from a behavioural point of view, including their myopic or far-sighted tendencies. The details of this modelling will be presented in Chapter 4. Along with Erev and Barron [2005], other investigations have effectively used reinforcement learning computational models to describe experimental decision-making behaviour, but these do not include time discounting into their learning model (Frey et al. [2015]). The present investigation fills this gap by adopting more complex models and a probabilistic model-selection methodology, respectively Q-learning and maximum likelihood estimation.

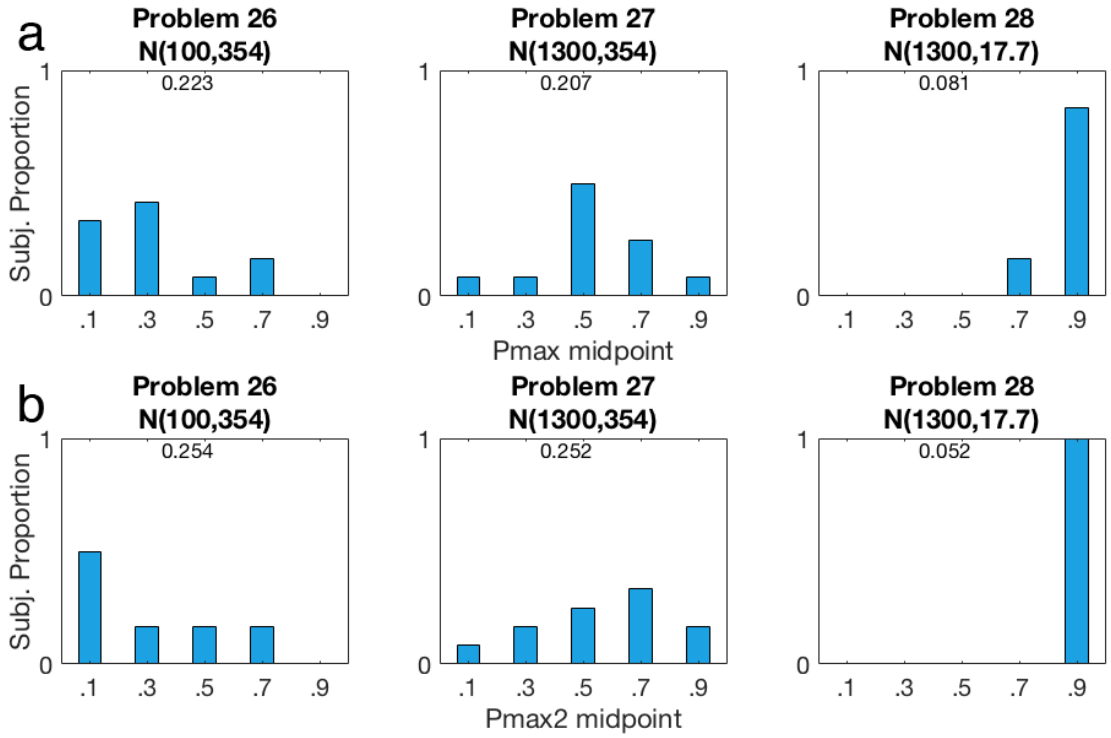

Fig. 3.5 Distributions of the subjects according to their proportion of maximisation choices (Pmax) in all trials and in the second block of trials. The x-axis represent the midpoint of the proportion of maximisation choices, while the y-axis represent the proportion of decision-makers. The titles report the original problem number (from Erev and Barron [2005]), and the gamble with the higher expected value, where N(x,y) is a draw from a normal distribution with mean x and standard deviation y. The values inside the charts, on top of the figure indicate the standard deviation of the Pmax. Panel a) presents this information estimated for all the 200 trials, panel b) shows the same information calculated on the second block of trials, the last 100 interactions and is a reproduction of Fig. 5 from Erev and Barron [2005].

Chapter 4

Methods and Modelling

This chapter presents the methodology adopted in this work, including the models proposed, their fitting procedure and the approach adopted to evaluate said models. The first part of the chapter presents the work focused on testing a series of candidate models, to describe a dataset of choices and outcomes deriving from an experimental setup from Barron and Erev [2003], replicating some of the conditions from Thaler et al. [1997]. 36 subjects subdivided into 3 groups, one for each condition. The subjects interacted for 200 trials with a money-machine, by selecting an option and received a payoff during each trial. This data is analysed with a descriptive modelling approach on a subjective level. This method is chosen in order to estimate the best description of learning process and choice strategy for each subject in order to understand how these are influenced by other factors, such as the outcomes of the options chosen by the subjects and their variability.

The second part of this chapter is focused on a quasi-field scenario. Data consisting of a trading simulation from an online game is used to test some of the models developed in the first part of the work. The subjects interacted with the game with different levels of engagement, with the number of trials per subject ranging from 16 to 107. The subjects would buy some of the available stocks and proceed to sell them at a later point in time. In a similar fashion to the previous work, the data is analysed on a subjective basis, following a descriptive modelling approach.

4.1 Subjective value perception, payoff variability, myopia and performance

The experimental conditions adopted in Thaler et al. [1997] and Barron and Erev [2003], further examined in Erev and Barron [2005], are good scenarios in which to investigate

individual strategies. The aim of this work is to capture subjective strategies so to allow a subject-based investigation. This proposed methodology allows the testing of the hypotheses of this work, including whether myopia can be considered an explanation of deviations from maximisation, specifically in the context of a binary choice task with partial feedback such as the experimental conditions considered in chapter 3. The procedure adopted involves estimating the best fitting reinforcement learning model for each participant, considered to be the most representative of the choice set for each subject. This is achieved by determining the set of parameters best describing their behaviour. Each parameter in these models relates to a certain aspect of a subjects learning and decision-making behaviour. The parameters to be estimated with this method are the following:

α : the learning rate parameter, representing the speed with which a subject learns;

β : the inverse temperature, representing the greediness of a subject when picking one of the available actions;

γ : the discount factor, representing the far-sightedness of a subject.

The strategy of each subject can be therefore described by a set of parameters, which can be then used to test the hypotheses proposed. By evaluating subjects' strategies singularly it is possible to compare individuals' behaviour on a quantitative basis. This is a novel approach to analyse the data, as previous efforts were focused on population-wide descriptions even if large differences had been observed among the subjects.

4.1.1 The effect of previous outcomes on decision-making

The first hypothesis of this thesis is focused on the indication, provided by prospect theory, that individuals evaluate options according to a relative, non-fixed reference point (Kahneman and Tversky [1979]). In accordance with this view, a potential way to determine the current reference point is to consider previous experiences. This part of the research attempts at clarifying whether past decisions outcomes affect future choices. This concept is closely related to the "house-money effect" and "break-even effect", the increased risk-seeking attitudes of subjects after incurring respectively in gains or losses. According to behavioural economics, both these mental accounting phenomena could affect the behaviour of subjects in this type of task (Erev et al. [2008]; Hsu and Chow [2013]; Thaler and Johnson [1990]). The behavioural data from Barron and Erev [2003], which replicates the experimental conditions in Thaler et al. [1997] is analysed in this thesis, in order to clarify whether these phenomena could be affecting subjects in this decision-making tasks. To do so, a

descriptive modelling framework is adopted and systematically modified to include different environmental scenarios. When modelling a reinforcement learning task, the structure of the environment is designed to include the information which are believed to be used by the agents. Binary decision-making tasks are often modelled with descriptive learning algorithms focusing only on the quality of the available options, thus neglecting how the information about prior outcomes could affect the subjects in other ways. In previous works on the same data, the subjects are modelled as RL agents and are assumed to use their decisions outcomes to update their beliefs on which option is the best at any point in time (e.g. Barron and Erev [2003]; Erev and Barron [2005]).

Using Kahneman and Tversky's prospect theory shifting reference point as a foundation concept, this thesis proposes that previous outcomes are considered, not only to update choice preference, but also to assess the current state subjects consider themselves to be in. While in previous work the environment is considered to be state-less, the modelling proposed in this thesis involves expanding the state space to account for previous interaction information. There is an exception in Plonsky et al. [2015], where dynamic environments are analysed. In this work, outcomes depend on the current "state of nature" and the states are described by a Markov chain with transition probabilities unknown to the agent. Their definition of state-space is similar to the one developed in this thesis but it is substantially different because the transitions depend on probabilities defined by the authors. In the state-space proposed in this thesis instead, the transition probabilities are unknown and can only be estimated from the outcomes received, which are draws from Gaussian distributions.

Considering the binary decision task, and looking from a decision-maker point of view, previous outcomes information is provided by the money-machine. As shown in the example in Fig. 4.4, the money machine shows the amount of virtual money gained or lost with the last decision and keeps track of the total amount accrued to that point in time. This hypothesis assumes that decision-makers are influenced by either of these two values. Because two values are presented, this first hypothesis can be dissected into two tests. The first concerned with testing whether the full history of previous outcomes influences the strategy the subjects follow by shifting their reference for future actions. Similarly, the second investigation tests whether subjects' reference point is influenced by the latest-outcome. Both parts of this hypothesis will test a version of the enhanced state-space against the traditional single-state scenario. Three versions of the environment modelling are developed, each one representing a different type of mental accounting. In the baseline case, subjects are assumed to consider each choice solely on the value of the option learned over time. In this scenario subjects do not consider their situation to be dependent on payoff history. In contrast to this simplistic scenario, two enhanced configurations are proposed and developed. The first based on the

sum of previous outcomes, the second only on the latest one. The information for both is present in the graphical interface the subject used in the task (Fig. 4.4).

Both these setups encompass a two states environment, with the current state of the system being determined based on either of the two values presented by the interface. In the first case the subjects are assumed to determine their current state considering the entirety of previous outcomes they experienced, the sum of gains and losses resulting from their previous choices. Subjects have this information in the form of the sum of the payoffs obtained (Fig. 4.2, panel a, third row: "Total"). In the second modelling, only the latest outcome is used to determine the reference point of a subject. Subjects have access to this value from the money-machine (Fig. 4.2, panel a, second row: "Outcome").

These state-space models represent a potential link between reinforcement learning modelling framework with prospect theory's (PT) reference points. The ability of the reference system to shift and adapt to the psychophysical perception of an individual is encapsulated in the RL model by defining the state-space accordingly.

The baseline state-less environment together with the two binary-state spaces gives rise to three scenarios. These scenarios will be fitted to the data and their descriptive value estimated probabilistically with the maximum likelihood estimate methodology described previously. The results of this model fitting procedure will be used to compare the scenarios and test both parts of the first hypothesis. More detailed information about the structure of the models tested in this descriptive study is provided in modelling section 4.

4.1.2 The effect of subjective perception on option values

Descriptive modelling attempts have been made in the past with the intent to capture how people perceive and internalise the outcome of their decisions, or the prospects of the available options. Among the many models a explanatory account of decisions from description is Prospect Theory. Loss aversion (LA) is one of the most representative and celebrated ideas introduced by Prospect Theory. LA concept refers to the decision-makers' subjective perception of losses to be about twice as strong as gains of similar magnitude. The way this effect is captured by PT is by introducing a kink in the origin of the value function. The value function is a mapping of outcomes to the perceived value of that outcome and the origin is the point of reference from which outcomes are assessed as profitable or detrimental by a decision-maker. The subjective value function adopted by PT, portrayed in figure 4.1, is a transformation function which takes as input the objective value of a payoff and translates it into a value representing the perceived utility of that outcome. It is worth noting that positive outcomes will be always perceived as positive yet decreasing in subjective value and viceversa for the negative outcomes. This diminishing sensitivity to the magnitude of the

outcomes is encoded in PT value function. Empirical evidence showed people behave as if risk-averse in the gains domain while behaving as if risk-seeking in the losses, leading to the the S-shape of the function which describes the pattern of diminishing utility of gains and diminishing discomfort for losses.

In modelling studies a transformation function modifies the value of received payoff, mapping it to the corresponding subjective perceived outcome. The second hypothesis regards the subjective perception of the option values. This is tested by adopting different types of subjective function and studying the most representative of the subjects in this task. PT value function is a good candidate of reward function and represents a relatively sophisticated description of the subjects' internalisation of action values. The PT value function adopted in this thesis will leverage the parameters estimated in previous studies and rely on those so that the current work can focus on estimating the quantities of interest for the hypotheses at hand. Therefore, the parametrisation of the PT value function follows the parameters from Tversky and Kahneman [1992]: $\alpha_{PT} = \beta_{PT} = 0.88$ and $\lambda_{PT} = 2.25$. In order to test whether this is indeed an appropriate descriptive reward function, it will be tested against least sophisticated functions. The simplest reward function to be tested is the identity function, which does not modify the reward values and can be considered the baseline reward function as it is the least refined. The adoption of raw payoffs within the modelling of learning process and decision-making is still widely used in descriptive modelling studies (e.g. in Erev and Barron [2005]). Another interesting reward function to test is a transformation which fully saturates the value of payoffs. This can be achieved by adopting an horizontally asymptotic sigmoidal. This last transformation function represents an extreme version of decreasing utility, which for very high (or very low) values of payoff saturates to a certain value. Such function would be able to capture extreme desensitisation to gains and losses. The null hypothesis is that there is no difference between the three proposed reward functions. The expectation for the outcome of the test of this second hypothesis is that the prospect theory's subjective value function will be the transformation function best characterising the subjects in the dataset.

4.1.3 The effect of payoff variability on learning speed and choice randomness

In Myers and Sadler [1960] a "card flipping" task was used to study decisions from experience. In each trial, the participant had two cards, the first would be flipped and reveal a payoff (the safe option in their scenario). The subject would then either accept that payoff or risk flipping the second card. The paradigm adopted in their work was of minimal information:

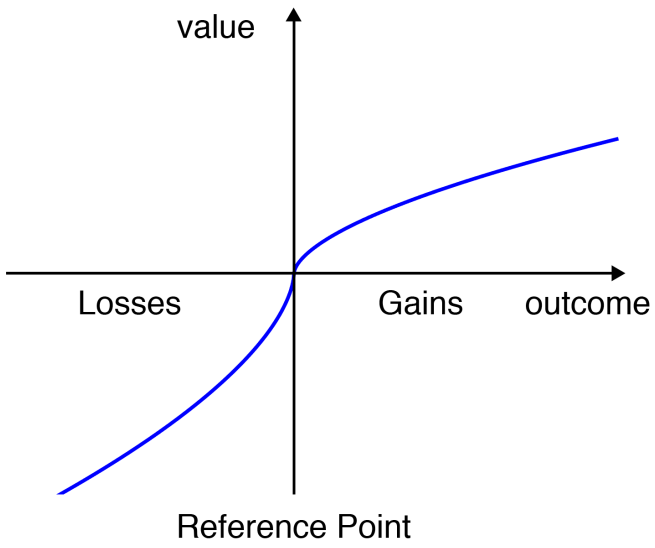

Fig. 4.1 An example of subjective value function from prospect theory, with the parameters estimated in Tversky and Kahneman [1992]: $\alpha_{PT} = \beta_{PT} = 0.88$ and $\lambda_{PT} = 2.25$.

subjects received information about the outcome of the cards flipped but were not shown the forgone payoff of the risky alternative, in the case the safe card was chosen. Their results showed that the proportion of maximisation choices was reduced when the variability of the risky option was increased (Myers and Sadler [1960]). Years later Busemeyer and Townsend [1993] named this phenomenon the "payoff variability effect", and along with other studies contributed to point out its robustness (Erev and Barron [2005]; Erev et al. [2012]; Haruvy et al. [2001]). The third and fourth hypotheses of this dissertation regard the payoff variability (PV) effect and its relationship on learning and decision-making.

Payoff variability is a measure of the fluctuations in the outcomes experienced by a subject. Higher PV can lead to random choice behaviour (Myers and Sadler [1960]), potentially originating from a slower learning effect (Erev and Barron [2005]). This can be ascribed to the fact that in highly uncertain environments, an otherwise reasonable exploration can be perceived as counterproductive, leading to purely random and non-exploratory behaviour (Erev and Barron [2005]). The relationships between variability in the observed outcomes and learning speed or decision-making randomness have not been tested systematically in previous works. This thesis proposes a methodology to test two hypotheses concerned with these two aspects of the learning and decision-making in the task at hand. The method proposed is focused on using those components of the descriptive models which capture and measure learning speed and choice randomness. Payoff variability is defined in Erev and Barron [2005] and reported in this thesis in Chapter 3, section 4, Eq. 3.9. In their

methodology the PV value is estimated on a trial-by-trial basis and serves as a scaling factor in the proposed model's action-selection policy. The estimation of PV in this thesis does not need to track its evolution in time, a final value summarising each subject's experience can be used instead. Therefore a simplified version of PV consisting in the standard deviation of the outcomes received by a subject over the course of the 200 trials will be adopted:

$$PV_{subj} = \sigma_{subj} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(o_i - \mu)^2} \tag{4.1}$$

where o_i is the outcome received during the $i-th$ trial, N is the number of trials a subject interacts with the money-machine and $\mu = \frac{1}{N}\sum_{i=1}^{N} o_i$ is the mean of the outcomes received by the subject.

The third hypothesis involves studying the relationship between payoff variability and speed of learning to clarify whether increased variability in the payoffs obtained by a subject could result in impaired learning; the indication from literature is that an increase in payoff variability leads to a slowing in the subjects learning process (Busemeyer and Townsend [1993]; Erev and Barron [2005]). To test this hypothesis the subset of models best fitting the choices is identified for each subject in the dataset. All models include a learning rule which features a parameter representing the speed of learning. This parameter quantifies the proportion of information subjects use when building their beliefs about action values. The learning speed parameter estimate of each subject paired with the corresponding payoff variability value will allow for a direct comparison of these two quantities. The correlation between the PV and the learning rate parameter estimate values, and the significance level of such relationship will be estimated in order to provide a statistically meaningful answer to the question of interest. The expectation is that the results will confirm previous investigations findings (Erev and Barron [2005]), that there is a significant negative correlation between the two quantities.

In a similar fashion, the fourth hypothesis focuses on the relationship between payoff variability and the subjects' action-selection strategy. More precisely, it investigates whether an increase in perceived payoff variability leads a subject to a more random choice behaviour. All the proposed models that will be fit to the data include a probabilistic action-selection policy, featuring a free-parameter that estimates the degree of randomness in a subject's choices set. This parameter is referred to as "inverse temperature" and the more random choice is in the data, the close this parameter estimate is to 0. Higher values instead, correspond to more exploitative behaviour, that is selecting the option which is believed to be the best at a particular time without exploring the alternative. This behavioural pattern is also

commonly referred to as "greedy". The expectation for this test is that there is a significant negative correlation between PV and the action-selection free-parameter, implying that when the obtained payoffs variability is high the choice behaviour is more erratic.

Two components of the modelling proposed in this thesis are of central importance for testing these two hypotheses. These were introduced in Chapter 2: the learning model and the probabilistic action-selection policy. The estimation of the free-parameters of interest is done following the model fitting procedure discussed earlier in this chapter, in section 4.1.14. The best fitting models for a subject are identified according to the Akaike weights methodology presented in subsection 4.1.15 and formally defined in eq. 4.18 and eq. 4.19. This setup allows to identify a single value for each free-parameter starting from the estimates of each model in the best models subset and weighting these by the amount of evidence in their favour. For both hypotheses the expectation is that payoff variability has a statistically significant ($p \leq 0.05$), negative ($R < 0$) correlation with both learning speed and choice. The implications of these hypotheses being confirmed will strengthen previous literature indications that higher uncertainty in the payoffs observed has an effect on speed of learning and decision-making.

4.1.4 The effect of myopic behaviour on task performance

The fifth hypothesis targets the relationship between the degree of far-sightedness in subjects' behaviour and their performance in the task. An intuitive measure of performance in the task studied can be identified in the total accrued rewards. Another measure for performance is the amount of time the best option has been selected, which directly influences the sum of payoffs. This is what has been used in previous work on the same dataset (Erev and Barron [2005]). Because these two are both meaningful ways to assess task performance they will both be used in this test.

Myopic behaviour can be described as choice behaviour which seeks short-term goals, disregarding the long-term implications of present-time choices. The opposite of myopic behaviour is far-sighted behaviour and in Thaler et al. [1997] terminology it is referred to as long-term framing relative to feedback frequency. Far-sighted individuals are not concerned with gaining the highest immediate reward but with maximising the potential long-term return, which sometimes includes exploration to discover better and previously unknown alternatives. The null hypothesis for this fifth investigation is that there is no correlation between the amount of far-sightedness subjects show and their task performance. The alternative hypothesis is that far-sighted subjects achieve higher performance for the task in analysis.

The cumulative rewards based performance measure is defined as the sum of outcomes accumulated during the 200 trials the subject interact with the money-machine:

$$CR = \sum_{i=1}^{N} o_i \tag{4.2}$$

where o_i is the outcome of the choice at the $i-th$ trial and $N = 200$ is the total number of trials. This measure represents a possible way subjects assess their performance: by using the value "total" provided by the money-machine interface, as in Fig. 3.1 in the previous chapter.

The myopic tendencies of the subjects are quantified by the value of a free-parameter in the learning rule adopted by the models proposed. Specifically, this free-parameter is featured in the Q-learning update rule and it can be viewed as a scaling factor for the term representing potential future rewards. It is not necessary to restrict the subjects set to only those who are best fit by a Q-learning model in order to test this hypothesis. All models can be considered in the estimation because the average-tracking rule, which is also referred to as immediate rewards learning, can be considered a particular case of Q-learning, in which the parameter regulating the value of future rewards is set to 0, effectively considering average-tracking a nested version of Q-learning. The Akaike weights methodology can be adopted directly to the results of each discount factor parameter value by considering this value for the nested model to be $\gamma = 0$ and keep it into account when evaluating the weighted-averaged single estimate used to test the hypothesis. The hypothesis testing method will be similar to the previous two, evaluating the correlation between the discount factor parameter estimate and the final accumulated payoffs for each subject. The expectation for this hypothesis is that there is a positive ($R > 0$) and significant (p-value ≤ 0.05) correlation between the discount factor parameter values of the subjects and their payoffs performance. In other words, the expectation is that subjects who are best fitted by short-sighted models achieve lower cumulative payoffs in the task; viceversa far-sighted subjects have better performances. This prediction can be interpreted as the following pattern: those subject who care more about their final score explore more and learn to identify the option which yields the best expected outcome (long-term reward) instead of focusing on the option which gives the lowest outcome (short-term reward) but requires less exploration to be fully identified.

The implication of this hypothesis holding true is a confirmation of the suggestions from previous work on myopic behaviour, helping to bridge the gap between the indication that framing and feedback frequency impair financial decision-making at the population level (Thaler et al. [1997]) and the evidence that subjective performance is indeed affected by far-sightedness in binary choice tasks at individual level. As noted in the previous chapter, the

idea that myopia influences performance in decision tasks has been put forward in literature by Benartzi and Thaler [1995] and Thaler et al. [1997]. The present descriptive modelling study of the relationship between the learning rule discount-factor, representing the degree to which each subject is concerned with future rewards on a subjective level is novel to the best of our knowledge.

4.1.5 Models details

The proposed models are based on the reinforcement learning framework, therefore this section begins by defining the components of the framework corresponding to the real life actors and features of the experiment. The details about the experiment setup are presented in chapter 3. It involves a repeated choice task on a binary option money machine, spanning 200 interactions. The objective of the task is to maximise the sum of payoffs, which will be reflected in the actual monetary payoff each subject received from the experimenter. Subjects do not know the underlying payoff distribution associated with the two buttons. They receive each payoff after selecting an option. This value is summed to their previous balance and presented together with the latest payoff. Forgone payoffs, the outcomes of the option not selected, are not shown. The experimental scenario can be modelled as a learning problem where the subject, or decision-maker, is the agent and the experimental design is the environment in which the agent/subject[1] operates. The environment of the model can be concretely derived from the observable features of the experiment.

4.1.6 Choices and Actions

The left and right buttons of the task interface shown in 4.2 (panel a, first row) can be mapped, in a straightforward fashion, to the actions the agent can take in the modelled world. The action space is therefore directly encoded from the two button choices to the two actions. Subjects do not have information about the quality of the options available as the buttons are unmarked and unlabelled, but for the sake of clarity in this thesis they will be referred to as $Left$ and *Right*. Each action is randomly associated to the high or low expected return distribution. There are three conditions of payoff distributions. The full experimental conditions are described in chapter 3 but will be summarised here for convenience. In the first condition the returns linked to option $High$ (the option with the highest expected returns) are drawn from the Gaussian distribution $\mathcal{N}(100, 354)$ while the *Low* option outcomes are draws from the Gaussian distribution $\mathcal{N}(25, 17.7)$ truncated at 0. In the second condition, both distributions are shifted of $+1200$ points and maintain the

[1] agent and subject are synonyms of decision-maker in this section

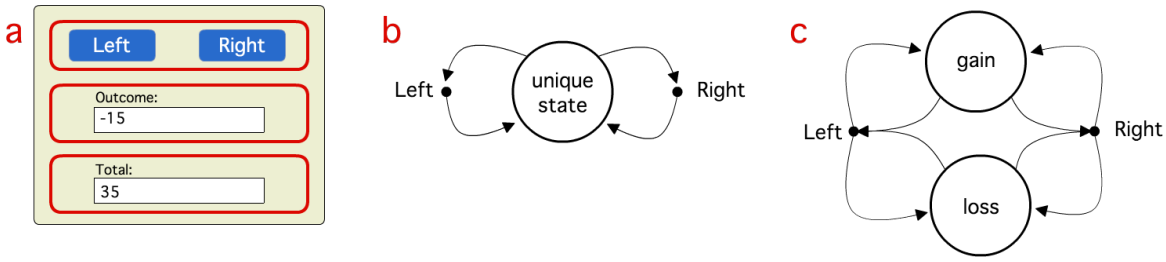

Fig. 4.2 Drawing a parallel between the actions and states of the experiment and the details and dynamics of the MDP modelling. Actions are portrayed with labelled filled dots and states with labelled circles. Panel a) The task graphical interface with the buttons (first row), latest outcome (second row) and sum of previous outcomes (third row). Panel b) The simplest state-space, referred to as "single-state", is modelled with an absorbing state where the agent is assumed to be no matter the outcome of his actions. In this scenario the agent is concerned with learning only the value of each action. Panel c) The two-state modelling can be defined either on the latest payoff or on the sum of all payoff received. An agent can remain in a state or move to the other one depending on the definition of the state-space and the outcome deriving from the choices made in the past. These two configurations are going to be compared to the single-state configuration to test the first hypothesis and determine whether subjects are influenced by previous outcomes when making decisions.

same standard deviations: $High \sim \mathcal{N}(1300, 354)$ and $Low \sim \mathcal{N}(1225, 17.7)$. The third and last condition, originally used to test payoff variability effects on the entire population, is characterised as $High \sim \mathcal{N}(1300, 17.7)$ and $Low \sim \mathcal{N}(1225, 17.7)$. This last configuration greatly reduces the uncertainty in decisions at the population level (Erev and Barron [2005]) and will help test the third and fourth hypotheses of this thesis, regarding the influence of payoff variability on learning speed and randomness of action-selection. An intuitive visualisation of this connection is shown in Fig. 4.2.

4.1.7 State-space

When modelling the environment of a problem, the states should include all the information required by an agent to make a decision at any point in time. This is the Markov property. In this thesis, the baseline model captures the simplest scenario, in which the agent is concerned only with the value of the actions without assessing the current state of the world. The learning problem is simply to learn which of the available actions is the best. This is referred to as single-state and can be represented by a trivial state space model assuming a dummy state, depicted in Fig. 4.2, panel b. This version of the model reflects the typical two-armed bandit problem setup adopted in literature. In order to test the first hypothesis, that

information about previous interactions influences decision-makers, the state-space needs to be enhanced.

There are two ways to define this enhanced state-space, each considering a different portion of the graphical interface. During the interaction (Fig. 4.4), the subjects are shown the payoff obtained from the last choice and the accumulated payoffs, which is updated after each choice. Both these quantities are directly observable by the subjects. The second row in Fig. 4.2 panel a) highlights the portion of the money machine presenting the outcome of the latest choice, while the third row of the same figure highlights the total accumulated payoffs. These two quantities are used to define the states-space of the model relating to each of the two parts of the first hypothesis.

As introduced in chapter 2, prospect theory (PT) suggests that the reference point from which individuals assess the outcome of their decision shifts subjectively, this property is also known as prospect theory's relative reference point assessment (Kahneman and Tversky [1979]). PT value function captures the subjective value of prospects based on a reference point which is not fixed on general terms but moves based on the subjective perception of these values. Integrating the ability of this reference point to shift into the learning model used to describe decision-makers' behaviour, constitutes the novelty in the approach proposed in this work. In an effort to integrate this feature of PT into the reinforcement learning framework, the modelling of the state-space will reflect the notion of changeable reference point. Plonsky et al. [2015] work on dynamic environments considers states configurations that depend on previous outcomes, focusing on the influence of the transition probabilities. Integrating the reference point information into the learning model through the state-space is, to the best of our knowledge, a novel idea that could help bridge the gap between prospect theory's shifting reference points, the reliance on recent samples observed in literature and the reinforcement learning computational modelling framework. There is virtually no upper bound to the amount of information that can be considered in the modelling process. Previous life experiences, temperature of the room, hunger or thirst of a subject, can all affect the state of a subject and the following decisions. According to the Markov property though, the representation of a real-world scenario with a modelled version needs to capture all and only the meaningful aspects of the learning problem. The house-money effect and the break-even effect are potential explanations of the reference point shifting in either direction, more or less risk-seeking depending on previous gains/losses. According to these phenomena, a subject whose wealth increased during the trials will consider the value of the available options in a different way than a subject who obtained poorer previous outcomes. Therefore, in the proposed model it is assumed that previous outcomes affect a subject's reference point, leading to the representation of previous outcome experiences with a binary state-space.

According to the two effects documented, a subject could become more risk-seeking in case he gains money because of not yet having factored in the gains as owned (house-money effect). In the opposite circumstance, if a subject loses money from previous interactions he will feel the pressure to regain the lost money at the cost of being more risk-seeking (break-even effect).

Two cases can be identified for this experiment-model translation: in the first the subject considers the sum of all previous outcomes as reference point; in the second case the subject uses the outcome from the previous decision as a reference point. The state-space defined on total accumulated payoffs shifts according to the sign of total accumulated rewards, starting from the initial wealth, considered to be 0 for all subjects. The other type of discretisation refers to a change of state based on the sign of the latest experienced outcome.

Indication that the latest transaction outcome could model decision-making well comes from experimental studies identifying recency to be a strong candidate to explain decision-making biases (Erev and Barron [2005], Hertwig et al. [2004]). In these works recent outcomes have been shown to be highly predictive of future choices and to be able to explain, together with undersampling, phenomena like the underweighting of rare events. Both these binary arrangements are also supported by the graphical representation of the task. In the interface of the task both these information are presented to the subjects, in the form of last choice outcome and total accumulated payoffs (Fig. 3.1 and 4.2). These modelling gives rise to three configurations: state-less, two-state based on full history and on latest outcome. The model fitting procedure will test each arrangement to determine the setups likelihood of correctly describing each subject. This will allow for the testing of the first hypothesis. For both binary scenarios, the state-space S of the experimental task is formally defined as:

$$S = \{gain, loss\} \tag{4.3}$$

where *gain* and *loss* depend either on the history of rewards or on the last reward. Let us consider the full history scenario, in which the sum of all previous outcomes defines the current state. The balance B_t is defined as the sum of the outcomes received up to the interaction at time t:

$$B_t = \sum_{i=0}^{t-1} o_i \tag{4.4}$$

where o_i is the outcome received during the $i-th$ interaction.

It follows that at time $t+1$, the agent can be in two situations:

$$state = \begin{cases} gain & \text{if } B_t \geq 0 \\ loss & \text{otherwise} \end{cases} \tag{4.5}$$

This information is provided by the task graphical interface, Fig. 4.2, panel a, third row. This is the first of the two states-space setups and it serves to test the first part of the first hypothesis, that individuals decision-making is influenced by information about the sum of their prior achievements. For the scenario in which the states are defined based on the latest outcome, the agent's state at time-step $t+1$ is defined as:

$$state = \begin{cases} gain & \text{if } o_t \geq 0 \\ loss & \text{otherwise} \end{cases} \tag{4.6}$$

meaning that the sign of the previous outcome determines the current state of the environment. The intuitive association of this version of the state-space based on the latest outcome to the information presented to the subjects in the task is portrayed in Fig. 4.2, panel a, second row.

These two methods to determine the current states, together with the simpler case in which there is only one state (two-armed bandit problem) represent three configurations of state-space to be examined. The Markov property holds true for all the configurations of states proposed in this work. In the single-state case, modelled with a single absorbing state, this is trivial as there is no history for the states, while information about actions and associated rewards are incorporated in the policy. Previous rewards obtained for each action are encoded in the option values the agent learns. In the full history two states scenario, the sum of accumulated payoffs determines the state in which the agent is. This configuration holds the Markovian property as well, because the balance B_t provides enough information to the subjects to make choices in the future in the same way the position of the pieces on a board of checkers allows a player to make a move. The dynamics of how the player or a subject got to the current state is lost but they both have sufficient information to make the next choice. In the relative two states scenario, instead, the Markov property is inherently satisfied, because the agent considers the latest outcome alone to determine the state.

Each combination of state, action and following state is associated with a transition probability and expected reward. An illustrative example considering a transition from the state *gain* to the state *loss* by taking action $Left$:

$$s = gain,\ s' = gain,\ a = Left,\ P^a_{ss'} = \rho,\ R^a_{ss'} = R^{Left} \tag{4.7}$$

where ρ is the probability of staying in state *gain* by taking action $Left$. The Markov Decision Process developed for this task is graphically represented in Fig. 4.2, panel c. The transition model for the single-state scenario (Fig. 4.2, panel b) is trivial, because for whichever action the agent will stay in the dummy state with probability 1. The transition function is not statically defined by the environment, unlike in Plonsky et al. [2015] where the probabilities for the triplets are predetermined. The transitions probabilities emerge from the scenario of state discretisation adopted and the outcome of each action. As a result, there are no predefined values of transition probabilities for each combination of state-action-state.

In the experimental design analysed in this work the MDP is finite as there is a finite number of interactions (200 trials), even if the subjects were not told that they will interact this precise number of times, which means that from the subjects' point of view the task can be considered as continuing. This characterisation of the task as continuing allows the use of temporal-difference methods, such as Q-learning. To reconcile this notion with the adoption of a finite MDP, it can be assumed that the final timestep $t = \infty$ is an absorbing state at the end of the experiment ($t = T = 200$).

4.1.8 Reward functions

A numeric signal is required for the agent to update the actions values. Generally, in reinforcement learning tasks the environment provides this signal directly. In this experimental task the signal can be identified to derive from each choice's outcome. The goal of the experimental task was to obtain the highest cumulative payoff. In the reinforcement learning description, the reward represents the computational counterpart of the payoffs the subjects seek to maximise in the task. The subjective reward is defined on the payoffs obtained after each choice. This is convenient as the reinforcement learning framework operates with real-valued signals and the decisions outcomes in the experiment are real-valued. These signals can be then transformed by means of a reward function which translates the observed raw payoffs into subjective utilities. The reward functions proposed in this thesis are used to test the second hypothesis; that people are loss averse and show diminishing sensitivity to gains and losses, instead of being insensitive to such effects. To test this hypothesis, three reward functions are adopted and fitted separately to the choice data. Details about the three proposed reward functions taken into consideration for this work, in order of complexity, are provided in the next three subsections.

Identity

The simplest reward function in exam is the identity function. It leaves the value and sign of the payoff unaltered and feeds this value as reward signal to the learning model. This can be considered the baseline for the modelling, representing an unrefined perception of each outcome as-is. People who show a linear appreciation of reward values will be best captured by this reward function as it does not scale nor it reduces the effect of increasing (or decreasing) values. It is also worth noting that it carries a high potential for numerical instabilities.

Hyperbolic tangent sigmoid

The second proposed reward function is the hyperbolic tangent function; this sigmoidal function possesses some interesting features. This function maps unbounded inputs to a range of values between -1 and $+1$ and is parametrised to change the slope of the S-shape. Both range and slope are adjustable and this is convenient because it allows to customise the function to the range of payoff values of the task. These features permit an easier handling of extreme values by squashing them into a narrower range, compared to the initial values, at the same time providing a handy way of capturing the variability of the payoff values in the new range, by adjusting the slope. The parametric hyperbolic tangent (Tanh) is defined as:

$$tanh(o_t) = \eta \cdot \frac{1 - e^{-o\omega}}{1 + e^{-o\omega}} \tag{4.8}$$

where o is the outcome received by a subject at any step, η is the minimum or maximum value after the transformation and ω is the parameter regulating the squash of this sigmoidal function. This is not a free parameter because it is kept constant over all the fitting procedures. The values adopted are $\omega = 1/500$ and $\eta = 50$.

Prospect theory's subjective value function

The third reward function to be tested is prospect theory's subjective value function, which translates the reward obtained from a decision to the subjective utility perceived for such reward (Kahneman and Tversky [1979]; Tversky and Kahneman [1992]). PT value function is characterised by three features. It is concave in the gain domain, convex and steeper in the loss domain, capturing the well documented diminishing sensitivity and loss aversion individuals show in decisions from descriptions (Fig. 4.1).

The parameters of PT's value function adopted in this thesis are the ones estimated and used in previous works (Benartzi and Thaler [1995]; Thaler et al. [1997]; Tversky and

Kahneman [1992]), and will be denoted with subscript PT to avoid ambiguity with other parameters in other parts of the modelling. These parameters are fixed and therefore they do not count towards the modelling fitting parameter search of this thesis. Adapting eq. 2.5 from chapter 2, the reward r obtained from outcome o at time t is:

$$r(o_t) = \begin{cases} o_t^{\alpha_{PT}} & \text{if } o_t \geq 0 \\ -\lambda_{PT}(-o_t)^{\beta_{PT}} & \text{if } o_t < 0 \end{cases} \tag{4.9}$$

with $\alpha_{PT} = \beta_{PT} = 0.88$ and $\lambda_{PT} = 2.25$ (Benartzi and Thaler [1995]; Thaler et al. [1997]; Tversky and Kahneman [1992]). These values of parameters produce the typical prospect theory value function graphically presented in chapter 2, Fig. 2.2 and reproduced here, for convenience in Fig. 4.1. This function depicts individuals presenting risk aversion for choices in the gain domain and risk seeking preferences for gambles in the loss domain. The parameter λ_{PT} captures the loss-aversion property, which makes a loss of a certain quantity perceived more than twice as bad as if the same quantity would be perceived as good, had it been a gain. Risk-aversion in the gain domain and risk-seeking preference in the loss domain are less pronounced, and sometimes reversed, in decisions from experience. This has been documented in Erev and Barron [2005]. In this regard, allowing three different reward functions, each one capturing a distinct mapping of payoffs to subjective values allows a more individualistic descriptive modelling. Prospect theory has another layer of pre-choice mental accounting which decision-makers are assumed to undergo when evaluating prospects. They are thought to apply a subjective probability weighting to the objective probabilities stated in the described choices. As this feature is only qualitatively described in Kahneman and Tversky [1979], but does not provide a numeric formulation, this would be hard to apply in the proposed framework. Therefore, this is proposed as a future development in chapter 6, along with more proposals deriving from the insights gained from this study.

Reward functions comparison

The three reward functions proposed are portrayed together in Fig. 4.3 for comparison. The identity function (red line) represents the simplest reward function, also adopted in Erev and Barron [2005]; it is therefore used as baseline for the comparisons. It makes no assumptions about the subjective perception of action values or about any other sort of effect (e.g. loss aversion or diminishing sensitivity) but it carries a risk of numerical instabilities. The hyperbolic tangent (green line) is a sigmoidal function that allows to capture the saturation of utility known in literature as diminishing marginal utility (Bernoulli [1954]; Dayan and Niv [2008]; Jensen [1967]; Kahneman and Tversky [1979]; Loewenstein et al. [2008]; Real

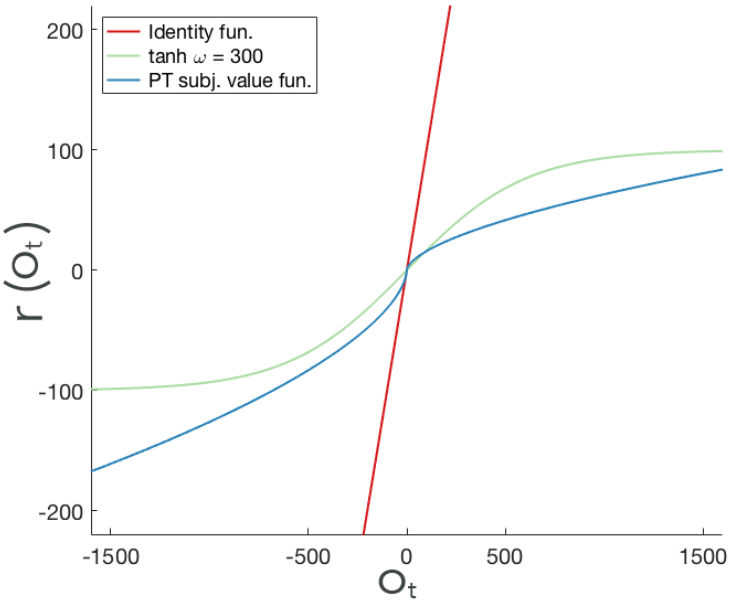

Fig. 4.3 The three reward transformation functions proposed. On the x-axis the original payoff obtained by a subject choice, on the y-axis the transformed subjective utility of that reward. The identity function (in red) leaves unaltered the reward value and does not saturate. The hyperbolic tangent function (in green) saturates, reducing the range of rewards. Prospect Theory's subjective value function (in blue) shows a monotonic trend with a steeper slope in the losses domain capturing both loss aversion and diminishing sensitivity (risk-aversion in the gains and risk-seeking in the losses domains).

[1991]; Sharpe [1964]; Tobler et al. [2007]). This sigmoidal function has, to the best of our knowledge, never been tested on choice data for similar tasks before. The most refined reward function tested is PT's subjective value function (blue line), which features the ability to capture loss-aversion by punishing outcomes in the negative domain more strongly than how equivalent outcomes in the positive domain are rewarded. This choice represents a potential link between the reinforcement learning framework and prospect theory, together with the state-space modelling representing PT's shifting reference points. Hertwig et al. [2006] already adopted this reward transformation when analysing choices from description. In their work, as well as in this thesis, the parameters of PT's value function are assumed to be the ones estimated in Tversky and Kahneman [1992]. These three reward functions are adopted to better capture the various subjective perceptions within the choice data analysed, allowing the second hypothesis to be tested.

An alternative to these functions, which are deterministic and in some cases parametrised is a stochastic reward function, which would transform the reward probabilistically. In prospect theory, for example, the value of the available choices is estimated as a combination

of the subjective value and a probabilistic weighting, denoted with $p_{PT}(p)$, as shown in eq.2.6. This extension is not implemented in the current work since a probabilistic action-selection policy is already in place for some of the models to be tested. The details of this component will be presented later in this chapter, in section 4.1.11.

4.1.9 Learning Models: Average Tracking and Q-learning

The adaptive learning process by which the subjects learn which option provides the best outcomes and leads to the highest overall payoff is modelled computationally with the following two learning rules. A simple average tracking rule which considers immediate rewards only, and Q-Learning as a more advanced learning model. The first rule is equivalent to the "fast best reply" cognitive strategy in Erev and Barron [2005], and is formally defined as:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} - Q(s_t, a_t)] \tag{4.10}$$

where $Q(s_t, a_t)$ is the value of the option/action selected at time-step t, r_{t+1} is the outcome of that action and α is a step-size parameter. The term in the square brackets represents the error between the target (reward obtained from an action) and the previous estimate (expectation of the reward obtainable). This quantity is scaled by α which can be considered to tune the quantity of information of the error that gets propagated in the current belief of the agent. According to this simple learning rule, the agent updates the estimate of the available actions values directly with the numeric signal deriving from the outcomes of the interactions.

Q-learning is a more advanced learning model which extends the previous method by introducing a temporal difference term, an estimate of the return (sum of rewards) that can be obtained choosing the best actions from the following time-step onwards. This term is scaled by a value, termed discount factor and denoted with γ, which regulates the influence of future rewards on learning and can be used as a measure of the myopic tendencies of each participant. Q-learning, as delineated in chapter 2, section 2.25, is widely adopted in literature because of a large body of evidence describing it as the computational account of the neural processing counterpart in the brain. This accountability together with a dedicated parameter which selectively tunes the temporal difference influence on the learning are the reasons underlying this modelling decision. Moreover, the structure of the state-space proposed, with previous outcomes modifying the current state a subject is considered to be in, represents an attempt to link the house-money effect and the break-even effect to this descriptive account of behaviour. Q-learning is also proposed to account for this linking,

since it allows agents to take into account future rewards when making decisions in a more complex binary state-space. For convenience of reading the Q-learning learning rule is repeated here:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

The simple model represents a baseline for testing as it has been used satisfactorily in previous work. The Q-learning model is equivalent to the average tracking method in the case $\gamma = 0$, therefore the baseline model can be considered nested within Q-learning update. The average tracking method represents a simple learning procedure, the agent updates the internal action values using new evidence and the pace of this process is conveniently governed by a parameter. This simple approach, though, lacks the power to capture a potential adaptive learning component, which is the concern with possible future outcomes. Hence the need to introduce an enhanced model with a dedicated quantity designated to represent this concern.

As for many other real-world tasks it is not necessary to learn the full state transitions model as this is often computationally intensive, and intractable in certain cases. Even if the computational cost of building an environment model for the experiment in analysis is negligible, previous literature indicated that for such simple task individuals tend to rely on faster brain computation circuitry (Daw et al. [2005]). The rationale behind this choice is that model-free reinforcement learning techniques provide a well studied account of brain areas such as dopaminergic neurons and their dorsolateral striatal projections (Daw et al. [2005]; Doya [2007]; Houk and Wise [1995]; Schultz et al. [1997]).

4.1.10 Initialisation

Considering the modelling of the environment comprising two states and two actions, their combination gives rise to a two-by-two matrix storing the agents beliefs about the value of each state-action pair. These values will be initialised to 0 because the experimental design did not give any indication to the subjects of the range of values the rewards can take. It cannot be assumed therefore that individuals begin their task interaction with high or low hopes, respectively optimistic or sceptic initialisation. Another potential way of setting the initial values is to set them to a high value. This is called optimistic initialisation (Sutton and Barto [1998]) and is generally used to encourage exploration. These high initial values are readily adjusted towards their real values after the agent first interactions but have the effect of pushing an agent to "believe" actions are better than they are. Optimistic values are a method to improve performance in a reinforcement learning problem. The main objective

of this thesis is to develop and test descriptive models to better understand the underlying decision-making processes, therefore this method of initialisation will not be used.

4.1.11 Action-selection: Soft-Max

Subjects in the experiment face a exploration-exploitation trade-off problem when choosing between the two options. After collecting a certain amount of information, the subject knowledge about which option is best to pick might be biased by an unlucky streak of unfavourable outcomes. A subject might think they possess enough information to make choices in the future, that is exploiting the knowledge gained. But to gain a better understanding of which option leads to the best payoff, it might be necessary to explore by sampling the choice believed to yield the poorer payoff. Undersampling has been proposed to be the cause leading subjects to a biased learning of which option is better, in turn leading to suboptimal behaviour (deviation from maximisation) (Barron and Erev [2003]; Erev and Barron [2005]; Erev et al. [2008]; Hau et al. [2008]; Hertwig and Erev [2009]; Rakow et al. [2008]). To capture the degree to which each subject is more or less prone to exploration the learning model needs to be combined with a probabilistic action-selection model. The proposed policy for action-selection is Soft-Max (chapter 2, section 2.2.1, eq. 2.24). An example of the formulation of this rule adapted to the model proposed, to calculate the probability of picking action $Left$ is:

$$P(Left) = \frac{e^{Q(s,Left)\cdot\beta}}{\sum_{a\in A(s)} e^{Q(s,a)\cdot\beta}} \tag{4.11}$$

where $P(Left)$ is the probability of selecting the action with the highest value (not known to the subjects), e is the exponential function, $Q(s,Left)$ is the value of taking action $Left$ when the agent is in state s, $A(s)$ is the set of actions available when the agent is in state s, and β is the inverse temperature free-parameter which regulates the greediness of the strategy.

This parametric, probability based rule translates the subjective value of actions into a probability of the actions being selected by the agent. The parameter β allows this descriptive model to fit the behaviour of the singular subject, identifying a quantitative representation of the degree or exploration, which can also be read in terms of randomness. Another way of reading this dimension continuum is the amount of sampling achieved by a subject, with undersampling behaviour being linked to exploitative individuals.

At this level of detail in the modelling of the task, it is assumed that the level of abstraction is articulated enough to capture the preferences for the available options expressed by the subjects, in a probabilistic fashion. In future extensions of this work it would be interesting

to construct a more complex version of this part of the system, which could factor in the probability weighting function presented by prospect theory and its improvements (Kahneman and Tversky [1979]; Tversky and Kahneman [1992]).

4.1.12 Models Summary

The combinations of learning rules, reward functions and state-space setups proposed gives rise to a number of arrangements to be fitted and assessed against each other and in relation to previous modelling efforts. These are summarised in table 4.2. Model number 1 (Single-state with Average Tracking and using the raw payoffs as reward signal) is similar to the two models tested in Erev and Barron [2005], "slow best reply" and "fast best reply" which have been defined and analysed in Chapter 3, section 4. This represents the starting point for the tests and will help determine which model best captures the individual subjects behavioural data. These procedures are explained in detail in the next section.

States Space		
Single-state	**SS**	$S = \phi$
Two-state (latest outcome)	**LO**	$S = \{gain, loss\};\ s = \begin{cases} gain & \text{if } B_t \geq 0 \\ loss & \text{otherwise} \end{cases}$
Two-state (full history)	**FH**	$S = \{gain, loss\};\ s = \begin{cases} gain & \text{if } o_t \geq 0 \\ loss & \text{otherwise} \end{cases}$
Learning Rule		
Average Tracking	**AT**	$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} - Q(s_t, a_t)]$
Q-learning	**QL**	$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$
Reward Function		
Payoff (raw)	**ID**	$r(o_t) = o_t$
Hyperbolic tangent	**TH**	$r(o_t) = tanh(o_t) = \eta \cdot \frac{1-e^{-o\omega}}{1+e^{-o\omega}}$
PT subjective value function	**PT**	$r(o_t) = \begin{cases} o_t^{\alpha_{PT}} & \text{if } o_t \geq 0 \\ -\lambda_{PT}(-o_t)^{\beta_{PT}} & \text{if } o_t < 0 \end{cases}$

Table 4.1 Models components, their abbreviations and their mathematical formulations.

	States Space	Learning Rule	Reward Function	Parameters
1	SS	AT	ID	α, β
2	SS	AT	TH	α, β
3	SS	AT	PT	α, β
4	LO	AT	ID	α, β
5	LO	AT	TH	α, β
6	LO	AT	PT	α, β
7	LO	QL	ID	α, β, γ
8	LO	QL	TH	α, β, γ
9	LO	QL	PT	α, β, γ
10	FH	AT	ID	α, β
11	FH	AT	TH	α, β
12	FH	AT	PT	α, β
13	FH	QL	ID	α, β, γ
14	FH	QL	TH	α, β, γ
15	FH	QL	PT	α, β, γ

Table 4.2 Models summary. Combinations of state-space, learning rule and reward function. The action selection is soft-max for all the models proposed.

4.1.13 Fitting procedure

To test the hypotheses proposed it is necessary to determine which model best describes each individual's strategy, this is achieved by means of a widely adopted model fitting method, known as Maximum Likelihood Estimation. This probabilistic approach evaluates models with different sets of parameters against the data, and estimates the likelihood of each model to be the one generating the data.

4.1.14 Maximum Likelihood Estimate

The Maximum Likelihood Estimate method is widely used in literature to assess how likely a model is to fit a specific dataset. The formalisation of this approach adopted in this thesis follows the description given in Daw [2009]. This method is a good choice because it is based on a probabilistic approach which works well with the probability based modelling developed in this thesis (e.g. Soft-Max). The fitting procedure results in a numerical value, termed the maximum likelihood estimate (MLE), that summarises how likely a model is to generate the data. In the context of the current modelling effort, the free parameters of the models to be tested describe quantitatively the subjects' learning and decision-making strategies. To

evaluate the parameters that best represent a subject's behaviour, a series of models are fitted to the data and their quality is described by their MLE. The procedure for a general case is formalised as follows. Let M be the proposed model with θ_M, the associated vector of free parameters and D the data set, it is possible to compute the probability distribution over possible data sets D as $P(D|M,\theta_M)$. Then, by means of the Bayes' rule:

$$P(\theta_M \mid D,M) \propto P(D \mid M,\theta_M) \cdot P(\theta_M \mid M) \tag{4.12}$$

The term on the left hand side of the proportionality is the posterior probability distribution over the vector of free parameters given the data and the model. This quantity is proportional to the likelihood function multiplied by the prior probability of the parameters. As there is no prior knowledge of which parameters values best fit the data, the last term can be considered uninfluential in the proportionality. Therefore, by maximising the likelihood of the data given the parameters it is possible to know which parameters best fit the data (the posterior probability). The most probable values of θ_M is the maximum likelihood estimate and is denoted with $\hat{\theta}_M$. In the case of a dataset comprising choices, the MLE value linked to a specific vector of free parameters is calculated by iteratively updating the actions' values based on the outcomes of the decisions using the learning model in analysis. Subsequently, these values are used to estimate the probability of picking each of the two options. The probability of the entire dataset is the product of the probability of each choice $c_t = c_1, \cdots, c_T$. Supposing $Left$ and $Right$ are two available actions, the probability of each choice c_t is calculated with the Soft-Max rule (eq. 2.24). These values are then multiplied as follows:

$$\prod_t P(c_t = Left|Q_t(Left), Q_t(Right)) \tag{4.13}$$

with c_t, the choice at time-step t, being action $Left$ as in the example in Fig. 4.4.

The calculation of this product can lead to computational underflow issues due to the multiplication of probability values in the range $(0,1)$. To obviate this potential drawback it is possible to take the logarithm of this quantity and compute the sum of the logarithms of the probabilities instead:

$$\log \prod_t P(c_t = high|Q_t(high), Q_t(low)) = \sum_t logP(c_t = high|Q_t(high), Q_t(low)) \tag{4.14}$$

where log is the natural logarithm function (base e). This operation is allowed as the logarithm is a monotonic function which does not alter the relationship between two estimates. A value in the range $(0,1)$ is simply translated to a value in the range $(-\infty, 0)$

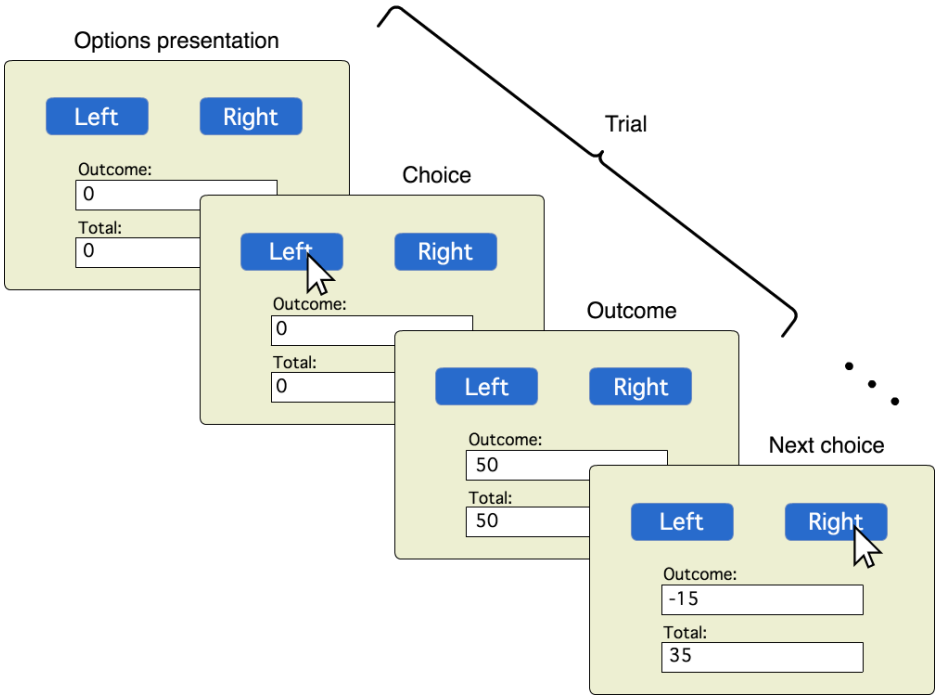

Fig. 4.4 Example of interaction with the money-machine game.

maintaining the magnitude relationships among values. The lower limit is unlikely to be reached in this specific case, as it is uncommon for options to have probability of being chosen $P(c_t) = 0$. The result of this transformation is called log likelihood; traditionally the negative of this value (negative log likelihood) is used instead, making the objective of the fitting procedure the minimisation of this quantity. The aim is then to minimise this positive estimate to best fit the data. This reversal is convenient in the context of the fitting procedure, because the optimisation routine is based on minimisation techniques that use gradient descent algorithms to estimate the parameter vector best fitting the data. The MLE, denoted as $\hat{\theta}_M$, is the set of parameters that maximises the likelihood of the observed data being generated by the model. Let $\hat{\theta}_p$ be the maximum likelihood estimate for subject $p \in \mathscr{P}$ where $\mathscr{P}$ is the set of subjects in the dataset. The best model for each subject is estimated individually by minimising the MLE among the set of potential parameter vectors. The models evaluated for each subject represent a descriptive account of the individuals strategies, captured by the explanatory quantities of the estimated parameters. This technique results in a characterisation of decision-makers behaviour that allows for further subjective analysis and hypotheses testing.

From a practical standpoint, applying this methodology consists in multiplying the logarithm of the probabilities of the actions selected by subjects at each time-step. Then using the outcome associated with the action selected at the current time-step, together with

the learning rule, to update the value of the state-action pairs, which in turn are used to update the probability of the actions using the soft-max rule. This dual update is iterated a number of times equal to the trials of interaction of the subjects.

To understand whether the models proposed in this work are a good descriptive account of the behavioural data they will be compared to a baseline random model. This comparison is necessary since the models proposed are all variation of a reinforcement learning model and the MLE score for each one of these is a comparative measure of a model fitness to the data, but does not indicate how good it is on an absolute scale. The MLE for the random model is denoted with (MLE_{Rand}) and is estimated in the same way as the rest of the models with the exception that the model offers a 50-50 chance of selecting either of the two available actions. The probability of picking an action is $P_a = 0.5$ and there are 200 trials, therefore the MLE of the model is defined as:

$$MLE_{Rand} = \log \prod_{t}^{N} P(c_t) = \sum_{t}^{N} \log P(c_t) = N \cdot \log(0.5) = -138.63 \tag{4.15}$$

where $N = 200$ is the number of trials and $P(c_t) = 0.5$ is the probability of the model selecting each option. This leads to a value of $MLE_{Rand} = -138.63$ which represents the baseline MLE score: models which present an MLE higher than this can be considered better than random. Because MLE on its own does not account for model complexity the significance of the model comparison with the random baseline model is tested by means of Akaike Information Criterion (AIC; Akaike [1974]), a complexity penalising method compensating the Likelihood of models which include different numbers of free-parameters. This methodology will be adopted for hypothesis testing, therefore its details will be extensively discussed in the following section.

4.1.15 Model comparison

Two model-selection criteria are used in literature to assess which model best describes the data: the Akaike Information Criterion (AIC; Akaike [1974]) and Bayesian Information Criterion (BIC; Schwarz [1978]). These scores are used to measure the model fitness to the data and provide methods of hypothesis testing that are extensively used in literature (Burnham and Anderson [2002, 2004]; Iigaya et al. [2016]; Lau and Glimcher [2005]; Symonds and Moussalli [2011]). The MLE values alone cannot be used to determine the best fitting model because they do not provide a parsimonious approach to model selection, and are known to favour overly complicated models (Gelfand and Dey [1994]). Conversely, AIC and BIC allow the comparison of non-nested models, keeping into account the number of

parameters and penalising overly complex models (Burnham and Anderson [2002]; Raftery [1995]). The AIC score of a model M is calculated as:

$$AIC = 2 \cdot \kappa - 2 \cdot \log(\hat{\theta}_M) \tag{4.16}$$

where $\log(\hat{\theta}_M)$ is the maximum likelihood estimate, calculated as described in section 4.1.14 and κ is the number of free parameters of the model. The BIC score of a model M is calculated as:

$$BIC = \kappa \cdot \log(n) - 2 \cdot \log(\hat{\theta}_M) \tag{4.17}$$

where n is the number of choices made by the subject. The method to determine which of the models best describes the data is by evaluating the absolute difference of the AIC or BIC associated with each model (Burnham and Anderson [2002]; Raftery [1995]). The greater this difference, the stronger the evidence against the model with higher AIC/BIC. A difference of 2 is considered to show weak support while a difference of 6 or higher corresponds to strong evidence in favour of the model with the lower AIC/BIC (Burnham and Anderson [2002, 2004]; Kass and Raftery [1995]; Raftery [1995]).

There has been debate in literature on whether to use AIC or BIC for model selection. None of the two criteria is inherently better than the other and the difference is more philosophical than practical (Burnham and Anderson [2002, 2004]). The case for BIC being favoured over AIC is that it is grounded in Bayesian inference (Kass and Raftery [1995]; Raftery [1995]). Burnham and Anderson [2002] and Burnham and Anderson [2004] pointed out that the difference between the two approaches concerns only the prior distribution over the model set, which represents the initial belief of which model is more likely to be the one better representing reality. Moreover, it is also shown that both criteria can be derived under a frequentist, non-Bayesian procedure. The case for choosing one or the other is therefore strongly connected to the context in which the model selection is carried out. From a practical point of view, BIC applies a stronger penalisation to the number of parameters of a model, which grows logarithmically with the number of samples. BIC penalty is guaranteed to select the true model, if this is included in the set of models compared, as the sample size grows infinitely (Vrieze [2012]). AIC instead does not assume the true model is among the ones analysed. Being grounded in information theory, specifically on the Kullback-Leibler (K-L) divergence, AIC would intrinsically estimate the distance between the true model and the proposed models; being the true model generally unknown, such comparison is not possible but AIC can be used to rank the candidate models according to their relative differences (Vrieze [2012]). AIC is known to perform badly when the number of parameters κ is high

compared to the number of samples n. In such cases a corrected form of AIC is used, denoted with AIC_c. The fitting procedure performed in this work is carried out over a sample size $n = 200$ with a small set of parameters $\kappa = 2,3$. As the number of parameters is lower than the number of samples used for the estimation by a factor of 40 (Burnham and Anderson [2002]; Symonds and Moussalli [2011]; Wagenmakers and Farrell [2004]), the standard AIC formulation is adequate and will be used. The different nature and the distinct assumptions of each criterion make it hard to readily decide which of the two should be adopted: BIC is guaranteed to select the true model as n approaches infinity, as long as the true model is among the candidates, which is rarely the case; AIC is preferred in scenarios in which the underlying process generating the data is simple but the risk of selecting a model with higher complexity than the true model is high (Aho et al. [2014]). Other factors contributing to the preference of a criterion over the other one include whether the research has an exploratory vs confirmatory reason, whether the candidate models are more or less known, and whether the goal is to make predictions or to find the true model (Aho et al. [2014]).

A reasonable method for model selection based on AIC criterion is the use of Akaike weights. Adopting Akaike weights avoids bluntly disregarding models which are potentially good descriptive candidates, only because their AIC score is slightly higher than the best. Akaike weights are based on the likelihood of a model, given the data and the set of models R (Burnham and Anderson [2002] p.75). They can be interpreted as the quantity of evidence in favour of a model, representing the conditional probabilities for each model (Wagenmakers and Farrell [2004]). Akaike weights are calculated as follows:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta_r)} \tag{4.18}$$

where Δ_i is the AIC difference of model i and the best model (the model with the lowest AIC score). Two AIC based model selection criteria often used in literature are to either select the single best model or to consider the models subset with AIC score within 2 units of the best model, then proceed with a case-by-case inspection to determine potential overfitting due to a larger number of parameters (Burnham and Anderson [2002] p.131 and Wagenmakers and Farrell [2004]). Akaike weights allow to overcome this inconvenience by attributing a representativeness factor to each model. Considering two or more models which barely differ in their descriptive power as quantified by AIC scores, i.e. when $\Delta_i \leq 2$, to base inference on the parameters estimates of the single best model is likely to misrepresent the true parameter value (Burnham and Anderson [2002] p.150).

The Akaike weights can be used to compute a weighted average of the parameters of interest:

$$\hat{\theta} = \sum_{i=1}^{R} w_i \theta_i \tag{4.19}$$

where θ_i is one of the parameters of the model. The model averaged estimate of a parameter is the weighted average of the estimates up to the point where the sum of weights for the most likely models is ≥ 0.95, representing a 95% confidence set; by doing so the Akaike weight are construed as a posterior probability, given the data and the set of a priori models (Burnham and Anderson [2002] p. 169). This procedure is grounded in the Kullback-Leibler information theory and represents a best model confidence set based on the data, in a similar way to a parameter estimation of a confidence interval based on the model and the data (Burnham and Anderson [2002], p. 169). This weighting scheme also provides a convenient method to identify a single estimate for each parameter, as these will be needed to test the next hypotheses.

Having laid out the methodology that will be used for fitting and comparing models, this first hypothesis section is wrapped by offering the according predictions. The expectation is that enhanced state-spaces will better represent the choice data examined. Moreover, the state-space modelling based on the full history of previous outcomes is expected to be the most likely description of the reference point system adopted for most of the subjects. The comparison between reference point based on complete historical information and latest outcome alone is expected to also be significant, with the former being the most descriptive. If these expectations are confirmed by the results, this work will represent evidence of the possibility of connection between the reinforcement learning descriptive framework and prospect theory's relative reference points. In case the state-less solution is the most likely descriptive account of the data, the assumption that individuals use previous interactions information when making decision is wrong and this work would provide evidence that decision-makers in binary decision tasks follow a simplistic approach and do not make use of past information to shift their reference system.

4.1.16 Predictive value comparison with previous results

While the hypotheses to be tested in this chapter are focused on descriptive quantities regarding the subjects behaviour, it is also interesting to compare the predictive power of the models proposed in this thesis with the RELACS model developed in previous work on the same choice data (Erev and Barron [2005]). The procedures based on prediction accuracy measurements used in the RELACS paper will be adopted in this thesis to achieve

an informative comparison. The predictiveness of the model will be estimated in the same way as in the original RELACS work (Erev and Barron [2005]). Simulations will be run 200 times for each proposed model and for each subject. The search procedure adopted, consistent with the one adopted in Erev and Barron [2005], is to discretise the space of parameters and use the subset of values to perform a grid search on the permutations of such values. The free parameters vector is composed of the combinations of a set of values. Specifically, the range of values that the free parameters can take is:

$$\alpha, \beta, \gamma \in \{0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.5, 0.7\}.$$

This procedure will produce a set of choices for each simulation run; these will then be split in two blocks dividing the number of interactions evenly. The choices within each block will be aggregated and contrasted with the observed choice data. To do so in a comparable way with the original paper the proportion of maximisation choices Pmax is estimated for each subject as described in eq. 3.1 in the previous section. The mean squared deviation (MSD; or mean squared error, MSE) of the predicted and observed Pmax for each block is adopted as measure of prediction. For every subject in each condition, the MSD score associated with a tested model M and its relative set of parameters θ_M is defined as:

$$MSD_M = \frac{1}{2} \sum_{b=1}^{2} (Predicted\ Pmax_b - Observed\ Pmax_b)^2 \tag{4.20}$$

where b is the Pmax block. The best model according to MSD score is selected and and compared. The MSD scores presented in Erev and Barron [2005] are cumulative; they result from averaging the MSD aggregating over the problems analysed. The MSD score for the model proposed will be calculated cumulatively, averaging the best score for each participant, in order to be compared to the score presented in the original paper. It is not possible to directly compare the two measures because the score presented in their work spans 40 different problems, of which only a subset has been analysed in this thesis. Moreover the parameter fitting procedures are substantially different. This work intentionally aimed at a descriptive modelling on an individual level, while in Erev and Barron [2005] the parameters have been fitted on the aggregated dataset of subjects and experiments, focusing on a population-wide characterisation. Nevertheless, this comparison can give a qualitative insight into how good the models proposed in this work are for prediction when compared to the RELACS. The following section will describe the details about the modelling efforts proposed in this thesis along with the rationale behind the implementation choices. The last section will present the results of the hypotheses tests along with considerations and potential explanations.

4.2 Online Investment Game: Virtual Trader

Modelling Stock-market Investors as Reinforcement Learning Agents

The game examined is an online trading simulation called Virtual Trader[2]. The platform is publicly accessible and managed by IEX Media Group BV in the Netherlands; it allows players to register for free and participate in the trading simulation game. The players who subscribe to the game are endowed with an initial balance of GBP 100,000 and can trade the stocks featured in the Financial Times Stock Exchange 100 Index (FTSE100, pronounced "footsie") pool. During the time range when data was collected, 107 stocks were available to trade in the game. Players would compose a portfolio (i.e. a collection of assets) by selecting stocks which are believed to be likely to appreciate in value in the future. Players are ranked on the cumulative assets, which are composed of "holdings" and "cash". The holdings are the shares (i.e. stocks) owned by a player, which can increase in value following stocks' rise in price or, conversely, lose value when the stocks' price plummets. The cash component of a player's assets does not change in value over time; it consists of either the amount of money which was never invested or the result of shares being sold.

This game simulation follows the evolution of real world stock data, including the fluctuations of stock's prices and their price splits or reverse splits[3]. These events are reproduced in the simulation with a slight delay of about 10-15 minutes; this delay does not result in information advantages to players following real world data, as the Virtual Trader orders are fulfilled taking into account the delay. Players are provided with an interactive graphical interface (Fig. 4.5), which features a time series for each stock with different time resolutions.

The transactions made by the players are recorded and can be publicly accessed by browsing the website. The dataset adopted for this study was generated by automatically crawling the transaction web pages. The transactions considered are the ones that happened in the time-range between the 1st of January 2014 and the 31st of May of the same year. The reason for choosing this specific time-range is because during these months the Virtual Trader game organised a prize give away. Players' portfolios were ranked according to their cumulative assets and the best achiever for each month was gifted with men grooming products, as shown in Fig.4.6.

[2]http://www.virtualtrader.co.uk - Copyright IEX Media Group BV

[3]A company can decide to issue more shares, simultaneously reducing their price, in order to, for example, allow more liquidity in the market. For instance, a 2-for-1 split would double the number of a company's shares in the market, halving their price. This procedure affects all the outstanding shares, including the ones owned by investors.

Fig. 4.5 The interactive graphical interface for the Virtual Trader online game simulation. This example shows the price time-series for the Coca-Cola HBC stock. The online website offers an interactive interface: moving the mouse cursor along the x-axis locks the pointer on the curve; the cursor's position, identified by the blue dot, shows the date and price information for the stock. In this example, the 30th December 2013 Coca-Cola stocks were priced at GBP 1,747.00. Different time resolutions are available by selecting the relative time-frame in the Zoom tab on the top left corner of the interface. The options on the top right corner of the interface instead, allow a player to swap between visualising prices during the current trading day or on longer historical periods, like the one shown of 1 year.

The structure of the prize system in Virtual Trader allows for two possible rewards to be identified: the first being the psychological reward of being ranked among the top players, the second being the tangible prize awarded to the winner of the competition each month (i.e. perfumes, etc). These rewards are the objective the players are supposed to aim at during the interaction with the game. Hence, in this attempt to map this task to a reinforcement learning process, the players are represented by RL agents.

The transactions have been stored in a database so that they can be preprocessed and later used to fit models of learning and decision-making. Each transaction is stored as a row composed of 6 fields: the *date* of the transaction, the *type* (i.e. sell/buy), the *name* of the stock traded, the *volume* and unitary *price* of the stock traded and *total* consisting of the money involved in the transaction. The dataset initially comprised $\sim 100,000$ transactions and 3381 players, but these numbers were greatly reduced after preprocessing. In fact, after removing the many inactive players and those players who engaged only at the beginning and/or at the end of the time-frame considered, the dataset featured 1420 transactions and 46 players. The average amount of transactions per player is 30.87, with the most engaging player having produced 107 transactions and the least active having made 16 transactions. This variability in number of transactions represents one of the key differences between this

quasi-field study from a laboratory experiment. The distribution of transactions made by the players is shown in Fig. 4.7

4.2.1 Hypotheses and testing methods

Among the many instances of human decision-making tasks, stock-market trading is a complicated and interesting one. The objective for financial-market's investors is to increase their wealth, by investing in shares of a company when their belief is that the company's stocks will increase in value over time. Investors would then proceed to sell their stocks when these are believed to have reached their peak. The famous sentence "buy low, sell high" represents this type of investing strategy. Investors are characterised by all sorts of degrees of competence and different strategies adopted. Researchers in the fields of economics and psychology focused on studying the behaviour of investors, as described in chapter 2, section 2.1.1, 2.1.2 and 2.1.2. In this part of the thesis the focus is on a set of subjects who played the Virtual Trader online game, described in the previous section. These players are decision-makers within the stock-market simulation offered by the game. Even if the players are anonymous and there is no certain way to know their level of knowledge of the markets or investment experience, the Virtual Trader game presents an interesting structure and a promising scenario, on which to test the following assumptions.

Some findings from both economics literature, including Choi et al. [2009] and Huang [2012], indicate that non-professional investors appear to make financial choices which are correlated to the outcomes of their previous choices. Choi et al. [2009] studied American workers' behaviour in a real-life scenario; specifically, their commitment to savings investments accounts called 401(k). The investor in their dataset could be considered unskilled as saving investments are not generally among the preferences of active investors or day-traders, who represent the more skilled financial agents, since savings investments are characterised

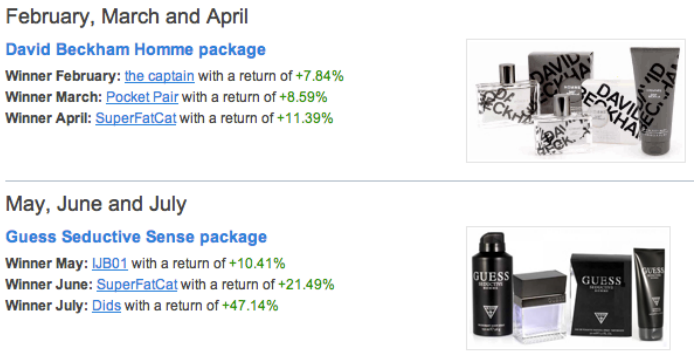

Fig. 4.6 Examples of prizes for the Virtual Trader online game simulation. The highest ranked player for each month wins a package of perfumes and other man grooming products.

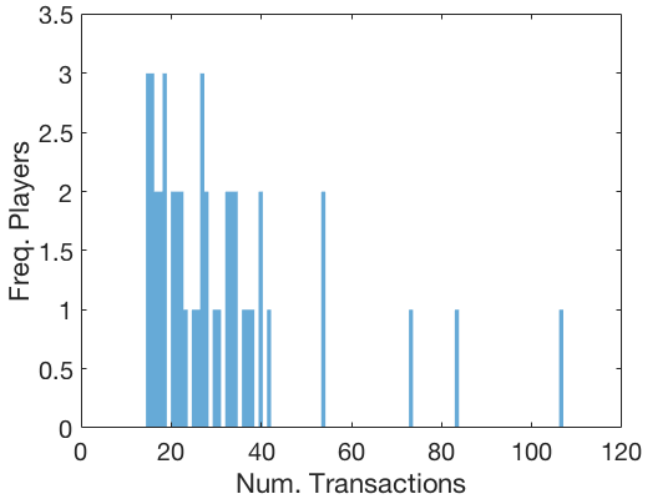

Fig. 4.7 Histogram of the transactions made by the players in the dataset. Transactions $total = 1420$, $maximum = 106$, $minimum = 16$, $mean = 30.87$, $median = 27$.

by low risk and low potential returns. The results in Choi et al. [2009] indicated that those investors who experienced a positive return from their saving account, increased their commitment to that particular savings plan the following year. Conversely, those investors whose saving accounts investments returned discouraging results, reduced their pledge to the 401(k) plan. In their paper, Choi et al. describe this pattern of behaviour as "naive" reinforcement learning. The term "naive", in Choi et al. [2009], is used in the sense that decision-making is based only on previous personal experience and not on other information which could lead to potentially higher future profit. By dissecting their observation, it is possible to identify two questions: one concerned with understanding whether unskilled investors learn by reinforcement, the other focused on investigating whether their learning is indeed naive and short-sighted.

4.2.2 Risk based stock classification

The indications lead to some interesting questions regarding the behaviour of unskilled investors. Firstly, it would be interesting to know whether they group the available stocks into classes, for example according to a shared feature. The first hypothesis is, therefore, focused on testing the assumption that the players of this trading simulation game perceive the tradable stocks as grouped into discrete classes of risk. In order to clarify whether this is indeed the case, we develop part of the modelling to test this assumption. To do so, a classification of stocks based on risk is developed and compared to 500 randomly generated classifications. The classification of stocks into discrete categories developed in this work is

based on a measure of the risk of an asset in comparison to the market, deriving from the Capital Asset Pricing Model (CAPM, Sharpe [1964]) and called the "beta coefficient". This measure provides an indication of the risk of an asset, in comparison with the market and is widely used in financial modelling (Beninga [2000]; Black et al. [1972]; Merton [1973]). By evaluating the volatility (standard deviation) of the returns for an asset and the relative benchmark, this measure expresses how sensitive the returns of the asset are in comparison to the returns of the benchmark. The beta coefficient, also referred to as "financial elasticity" of a security, is calculated using historical price data and represents how much the security price movements are correlated with the benchmark, which is usually defined as the market index[4] underlying the security considered. A security is said to have lower volatility than the market if the price movements are less pronounced than the market and viceversa, when a stock presents higher volatility than the market, it presents more marked price fluctuation in relation to the market.

The assumption that players perceive the available trading options as classified according to their risk is tested by comparing the best fitting model which uses the risk-based classification, against the best fitting model which uses one of the 500 random classifications. This comparisons are based on the Akaike Information Criterion measure of the models goodness-of-fit to the data. This set of comparisons is treated as a series of binomial outcomes and the Clopper-Pearson binomial confidence interval methodology (Clopper and Pearson [1934]) is used to evaluate the probability that the risk-based categorisation is better than the 500 randomly generated ones at describing the players. The first hypothesis will test the risk-classification assumption within a RL framework.

4.2.3 Reinforcement learning as a descriptive model

Since the test of the first hypothesis requires RL models to be fitted to the subjects, it is important to test whether the learning and decision-making behaviour players exhibit in this dataset does indeed follow a reinforcement learning pattern. Therefore, the second hypothesis is focused on testing whether the behavioural data collected for the players of the Virtual Trader online game can be described with a reinforcement learning model. In a series of economics experiments in Erev and Roth [1998], RL is a found to be a good predictor of the evolution of play in a series of economics experiments and games. In case this hypothesis is confirmed, this work would provide new evidence that RL accounts for unskilled investors'

[4] A stock market index is a mathematical construct which represents a portion of the stock market. Its value is often calculated as the weighted average of the stocks included. It is used as a representation of the market, against which to asses the return of other investments or single stocks. FTSE100 is an example of British stock index, Standards & Poor 500 is an American stock market index. They are both based on the market capitalisation of, respectively, the 100 and 500 largest companies in the relative stock market.

behaviour. Formally, the second hypothesis is that the Virtual Trader investment game players behave following a reinforcement learning pattern. To test whether this is the case, a simple average-tracking RL model will be fitted to the individual players decision data, which consist of the players choices made in the game and their corresponding outcomes. This model represents a myopic strategy, where a player is concerned with maximising the immediate rewards, but does not care about the potential long-term rewards. The fitting procedure follows the maximum likelihood estimate (MLE) methodology presented previously in chapter 4, section 4.1.14, eq. 4.12. The Bayes' rule used in this methodology is replicated here:

$$P(\theta_M|D,M) \propto P(D|M,\theta_M) \cdot P(\theta_M|M) \tag{4.21}$$

where D is the data, M is the model, and θ_M is the parameter set associated with model M. The best fitting models are identified by evaluating different parameter sets, with the aim of maximising the likelihood of the data being generated by the model tested, represented by the term $P(D|M,\theta_M)$. This quantity, multiplied by the prior, indicated by the term $P(\theta_M|M)$, is proportional to the posterior likelihood, represented by the left-hand side of the proportionality, $P(\theta_M|D,M)$. Therefore, if treating the prior as constant (since there is no a-priori knowledge of which parameter set is the most representative), maximising the likelihood achieves the same objective as maximising the posterior likelihood. This second hypothesis will be tested by comparing the RL models against a baseline random model which assumes equal probability of picking any available action. The parameter set associated with a model is denoted with θ_M, the maximum likelihood estimate is the parameter set which maximises the goodness-of-fit of a model to the data and is denoted with $\hat{\theta}_M$, therefore the MLE values for these two models are compared by means of a Likelihood Ratio Test (LRT; Huelsenbeck and Crandall [1997]; Wilks [1938]). The LRT is a statistical test which uses the ratio of the likelihood scores of two nested models, defined as the null model and the alternative model. The likelihood ratio quantity represents the amount of support in favour of the more complex model (Edwards [1972]). The LRT test takes into account the difference in the numbers of parameters, using it as a penalising factor for the more complex model, when estimating the evidence in its favour. Formally, the likelihood ratio test is defined as:

$$LR = 2 \cdot (\mathfrak{L}(\hat{\theta}_{alt}) - \mathfrak{L}(\hat{\theta}_{null})) \tag{4.22}$$

where $\hat{\theta} = MLE$ is the maximum likelihood estimate, *alt* is the unrestricted model, *null* is the restricted model, and $\mathfrak{L}$ is the log-likelihood of these values. This test statistic is χ^2

distributed with degrees of freedom depending on the difference of parameters between the null and the alternative model. The outcome of the LRT is evaluated against the χ^2 distribution with the number of degrees of freedom equal to the difference in parameter between the two models, and at the 5% significance level. If its value exceeds the critical value then the null model is rejected in favour of the unrestricted model.

If a subject best fitting RL model does not fit the data significantly better than a random model, then, for that subject, it is assumed that the strategy adopted in the game is not based on reinforcement learning. Once the most representative model is found, for each subject in the dataset, a χ^2 test is performed on the frequencies of the two models, to provide an answer to the hypothesis. If the number of players best fitted by a RL model is significantly higher than the number of players whose RL models failed to fit significantly better than random, then the null hypothesis is rejected. Conversely, if not enough players are described satisfactorily by a RL model, as opposed to a random model, the alternative hypothesis is rejected. The third hypothesis further investigates the players, to understand whether a more complex strategy is more descriptive of their behaviour, when compared to the random or myopic strategy.

4.2.4 Naive behaviour as short-sighted learning

Literature provides indications that individuals tend to follow personal experience, instead of other more sophisticated methods, when making investment decisions. For example, the analysis of saving accounts' data, in Choi et al. [2009], provided evidence that unskilled investors make decisions in a way which appears to overweight personal experiences, therefore referred to as naive reinforcement learning. Similar conclusions are found in Huang [2012]: investors are more likely to trade in those industries in which they had positive experiences (i.e. trades resulting in gains). Moreover, higher sophistication levels or longer time-ranges reduce this effect. Therefore, the third hypothesis is to test whether unskilled investors behave in a naive way, by short-sightedly aiming to maximise their immediate rewards. The method to test this hypothesis involves fitting RL models which represent different strategies. Following the same structure as in the second hypothesis test, the Virtual Trader players choice data will be fit with a more complex RL model which characterises far-sighted behaviour. The average-tracking RL model adopted previously in this work, outlined in section 4.1.9, captures myopic behaviour because its update rule is only focused on current rewards, hence the name immediate-rewards RL.

Far-sighted strategies instead, can be captured by temporal difference learning (TD-learning), which includes a future rewards term in the learning rule. The model supposed to capture long-term strategies is Q-learning, first introduced in chapter 2, section 2.25 and

adopted in the modelling performed in the previous chapter, in section 4.1.9. Q-learning is a widely adopted implementation of TD-learning, which a large body of evidence describes as a computational account of the dopaminergic activity in areas of the brain associated with decision-making and learning (Cohen et al. [2007]; Joel et al. [2002]; Lohrenz et al. [2007]; O'Doherty et al. [2003]). More details about the learning rules used by these models are provided later in this chapter, in section 4.2.8. To test this hypothesis the myopic RL model and the far-sighted RL model are fitted to the subjects' data, and a likelihood ratio test is carried out on the corresponding MLE values. The most representative models is evaluated by means of the Likelihood Ratio test, as the immediate-reward RL model is nested in Q-learning. Finally, two χ^2 tests will be carried out on the frequencies of each model within the players dataset, to provide an answer to the hypothesis. The number of players well fitted by the far-sighted model will be compared first to the myopic RL model and then to the random model. The first comparison will provide an indication of how many subjects adopted a far-sighted strategy in the game. The second comparison will indicate whether the players are actually better represented by a far-sighted strategy. If the number of players better fitted by the model representing a far-sighted strategy is higher than the alternative simpler models in both tests, then the hypothesis that unskilled investors behave in a naive reinforcement learning way is rejected. Alternatively, if the number of players fitted by a myopic strategy model is higher, then it can be concluded that unskilled investors are indeed naive. Finally, if the tests indicate that the players are not well fitted by either of these two strategies, then no indication can be offered on whether players follow a reinforcement learning process, being it naive or far-sighted. The details of the models used to test these hypotheses are presented in the following section.

Models details

Similarly to the binary decision task models developed in the previous chapter, this task is modelled using the reinforcement learning framework; this section describes the components adopted to model the scenario and the rationale behind them. The game analysed consists of repeated interactions with a financial stock market simulation, which follows the Financial Times Stock Exchange 100 Index price movements. A player corresponds to a reinforcement learning agent while the trading simulation game represents the observable portion of the environment.

4.2.5 State-space

The investigation of state-space configurations performed in the previous chapter provided some indication about the type of state-space which best describes the mental accounting operated by the subjects in the binary-choice task analysed. The environment for the Virtual Trader game is modelled as a binary state-space because this proved to be the scenario which best represented subjects in the investigation carried out for the previous task, both in absolute terms and when statistically compared to the other proposed state-space cases. Moreover, by using only one state-space modelling, this investigation can focus on other research questions. The state-space consists of two states: $S \in \{gain, loss\}$, shown in Fig.4.10. The players' state of the world is determined by their balance, which is the sum of the outcomes from the transactions made up to that point:

$$B_t = \sum_{i=0}^{t-1} o_i \tag{4.23}$$

with o_i being the i-th transaction outcome and B_t being the balance at the time t. The state of a player at time t is defined by the balance sign:

$$state_t = \begin{cases} gain & \text{if } B_t \geq 0 \\ loss & \text{otherwise} \end{cases} \tag{4.24}$$

A player's current balance is estimated as the outcome of the sell transactions. Purchase transactions only offer information on the amount spent but this is not enough to tell how good the transaction is. Therefore, the current state depends only on the sales outcomes, a player's cash flow; while the holdings, which are the shares owned by a player at any time, are not accounted for.

4.2.6 Reward signal and transformation

The decision to use the players cash but not their holdings has also implications in the definition of reward. The information accessible by the players needs to be considered when defining what is considered to be the reward signal. In Virtual Trader, players are presented with the outcomes of their trades in the game graphical interface. Therefore, a reasonable way to adapt this information to the RL modelling is to consider the outcome of each transaction as the reward signal for the RL agent. The assumption that the reward signal is based on the cash flow arising from the sales is also linked to a well known phenomenon known in literature as the "disposition effect" (Barber and Odean [2013]; Odean [1998]). Investors tend to sell securities which appreciated since the time they were bought, while holding positions

which lost value since purchase time. The disposition effect is more evident for individual investors, but is present also in bigger investing firms such as corporations or mutual funds (Barber et al. [2007]; Brown et al. [2006]; Frazzini [2006]; Grinblatt and Keloharju [2001]; Heath et al. [1999]; Odean [1998]; Shapira and Venezia [2001]). Therefore, the "holdings" component of a player's wealth, in the Virtual Trader simulation is not considered as part of a reward signal, while the sell transaction outcomes are considered, since these are likely to represent a player's will and because it realises a measurable gain (or loss). Furthermore, the impracticality of estimating the exact value of a player's portfolio holdings at each transaction time is another reason why only the cash component is considered. The scope of the model is reduced to what can be considered a "Sell-model".

The reward is defined on the sell transactions, with the purchasing transactions used to keep track of the player's portfolio and to estimate the price spread at the time of sell. The reward r at time-step $t+1$ is defined as:

$$r_{t+1} = v_{t+1}\left(p_{t+1} - \frac{1}{t}\sum_{i=1}^{t} v_i p_i\right) \tag{4.25}$$

where v_i is the volume and p_i the price of the stock sold at the i-th time step. The difference in the brackets represents the price spread, where p_{t+1} is the current sell price, while the summation term represents the weighted average price of previous purchases, with the prices weighted on the volumes of previous transactions. The result of this calculation r_{t+1} is the raw reward, which is then transformed via the hyperbolic tangent.

Three reward transformation functions have been proposed and tested in the previous work chapter; in this chapter instead, the modelling focuses only on the hyperbolic tangent. For the binary task examined, this function has been shown to be the least representative reward transformation in the development of descriptive models of decision-making. Nonetheless, the hyperbolic tangent presents properties which make it a good choice for the current modelling attempt; its sigmoidal nature captures decreasing utility in both gains and losses domains, and its saturating property restricts the magnitude of the rewards in the range $[-1,+1]$. This last property is necessary to analyse this dataset, because the other non-saturating functions, previously tested, lead to numerical instabilities, due to the extreme values in the range of original rewards. The range of the original rewards is $[-5695, 79264]$, with mean = 477 and median = 93. The original and transformed reward distributions are shown in Fig. 4.8.

Several attempts to fit the models using both the identity function and prospect theory's subjective value function produced numerical overflows, resulting in computational insta-

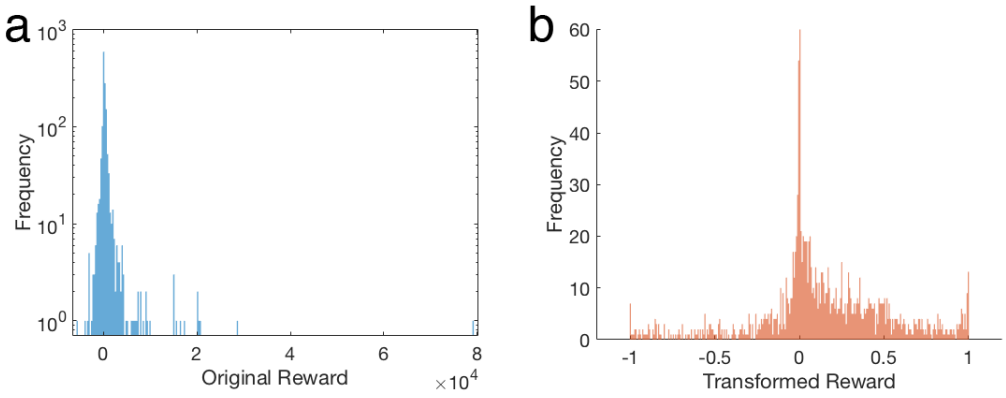

Fig. 4.8 The distribution of rewards before and after the hyperbolic tangent transformation. Panel a, the original range on the x-axis $[-5695, 79264]$ with frequencies on the y-axis and in log scale. Panel b, the new range on the x-axis, reduced to $[-1, +1]$, y-axis showing the frequencies in linear scale. As a result, rewards $r > 5,000$ are collapsed to 1. Similarly when $r < -5000$ the values are collapsed to 1.

bilities, which broke the model fitting routines. The hyperbolic tangent adopted is defined as:

$$tanh(r) = \frac{1 - e^{-r\omega}}{1 + e^{-r\omega}} \tag{4.26}$$

where r is the reward signal, e is the exponential function and $\omega = 1/500$ is the parameter which regulates the slope of this sigmoid function. The resulting slope allows to maintain most of the variability of the rewards, only flattening the extreme values which caused the numerical issues. This choice resulted in only 12 out of 1420 (<0.008%) rewards being collapsed on a single value, rewards above GBP 5,000 being flattened to 1: $r > 5,000 \rightarrow r = 1$.

4.2.7 Stocks and Actions

Players are endowed with a virtual lump sum of GBP 100,000 at the moment of subscription and are allowed to trade (buy or sell) stocks over time, with the objective of increasing the value of their portfolio.

The payoff distribution behind each stock is unknown and non-stationary, for example a security which appreciated at a certain time, providing a positive return to a player, might plummet in value later on. As this is not a controlled experiment, the number of available actions depends solely on the particular stock price movements and there is no a-priori good or bad choice, unlike in the binary choice task examined in the previous chapter.

The FTSE100 is a stock market index which includes the 100 companies with the highest capitalisation listed on the London Stock Exchange (LSE). At the time of data collection

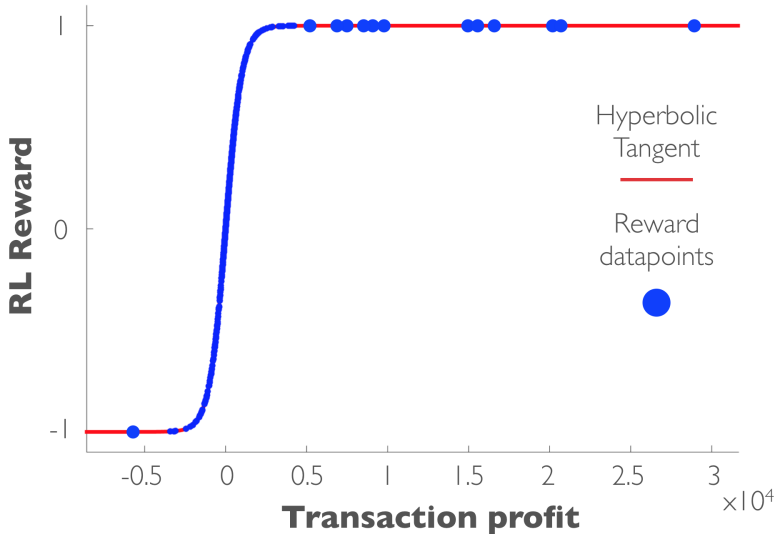

Fig. 4.9 The hyperbolic tangent function used to reduce the range of rewards from $[-5695, 79264]$ to $[-1, +1]$. The slope of the function has been chosen so that as much possible reward variability is still portrayed in the transformed values. Rewards $r > 5,000$ are collapsed to 1. Similarly when $r < -5000$ the value is collapsed to 1, which happened for only one transaction in the dataset.

the FTSE100 included "Royal Dutch Shell" stocks which are composed of class A and B shares therefore listing 101 available options. Moreover, because the index is evaluated every quarter of a year it could happen that new companies are added while others are excluded. During the time range considered for this study, these changes led to a total of 107 companies available to be traded in the game. If every tradable stock is considered as a potential action, and the players are supposed to be in either of the two states previously defined, the total amount of available actions would be $107 \times 2 = 214$. Considering such a high number of actions over a relatively short time-frame, during which players interacted on average 30 times, would represent a learning process potentially too hard to model. Therefore, a dimensionality reduction on the action space is required to make this model numerically treatable. The reduction of the action-space dimensionality offers a chance to test the first hypothesis, that players group the available stocks according to their risk.

The categorisation adopted to test this hypothesis and to tackle the action-space dimensionality reduction is defined on a measure of the stocks' risk, called financial elasticity and representing the risk of a security in comparison to the market. The Capital Asset Pricing Model (CAPM, Sharpe [1964]) is a security pricing model developed independently, by several economists including Merton Miller, Jan Markowitz and William F. Sharpe who jointly received a Nobel Prize in Economics for their work. The CAPM model is particularly suited for the purpose of assessing the risk of a security because it features a quantity called

beta coefficient, denoted with β_F[5]. This quantity measures the elasticity of a security in comparison to a benchmark index. This method of assessing an asset's risk is used in financial modelling to evaluate the risk of the components in a portfolio (Beninga [2000]; Black et al. [1972]; Merton [1973]). The beta coefficient risk measure β_F is defined as:

$$\beta_F = \frac{Cov(r_a, r_b)}{Var(r_b)} \tag{4.27}$$

where r_a are the returns of a security a, r_b are the returns of the benchmark index b, $Cov(r_a, r_b)$ is the covariance of these two quantities and $Var(r_b)$ is the variance of the benchmark index returns. To better understand how this measure represents the volatility of stocks in comparison to the market, it is useful to provide some examples. Gas and power related stocks are relatively safe investments as it is rare for hugely disruptive events to happen that can affect the price trend of these utilities stocks, hence such stocks present a low β_F. On the other hand, high-tech stocks are highly susceptible of events, such as innovative start-ups or new technologies being introduced, which would destabilise these markets; these stocks are characterised by a high β_F. A beta coefficient $\beta_F = 1$ corresponds to perfect synchrony in price fluctuations. When $\beta_F \in (0,1)$, the asset manifests lower volatility (or low correlation) of the price movements compared to the benchmark. If the value of $\beta_F > 1$ the security is characterised by higher volatility than the market. Applying this to the examples provided, high-tech stocks with a $\beta_F > 1$ would outperform the benchmark index during an upward trending market. This high value also means that when the market is going down the high-tech security would depreciate at a higher rate than the market. Hence, higher values of β_F coefficient indicate higher risk in the security. The beta coefficient is a measure of intrinsic, or systematic, risk which can also be estimated by regression:

$$r_a \approx \alpha_F + \beta_F r_b \tag{4.28}$$

where r_a and r_b are the returns of an asset and the benchmark, and α_F represents the active return (which is what skilled investors often attempt to optimise in their trades). The beta coefficient of each stock is calculated and used, in this thesis, to classify the stocks into three risk categories. This procedure reduces the action-space from 107 to 3 (classification described in 4.3, globally reducing the total of 214 actions available in 2 states, to 6 actions in 2 states). The estimation of the stocks' beta coefficient is achieved by gathering daily returns from the 1st June 2013 to the 31st of May 2014. This choice of time-range and price interval guarantees an extension covering the months before the players' interactions with the game,

[5]the original beta coefficient is denoted with β but to avoid confusion with other parameters previously introduced in this thesis, it will be referred to as β_F

also providing some overlap with the game time-range, and a daily granularity. The stocks are then ranked and subdivided in three categories: low- and mid-risk featuring 36 stocks and high-risk class which contains 35 stocks. The state-space is therefore composed of two states and three actions, with the relative Markov decision process (MDP) shown in Fig 4.10.

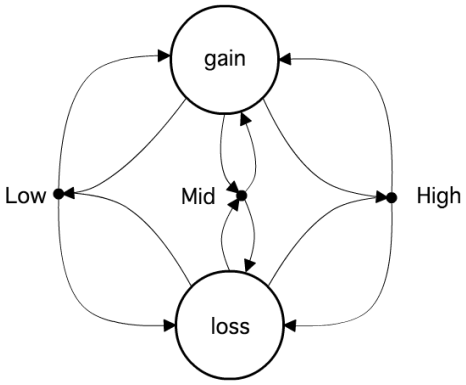

Fig. 4.10 The transition graph describing the Markov decision process used to model the structure of the Virtual Trader online financial trading simulation game. The actions the agent can take are represented by labelled black solid circles (Low, Mid and High), while the states are shown as labelled empty circles (gain and loss). The modelling is based on two states: loss and profit, which are defined on the sum of previous transaction outcomes. The modelling assumes three actions which correspond to the three discrete degrees of risk. The stocks are classified based on their financial elasticity, beta coefficient β_F, a measure of risk of a security in comparison to the market.

It is important to note that different numbers of classes have been tested in the categorisation of stocks, specifically 2 and 4 classes of risk; these produced results consistent with the 3 class categorisation and will therefore not be discussed further.

4.2.8 Learning models and policy

This dataset of choices and outcomes collected for this study has never been analysed before; no previous descriptive modelling attempt is known for such data. The purpose of the analysis carried out in this work is to understand whether the Virtual Trader players operated in a reinforcement learning manner and whether they have done so in a myopic way. To test the hypotheses of this chapter, the players are described as reinforcement learning agents who operate decisions following a probabilistic strategy. One way to assess if this is indeed the case is to compare the goodness-of-fit for the models proposed against a baseline. Since these choices have never been modelled before, a random model is used as baseline comparison. The attempt at modelling myopic behaviour is done by adopting the average-tracking (immediate-reward) reinforcement learning model. The comparison between such

Table 4.3 List of stocks classified according to their risk (financial elasticity, beta coefficient β_F)

Low-risk	Mid-risk	High-risk
Kazakhmys Plc	Johnson Matthey	Lloyds Banking Group Plc
Serco Group	Croda International	Kingfisher
Centrica Plc	Intu Properties	Aviva Plc
SSE	Carnival Plc	Hargreaves Lansdown Plc
G4S Plc	AstraZeneca Plc	TUI Travel Plc
Imperial Tobacco	BAE Systems	Burberry Group
Admiral Group Plc	Friends Life Group	ITV Plc
RSA Insurance Group Plc	Aggreko	Melrose Industries
British Sky Broadcasting Group	Vodafone Group Plc	Barclays
United Utilities Group	Tullow Oil Plc	Wolseley
Pearson	Rexam Plc	SABMiller Plc
Randgold Resources Ltd	Compass Group Plc	WPP Plc
Severn Trent	British American Tobacco Plc	Travis Perkins Plc
Morrison Wm Supermarkets	Rolls-Royce Holding	Standard Chartered
Sainsbury (J.)	Unilever	Antofagasta
Capita	Sports Direct International Plc	BHP Billiton Plc
Next	Smiths Group Plc	Schroders Plc
Reed Elsevier Plc	Marks&Spencer Group	Standard Life Plc
Royal Dutch Shell-B Shs	Land Securities Group	Glencore
Royal Dutch Shell-A Shs	Diageo Plc	Rio Tinto
Intertek Group	Experian	International Consolidated Air
National Grid Plc	BG Group Plc	Old Mutual
Sage Group Plc	IMI	Prudential
AMEC Plc	Hammerson	ARM Holdings
BP	Fresnillo Plc	GKN
Tesco	Shire Plc	easyJet Plc
Reckitt Benckiser Group	British Land	Royal Bank of Scotland Group
Smith & Nephew	William Hill	Persimmon
Bunzl	Weir Group	Vedanta Resources Plc
Tate & Lyle	HSBC Holdings Plc	CRH
Associated British Foods Plc	Wood Group (John)	Mondi Plc
GlaxoSmithKline	InterContinental Hotels Group	Anglo American Plc
Meggitt Plc	London Stock Exchange Group	Polymetal International
Babcock Intl Group	Legal & General	Evraz
Coca-Cola HBC	Petrofac Ltd	Aberdeen Asset Management Plc
BT Group Plc	Whitbread 'A'	

RL model and the random baseline will provide an answer to the second hypothesis, that Virtual Trader players behave following a reinforcement learning pattern.

From a probabilistic point of view, a random model is described with an agent whose action-selection policy selects, at any time-step, one of the available actions with equal probability. Formally, the baseline random model can be estimated as:

$$P(D|M_{Rand}) \;=\; MLE_{Rand} = \sum_{t}^{N} \log P(c_t) = N \cdot \log\left(\frac{1}{3}\right) \tag{4.29}$$

where D is the data, M_{Rand} is the random model, log is the natural logarithm function (with base e), N is the number of transactions a player executed, and $\frac{1}{3}$ is the probability of selecting an action in the scenario encompassing three available risk-classified actions.

The immediate-rewards RL model to be tested against this baseline uses the average-tracking rule described in chapter 4, section 4.1.9 and formalised in eq. 4.10, reproduced here for convenience:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} - Q(s_t, a_t)] \tag{4.30}$$

where s_t and a_t are the state and action at time-step t, $Q(s_t, a_t)$ is the Q-value associated with this state-action pair, r_{t+1} is the reward obtained from the interaction happened at time-step $t+1$ and α is a step-size parameter, known as learning rate, which indicates the amount of new information used to update the agent's belief about the Q-value of the current state-action pair. This rule is based on a bootstrapping method, because it builds estimates (left-hand side of the assignment) based on previous estimates (right-hand side of the assignment), eventually converging to the actual value when more and more experience is accumulated (formally denoted with $Q*$). This learning rule modifies the Q-values estimates by considering a portion (scaling by α) of the numeric signal from the immediate reward (r_{t+1}), and how much this value differs from the previous estimate. No information about potential future rewards is taken into account in this model, which explains why it is called "immediate-rewards".

The third hypothesis tests whether the Virtual Trader players who behaved in a reinforcement learning way did so by operating unsophisticatedly, as suggested by Choi et al. [2009] and Huang [2012]. In these work, unskilled investors are found to rely too much on their previous personal experience, attributing less importance to the long term implication of their actions. From a descriptive point of view, the immediate-rewards learning model previously introduced represents an attempt to capture the players who acted short-sightedly. A good candidate model for capturing far-sighted behaviour, representing non-naivety, is temporal difference learning (TD-learning). This learning model incorporates a future reward term in the update rule and provides a free-parameter to tune the degree to which an agent is concerned with future rewards. Q-learning, described in 4, section 4.1.9, formalised in eq. 4.1.9, is an off-policy implementation of TD-learning:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_{a} Q(s_{t+1}, a) - Q(s_t, a_t)]$$

where the term $\max_a Q(s_{t+1}, a)$ represents the future rewards obtainable following a greedy policy (i.e. selecting the actions believed to yield the best outcome in future interactions), and with γ being the discount factor parameter which adjusts how much this information influences the current update.

Q-learning update rule is also based on bootstrapping, with the future reward term being an estimate of the actual return (cumulative rewards) obtainable from the next time-step. The addition of a term representing future rewards is important because it allows this model to capture more complex behaviour, including strategies aimed at learning the behaviour leading to the best return in the long-term. This type of RL has been considered a computational account of dopaminergic neurons activations in the striatum and orbito-frontal cortex (Cohen et al. [2007]; Joel et al. [2002]; Lohrenz et al. [2007]; O'Doherty et al. [2003]). Q-learning is an off-policy implementation of TD-learning, which has been previously used to model instrumental conditioning (Dayan et al. [2006]), to generate individual behaviour in the development of an artificial stock market (Rutkauskas and Ramanauskas [2009]), or to optimise trade execution (Nevmyvaka et al. [2006]).

These two RL models, previously adopted for the analysis of the behavioural data in the experimental scenario studied in chapter 4, will also be adopted in this quasi-field study. In the previous work these were both fitted with the intent of finding the most descriptive computational account of subjects behaviour, in order to further study the characteristics of the captured strategies. In the present work they are tested against each other and against a baseline random model, to find if they can provide a good descriptive account of behaviour and, in such case, which one is the most descriptive. The comparison of the RL models against the random baseline will help clarify if Virtual Trader players learn by reinforcement with their interactions; the head-to-head comparison between the myopic and the far-sighted RL models will help clarify whether the unskilled investors in the Virtual Trader dataset behave naively, as suggested by literature.

Action-selection: Soft-Max

In a similar way to the previous part of the work, the hypotheses testing methodology is based on comparing measures of descriptiveness of the models tested. Maximum likelihood estimates are the starting point for such methodology. These are measures of the quality of a model in describing a dataset, based on a probabilistic approach. In order to estimate MLE scores it is necessary to adopt a probabilistic model of action-selection, which quantifies the probability of each choice made by players. This is conveniently achieved by using Soft-Max action-selection policy, previously described in chapter 2, section 2.2.1, eq. 2.24 and applied in the previous work chapter 4, section 4.1.11, eq. 4.11. This rule transforms an agent's

beliefs about the Q-values (quality of state-action pairs) into probability values. This process allows the estimation of a model likelihood, the probability that a model with a specific set of parameters generates a specific set of choices. The Soft-Max rule adapted to the current analysis is:

$$P(high) = \frac{e^{Q(s,high)\cdot\beta}}{\sum_{a\in A(s)} e^{Q(s,a)\cdot\beta}} \tag{4.31}$$

where $P(high)$ is the probability of selecting the action with the highest risk, e is the exponential function, $Q(s,high)$ is the value of taking action *high* when the agent is in state s, $A(s)$ is the set of actions available when the agent is in state s, and β is the inverse temperature free-parameter which regulates the greediness of the strategy.

4.2.9 Model Comparison

Equally complex models are compared to test the first hypothesis; these differ in that they are fitted considering the action-space built on different arrangements of stocks in discrete categories. The objective of this first inquiry is to find out if the risk-based classification of stocks proposed in this work is a good account of the players' risk perception. To accomplish this task, the model fitted using the risk-based classification is tested against models trained using the 500 randomly generated classifications. The Akaike Information Criterion (AIC; Akaike [1974]) is used to evaluate the goodness-of-fit for the models tested on either the risk-based classification or on the "scrambled" stocks (i.e. the randomly-classified arrangements). The Bayesian Information Criterion (BIC; Schwarz [1978]) has been tested as well; the AIC produced consistent results with the BIC, therefore only the AIC based results will be presented. 500 comparisons are performed for each of the 46 players. For each player, the AIC score of the risk-based model is compared to the AIC scores of the models fitted using the 500 scrambled stocks. The comparison is carried out between the best fitting models for each stock-arrangement; this means that the parameters sets representing the maximum likelihood estimate for one model could differ from the ones of another model. This comparison was chosen to be executed in this way to avoid biasing the scrambled arrangements with parameters sets which do not necessarily represent the most descriptive for these random classification. The result of these 500 comparison generate a binary vector, which can be seen as a series of trials, with each element indicating whether the risk-based model fits the data better than the i-th randomly-generated model:

$$v_i = \begin{cases} 1 & \text{if } AIC_{risk} < AIC_i \\ 0 & \text{otherwise} \end{cases} \tag{4.32}$$

where v_i is the outcome of the i-th comparison, with $i = \{1, \cdots, N\}$ and $N = 500$ the number of randomly generated stock arrangements; AIC_{risk} is the AIC score associated with the best model which uses the risk-based classification and AIC_i is the AIC score for the i-th randomised classification. Given the binary vector resulting from the AIC comparisons between risk-based classification and scrambled versions.

$$\vec{v}_p = [v_1, v_2, \cdots, v_i, v_N]' \tag{4.33}$$

where $\vec{v}_p$ is the vector of comparisons for player p, v_i is the i-th comparison result, which can be 1 in case the risk-classified model fitting performed better than then scrambled one and 0 in the opposite case, as described in eq. 4.32.

The Clopper-Pearson binomial confidence interval (Clopper and Pearson [1934]) is a method for calculating binomial confidence intervals; it is referred to as an "exact" method since it is not based on an approximation of the binomial distribution, unlike for example, the normal approximation method (Clopper and Pearson [1934]; Neyman [1935]). The Clopper-Pearson interval method is adopted in this work to estimate the probability, and the relative confidence intervals, that the proposed stock classification is better than the 500 random classifications. To test the first hypothesis, the resulting confidence intervals will be evaluated against the 50% chance threshold; players whose confidence interval lies fully above this threshold are likely to perceive the tradable stocks as belonging to discrete classes defined on the stocks risk, estimated with a measure of their volatility in comparison to the market.

In order to test the second hypothesis, that players behave in a reinforcement learning way, an immediate-reward RL model will be compared to a baseline random model, for each player in the dataset. In order to test the third hypothesis, that unskilled investors behave in a naive way, a temporal-difference RL model will also be fitted to the data, so that a comparison between naive strategy and far-sighted strategy can be performed. As described previously, average-tracking RL will be used as the computational counterpart of naive behaviour, while Q-learning will be adopted to represent the far-sighted strategy.

The best parameter set for the models is identified by means of a bounded gradient descent searching algorithm. Gradient descent is an optimisation algorithm which aims at finding the minimum of a function. For this work, the function to be minimised is the model fitting routine, which corresponds to the iterative estimation of Q-values and the subsequent

estimation of the probability of the actions taken by the players. The maximum likelihood estimate represents the output of this function and the objective of the minimisation. The gradient descent algorithm implemented to fit the models to the data is performed on a bounded space defined for the three free-parameters of the models and with 27 combinations of initial guess-points. The guess-points are the starting values each free parameter takes at the beginning of the gradient descent search routine. It is extremely important to distribute these initial points across the search space, because this reduces the chance of the gradient descent routine getting stuck in local optima.

The boundaries for the parameters in the gradient descent search procedure are $\alpha \in (0.0001, 2)$, $\beta \in (0, 50)$, $\gamma \in (0, 0.9999)$ (for the myopic model $\gamma = 0$).

The guess points are combinations of $< \alpha, \beta, \gamma >$ from the values in the following sets: $\alpha \in \{0, 0.5, 2\}$, $\beta \in \{0, 25, 50\}$, $\gamma \in \{0, 0.5, 0.9999\}$.

The immediate-reward RL model features only the first two free-parameters, α and β which represent the speed of learning and the propensity to select the best action (i.e. greediness). The MLE for these models will be compared to the MLE of the random model which can be calculated analytically as shown in eq. 4.29. The evaluation of random model MLE is similar to the one established in the previous chapter (eq: 4.15), with the difference that, in the controlled experiment scenario, the number of trials (i.e. subject's decisions) was fixed by the experimental design, while in this work it is derived by the quantity of transactions made by each subject. This leads to different values of MLE_{RND} for each player, depending on the number of interactions that particular player has produced.

The immediate-rewards RL model includes two parameters, while the random model has no parameters; since the random model can be considered nested, in the case when the free-parameters $\alpha = 0$ and $\beta = 0$, this comparison can be carried out with the Likelihood Ratio Test. The LRT comparison, similarly to the Akaike Information Criterion adopted in the previous work, penalises the more complex model. The probability distribution of this test statistic is approximated by a χ^2 distribution with the degrees of freedom ($d.o.f.$) equal to the difference between the number of parameters of the models tested (Huelsenbeck and Crandall [1997]; Wilks [1938]). The LRT is convenient because it also provides a straightforward way to estimate a p-value for the statistical significance of the comparison.

The third hypothesis will be tested in the same way by comparing a model representing long-term strategies, which features three parameters, including a long-term discounting factor, against the naive immediate-reward RL model. For this comparison the $d.o.f. = 1$, while in the comparison between immediate-reward RL and random model the $d.o.f. = 2$.

4.2.10 Models summary

Table 4.4 presents a summary of the three models which will be fitted to the Virtual Trader players dataset. The baseline is represented by the random model, which trivially considers the available actions equiprobable, not making any assumption about the subjects decision-making or learning processes and therefore not featuring any parameter. The random model does not represent any particular strategy but provides a starting point against which to compare the descriptive achievements of the other proposed RL models.

The second model is the immediate-reward reinforcement learning, also known as average-tracking, representing the naive strategy. This model features a simple learning rule which focuses only on the maximisation of the reward available at the current time-step. This model represents a short-sighted reinforcement learning model, which does not take into account future rewards, therefore representing the myopic, naive strategy of an agent focused on short-term goals. This simple RL model includes a step-size parameter representing the speed of learning, denoted with α, which scales the new evidence obtained when applying the learning rule.

The last model is Q-learning, an off-policy implementation of temporal-difference learning. This model learning rule incorporates a future reward term in the prediction error, which allows to capture far-sighted strategies. The information about future attainable rewards from the current time-step is scaled by a parameter called discount factor and denoted with γ, which is not included in the myopic RL model. Its value indicates the degree to which potential future rewards are considered when updating the belief about the current state-actions pairs. With values closer to 1 indicating a more far-sighted time horizon and values closer to 0 being indicating a more myopic approach. This type of model has the power to capture more complex and far-sighted strategies, as opposed to its nested myopic version.

Both the proposed RL models operate with Soft-Max as the probabilistic action-selection policy, described previously in this chapter, in section 4.2.8, eq. 4.31. This stochastic rule translates the value of state-action pairs into probabilities and involves a free-parameter called inverse temperature, denoted with β and representing the greediness of an agent when deciding which action to take; it tunes the probability of picking actions according to their previously computed values. The extreme values of this parameter result in either completely random behaviour (i.e. $\beta = 0$), or greedy behaviour (i.e. $\beta \to \infty$), selecting only the action believed to be the best and never exploring.

Each model is nested within the more complex one: the myopic RL is equivalent to Q-learning when one of the parameters, the discount factor γ is set to 0; similarly if both the parameters of the myopic RL, α and β, are set to 0 it reverts to a random model, as no learning can happen and the action-selection strategy is to pick an action randomly.

Table 4.4 Models summary. Each row represents a model to be fit to the behavioural data, consisting of the players transactions. The table shows the model name, the number of parameters featured and the type of behaviour captured.

Model	# parameters	Behaviour captured
Random	0	No strategy
Immediate-reward Reinforcement Learning	2	Myopic / Naive
Q-learning (TD-learning)	3	Far-sighted

Chapter 5

Results

The next section of this chapter summarises the conditions studied in the decision-making tasks from Barron and Erev [2003]; Erev and Barron [2005]. After this, both a qualitative and a quantitative analysis will show that the models tested in this work achieve comparable predictive performance as the more complex system developed in Erev and Barron [2005]. The second part of this chapter will examine the results of the descriptive modelling approach developed, applied to a more realistic decision-making scenario. The data for this part of the study is obtained from an online stock trading simulation, and includes the transactions made by the subjects and the associated outcomes.

5.1 Experimental binary decision task results

36 subjects are subdivided into three groups, each characterised by different payoff conditions. The underlying distribution of payoffs for the two option in each condition are:

Condition 1

High: Draw from the Gaussian distribution $\mathcal{N}(100, 354)$

Low : Draw from the truncated Gaussian distribution $\mathcal{N}(25, 17.7)$

Condition 2

High: Draw from the Gaussian distribution $\mathcal{N}(1300, 354)$

Low : Draw from the truncated Gaussian distribution $\mathcal{N}(1225, 17.7)$

Condition 3

High: Draw from the Gaussian distribution $\mathcal{N}(1300, 17.7)$

Low : Draw from the truncated Gaussian distribution $\mathcal{N}(1225, 17.7)$

The subjects interact with the money-machine (shown in Fig. 3.1) for 200 trials, by selecting one of the two options and receiving a payoff. This quantity is visualised temporarily in the lower part of the game interface. It is then summed to the total payoff, this information is shown constantly in the lower part of the interface.

5.1.1 Predictive value

This section presents the results of the comparisons between the models proposed and the RELACS, developed in Erev and Barron [2005]. A qualitative comparison is provided in Fig. 5.1. In this figure, the observed proportions of maximisation choices are compared with the ones predicted by the models proposed in this thesis alongside the RELACS (Erev and Barron [2005]). The observed choices show an upward trend in each condition, indicating that subjects learned over the course of the trials for each problem formulation. The markers for the Pmax values are coded with the symbols: diamonds for condition 1, squares for condition 2 and triangles for condition 3. Examining the observed choices in the leftmost graph, the Pmax in condition 3 is represented in the figure by triangles and shows both the highest initial Pmax in block 1 and the strongest increase over the two blocks. Condition 3 is in fact the simplest problem, in which the two choices are easily distinguished. The other two conditions exhibit an overall increment in proportion of maximisation choices but not as marked. This is due to the fact that the outcome variability in these conditions was higher. The RELACS predictions are summarised in the rightmost graph in the same figure. The Pmax predictions trends are qualitatively similar to the ones observed from the data, but slightly differ from the observed values. The markers are all positively shifted for the RELACS panel in relation to the observed data. The Pmax predicted by the models proposed in this thesis, similarly to the RELACS, capture the overall trend of learning over the four blocks in each condition. As opposed to the RELACS Pmax predictions in the first two conditions (diamonds and squares respectively), this thesis models predict values of Pmax closer to the observed choices. It is also interesting how in condition 2 (squares) the prediction captures the shift in preference, crossing the threshold of proportions at 50%. The predictions of the proposed models for condition 3 (triangles) are negatively shifted. This could indicate a potentially slower or more exploratory learning by the proposed models in condition 3. The subjects in this condition acted quite greedily in the first block, picking the maximising option more than 75% of the trials. Their preference for such condition increased to almost 100% in the second block, indicating that their behaviour became more greedy with time. In fact, in this condition the two payoff distributions are quite distinguishable, leading to more consistent choice.

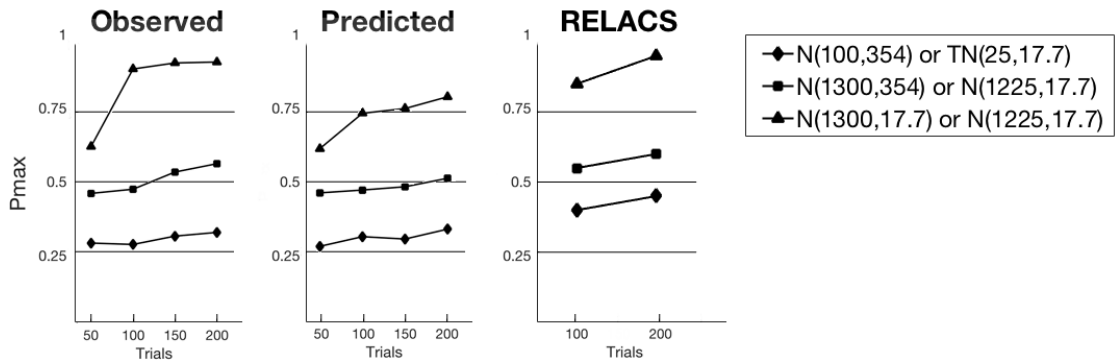

Fig. 5.1 Comparison of observed and predicted proportions of maximisation choice (Pmax). Each graph shows the evolution of the Pmax in four blocks of 50 trials each. The different symbols individuate the different conditions. On the x-axis the trials, on the y-axis the Pmax. The left panel shows the observed Pmax from the choice data. The central panel shows the Pmax predicted by the models examined in this thesis. The panel on the right reproduces the predicted Pmax from RELACS (Erev and Barron [2005]).

Considering a quantitative comparison, the MSD scores of the RELACS are reported as 0.0036 for the full model, which includes all the learning components as detailed in the previous chapter. There are also scores for variants of RELACS. A one-strategy, reinforcement learning based variant performed poorly, scoring 0.0213. For example, a two-strategy model which does not include the "slow best reply" rule presents a MSD of 0.0169. The excluded strategy represents the reinforcement learning component combined with a stochastic action selection rule. This dramatic increment in predictive score shows how important these components are. The models examined in this thesis are variations of these rules enhanced with notions from prospect theory. The MSD score achieved by this thesis models on the entire dataset is 0.0066. This score is the average of the three MSD scores achieved for every condition analysed. In the first condition the model scored 0.0021, in the second 0.0054 and in the third 0.0124. The predictive performance in the third problem is lower relatively to the other two conditions but still better than the RL or no-RL variants of RELACS. Even if the general RELACS MSD score is lower than the average MSD score for the proposed models it is necessary to point out that RELACS is a model with 4 free parameters while the proposed models all have 3 or 2 parameters. The predictive MSD score does not provide a way to account for this difference in comparing models. It is also worth noting how this predictive assessment is not the central focus of this thesis but a necessary comparison with previous modelling efforts. The rest of this work is concerned with studying the subjects' behaviour on an individual basis in order to understand its relationships with payoff variability (PV), subjective perception of value and myopic behaviour. For these reasons, the analysis in the next section follows a descriptive model fitting procedure. This is based on the maximum

likelihood estimate approach, which is a probability based assessment of the descriptive power of a model. Furthermore, this approach can be extended with AIC scores and weights, which provide a framework for model comparison, including in those cases where the models compared have different numbers of free parameters. The remaining sections of this chapter focus on providing the results obtained following this precise approach, elucidating the findings in relation to the hypotheses proposed in this chapter.

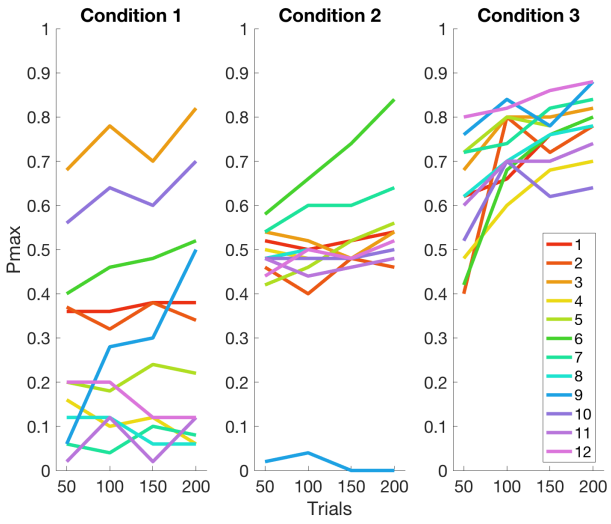

Fig. 5.2 Proportion of maximisation choices predicted by the model for each subject, aggregated over four blocks (50, 100, 150 and 200 trials marks).

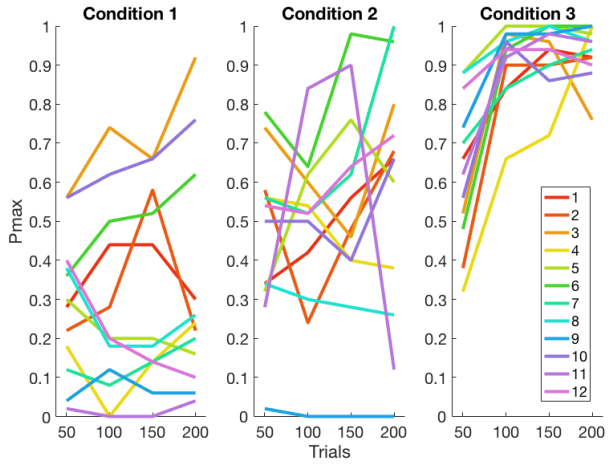

Fig. 5.3 Proportion of maximisation choices aggregated over four blocks (50, 100, 150 and 200 trials marks), reproduced from chapter 3 for convenience of comparison.

5.1.2 Descriptive value

The results of model fitting and hypotheses testing are presented and analysed in this section. The model fitting procedure estimated the descriptive performance for each of the 15 models deriving from the combinations of environments, learning rules and reward functions. A subset of these models is evaluated for each subject, in order to test the hypotheses presented at the beginning of this chapter. It is worth noting that, when comparing the models proposed with a baseline random model, 100% of the 36 subjects have at least 2 models which are better than random and have been considered for further analysis. This comparison with a baseline random model is carried out by comparing the AIC score of each tested model, with the AIC score of the parameter-less random model. The random model is assumed to have a probability $p = 0.5$ of picking either option. The MLE for such model is estimated according to the same methodology as in the other cases. The AIC is calculated with no penalty for model complexity, as the model has no free-parameters tuned.

$$\begin{aligned} AIC_{Rand} &= 2 \cdot \kappa - 2 \cdot MLE_{Rand} = \\ &= 2 \cdot 0 - 2 \cdot -138.63 = 277.28 \end{aligned} \tag{5.1}$$

where $\kappa = 0$ is the number of parameters and MLE_{Rand} is the MLE score for the random model which is calculated at the end of 4.1.14, eq. 4.15. The median AIC score for the proposed models is 201.95 while the mean is 191.5239. The AIC difference method was used to compare each reinforcement learning model fitted with the random baseline. A threshold of significance for this comparison has been identified in $\Delta_i = 5$, which is deemed as a considerable evidence in Burnham and Anderson [2002]. A further inspection with a threshold twice as high $\Delta_i = 10$ lead to only two subjects being worse than random with the majority of subjects being described by the models proposed significantly better than a random model (34 out of 36 subjects: 94.4%).

The Akaike weights methodology described in section 4.1.15 has been used to select the subset of models, in conjunction with the comparison with the baseline random model. Examples of this procedure's outcome for the first subjects in each condition, are shown in Fig. 5.4, with full figures reported in Appendix A. The figures show the AIC score for each of the 15 models fitted to each subject as a bar chart. These are presented in ascending order, with the lowest, leftmost bar representing the best fitting model. The colour of the bars indicates the configuration of state-space scenario and learning rule. The red bars represent the AIC scores of the models discarded. The coloured bars to the left represent the models whose summed Akaike weights crossed the 95% confidence set threshold. As it is often

the case in multiple model fitting procedures, there is no single best model for each subject, but a subset of the candidate models. These are the best fitting and the ones which best capture the behavioural data. The summaries of model scores and weights for the subjects 1, 5 and 10 in condition 1 are presented in tables 5.2, 5.3 and 5.4 (the rest of the subjects model comparison is reported in appendix A). In these tables the first column is divided in three acronyms according to the following scheme: Configurations are summarised with the following scheme:

Table 5.1 Abbreviations for model components

State space	Learning Rule	Reward Function
SS: Single state **FH**: Full history **LO**: Latest outcome	**AT**: Average Tracking **QL**: Q-learning	**ID**: Identity **TH**: Hyperbolic tangent **PT**: Prospect theory's value function

The three subjects reported in Fig. 5.4 and tables 5.2, 5.3 and 5.4 are chosen to illustrate the diversity of models describing the individual behaviour. Subject 1 is well fitted by five models, all based on a full-history state space. Subject 5 strategy is captured by two full-history and one single-state models, combined always with the prospect theory's value function. Subject 10 is described by 2 models based on the latest-outcome state space, both encompassing a Q-learning rule. Both predictive and descriptive analyses show that the reinforcement learning models adopted in this work achieve human-level performance on prediction and are a good descriptive account of their learning processes and actions for the task analysed.

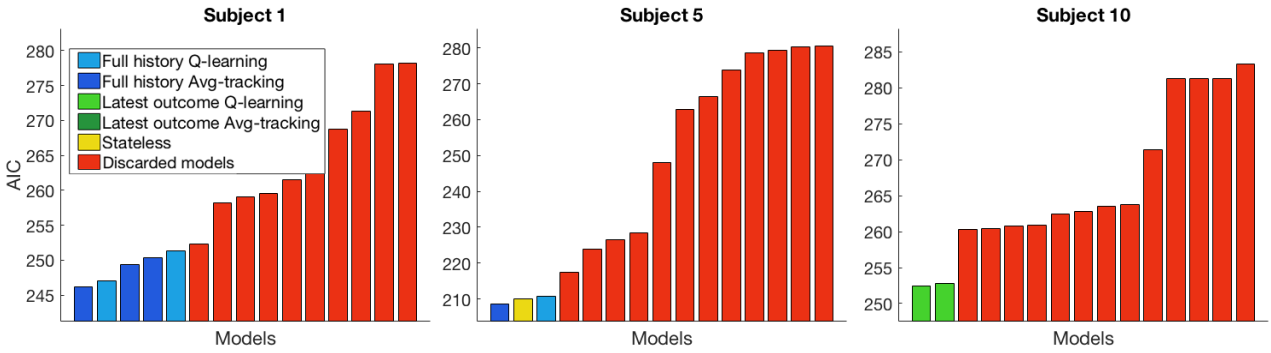

Fig. 5.4 Comparison of AIC scores of the 15 models fitted to subjects 1, 5 and 10 in condition 1. For this comparison of state-space arrangements, the three subjects have been chosen specifically to show that each subject is characterised by models based on different types of environment and learning rule.

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	1.000	0.003	0.000	-123.161	252.321	6.089	0.022
FH	**QL**	**TH**	**3**	**1.000**	**0.058**	**0.000**	**-122.692**	**251.384**	**5.152**	**0.035**
SS	AT	TH	2	1.000	0.038	0.000	-132.259	268.517	22.285	0.000
SS	AT	ID	2	1.000	0.002	0.000	-132.347	268.694	22.461	0.000
LO	AT	PT	2	0.056	0.010	0.000	-133.679	271.358	25.125	0.000
LO	AT	TH	2	1.000	0.019	0.000	-137.078	278.157	31.924	0.000
LO	QL	PT	3	0.919	0.002	1.000	-126.542	259.085	12.852	0.001
FH	**QL**	**PT**	**3**	**1.000**	**0.003**	**1.000**	**-120.499**	**246.998**	**0.766**	**0.317**
LO	QL	ID	3	0.817	0.002	1.000	-126.766	259.532	13.299	0.001
FH	**AT**	**PT**	**2**	**1.000**	**0.003**	**0.000**	**-121.116**	**246.233**	**0.000**	**0.466**
FH	**AT**	**ID**	**2**	**1.000**	**0.003**	**0.000**	**-123.161**	**250.321**	**4.088**	**0.060**
FH	**AT**	**TH**	**2**	**1.000**	**0.058**	**0.000**	**-122.692**	**249.384**	**3.151**	**0.096**
LO	QL	TH	3	0.813	0.038	1.000	-126.090	258.179	11.947	0.001
SS	AT	PT	2	0.062	0.014	0.000	-128.736	261.472	15.239	0.000
LO	AT	ID	2	1.000	0.001	0.000	-137.044	278.087	31.855	0.000

Table 5.2 Models summary: problem 1 - subject 1

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.001	1.000	1.000	-128.453	262.907	54.218	0.000
FH	QL	TH	3	0.022	1.000	1.000	-121.058	248.117	39.428	0.000
SS	AT	TH	2	0.873	0.014	0.000	-137.683	279.365	70.676	0.000
SS	AT	ID	2	1.000	0.000	0.000	-138.103	280.206	71.517	0.000
LO	AT	PT	2	0.004	0.140	0.000	-109.989	223.979	15.290	0.000
LO	AT	TH	2	0.003	1.000	0.000	-137.364	278.729	70.040	0.000
LO	QL	PT	3	0.276	0.007	1.000	-105.719	217.437	8.748	0.007
FH	**QL**	**PT**	**3**	**0.007**	**0.091**	**0.000**	**-102.344**	**210.689**	**2.000**	**0.195**
LO	QL	ID	3	0.315	0.004	1.000	-111.161	228.322	19.633	0.000
FH	**AT**	**PT**	**2**	**0.007**	**0.091**	**0.000**	**-102.344**	**208.689**	**0.000**	**0.529**
FH	AT	ID	2	0.001	1.000	0.000	-131.249	266.497	57.808	0.000
FH	AT	TH	2	0.004	1.000	0.000	-134.978	273.956	65.267	0.000
LO	QL	TH	3	0.271	0.108	1.000	-110.226	226.451	17.762	0.000
SS	**AT**	**PT**	**2**	**0.007**	**0.103**	**0.000**	**-103.023**	**210.046**	**1.357**	**0.269**
LO	AT	ID	2	0.000	1.000	0.000	-138.231	280.461	71.772	0.000

Table 5.3 Models summary: problem 1 - subject 5

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.000	1.000	0.000	-128.227	262.454	9.963	0.003
FH	QL	TH	3	0.002	1.000	1.000	-128.381	262.762	10.271	0.003
SS	AT	TH	2	0.002	1.000	0.000	-128.371	260.742	8.252	0.008
SS	AT	ID	2	0.000	0.700	0.000	-128.172	260.344	7.854	0.010
LO	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	28.769	0.000
LO	AT	TH	2	0.017	0.309	0.000	-129.781	263.562	11.072	0.002
LO	QL	PT	3	1.000	0.001	1.000	-132.700	271.401	18.910	0.000
FH	QL	PT	3	0.000	0.000	0.719	-138.629	283.259	30.769	0.000
LO	**QL**	**ID**	**3**	**0.144**	**0.004**	**1.000**	**-123.245**	**252.490**	**0.000**	**0.509**
FH	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	28.769	0.000
FH	AT	ID	2	0.000	0.700	0.000	-128.229	260.458	7.967	0.009
FH	AT	TH	2	0.002	1.000	0.000	-128.425	260.850	8.359	0.008
LO	**QL**	**TH**	**3**	**0.145**	**0.081**	**1.000**	**-123.381**	**252.762**	**0.272**	**0.445**
SS	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	28.769	0.000
LO	AT	ID	2	0.018	0.013	0.000	-129.909	263.818	11.328	0.002

Table 5.4 Models summary: problem 1 - subject 10

Hypothesis 1

The first hypothesis is concerned with clarifying whether decision-makers in binary choice tasks use historical payoff information to assess their current status when making decisions. This investigation stems from prospect theory's idea that decision-makers have a shifting reference point that moves according to previous experiences. This hypothesis proposes that subjects take into account the historical information shown by the money machine (Fig. 3.1 in chapter 3), in order to decide which action to take next. This information is the outcome of the latest choice and the sum of the payoffs accumulated over the course of all previous interactions. This section focuses on the results of this hypothesis' tests. To test this hypothesis with a descriptive modelling approach, three state-spaces have been identified (section 4.1.7) and will be used to represent the various approaches the decision-makers could adopt in the task.

The single-state scenario represents the case in which a subject does not consider previous outcomes in any form and therefore behaves only according to the perceived value of the available options. The alternative scenarios are both represented by two-state modelling; the latest outcome model captures the behaviour influenced by the last payoff value while the full-history scenario captures the behaviour influenced by the total accumulated payoffs information. The null hypothesis is that subjects do not use information about previous outcomes when making decisions. This can be investigated within this modelling framework by testing whether there is a difference in the distribution of single-state models and two-state

variations. Alternatively, a significant difference in the number of subjects described by either of the improved state-spaces would reject the null hypothesis. The prediction linked to prospect theory's shifting reference point suggestion is that the state-space environments based on previous information are more likely to be descriptive of the behavioural data because these scenarios describe the subjects' decision-making more realistically.

This hypothesis is centred on comparing the single-state scenario to the enhanced state-spaces. The full-history scenario is compared with the single-state one and with the latest outcome. To test this hypothesis, the choice data for each subject is fitted with each of the models deriving from the combinations of learning rules, reward functions and state-spaces. This setup allows the subset of best descriptive models to be identified. The ones which are best at describing the data are identified according to the AIC weights criterion. Firstly, the best model, which is the one with the lowest AIC score, is identified as the most descriptive. Then the Akaike weights for each model are calculated, as described in section 4.1.15, eq. 4.18. These are used to determine the subset of models which are less likely than the best model (lowest AIC), but are still likely to provide significant description of the choice data. The distance between the AIC score of the best model and the others is used to calculate the weights which represent a probabilistic account of the evidence in favour of each model. This methodology provides a confidence set for the best model, based on AIC scores which are grounded in the Kullback-Leibler information theory; this confidence set can be used in a similar way as to the evaluation of a parameter confidence interval based on a model and a set of data (Burnham and Anderson [2002], p. 169). The Akaike weights method are also helpful in estimating parameters more reliably, avoiding abruptly disregarding potentially meaningful parameter sets. This is crucial for testing the other hypotheses. This procedure results in a different number of models selected for each subject. An alternative approach would be restricting the model selection with a specific threshold, for example accepting only the three best models for each subject. This cut-off method was not adopted because it would introduce an arbitrary bias in the model selection unsubstantiated by literature. Moreover, arbitrarily capping the number of models considered would also break the Akaike weights confidence set theory.

The next step is to estimate the frequency of each type of scenario. The configurations tested and summarised in table 4.2 result in five configurations with equal number of models being tested. The first configuration is single-state and is combined with average-tracking learning rule (models 1-3 in table 4.2). For the baseline state-space scenario, this is the only combination possible. The alternative rule Q-learning, being a type of temporal-difference learning, requires more than one state to be applied. Therefore, both the two-state scenarios are combined with the two learning rules. This TD-learning rule is included in the modelling

to accommodate the possibility that subjects behave in a more far-sighted fashion, in the case their perception is described with a two-state environment. The resulting configurations combinations are latest-outcome combined with the average-tracking (models 4-6), latest-outcome and Q-learning (7-9), full-history in combination with average-tracking (10-12) and full-history combined with Q-learning (13-15).

The frequency of the model combinations fitted to the data and selected by the criteria described above resulted in the following distributions:

Single-state & Average-tracking: 57 models, 17.87%;

Full-history & Average-tracking: 60 models, 18.81%;

Full-history & Q-learning: 82 models, 25.71%;

Latest-outcome & Average-tracking: 50 models, 15.67%;

Latest-outcome & Q-learning: 70 models, 21.94%.

These results are graphically presented in Fig. 5.5.

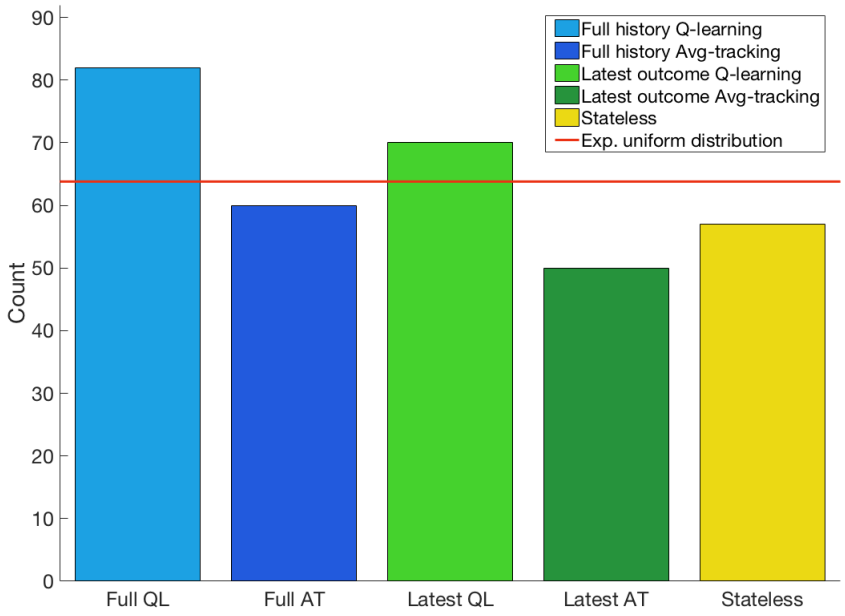

Fig. 5.5 The comparison of the state-space configurations. Each bar represents the model frequency for a specific configuration obtained selecting the subset of models with the sum of Akaike weights $\geq 95\%$. The red line is the expected count for a uniform distribution. Each bar is colour coded with the model type.

Multiple Pearson Chi-squared (χ^2) "goodness of fit" tests have been carried out on the frequencies of the subset of subjects' best models in order to test the first hypothesis. The results of the tests confirm that configurations of two-state scenarios combined with Q-learning are more likely to be representative of the decision-makers' behaviour when compared with single-state configuration or latest-outcome scenario combined with average-tracking learning rule:

The overall five-way χ^2 test shows a significant difference across the five categories:

Combined five categories: $\chi^2(4, N = 319) = 9.730, p = 0.045 < 0.05$

Further pairwise tests between the full-history model with Q-learning (QL) rule and the other configurations provide evidence against the single-state version and the latest-outcome model with average-tracking (AT):

Full-history QL vs Single-state: $\chi^2(1, N = 139) = 4.496, p = 0.034 < 0.05$;

Full-history QL vs Latest-outcome AT: $\chi^2(1, N = 132) = 7.758, p = 0.005 < 0.05$.

Full-history QL vs Full-history AT: $\chi^2(1, N = 142) = 3.409, p = 0.064 > 0.05$.

Full-history QL vs Latest-outcome QL: $\chi^2(1, N = 152) = 0.947, p = 0.330 > 0.05$.

The comparison between full-history QL and full-history AT does not indicate a significant difference. The same result emerges from the comparison between full-history QL and latest-outcome QL. Moreover, none of the remaining comparisons achieves a significant result (full details in the box at the end of this section).

These results provide evidence that a configuration encompassing a two-state version of the environment, in conjunction with a long-term learning strategy, is more representative than either a single-state version or an immediate reward learning strategy paired with a short-term memory state-space. From the rest of the comparisons there is no strong indication on which state-space modelling is more representative of the subjects' reference system in this binary choice task.

The results obtained reject the null hypothesis, that subjects do not use information about previous outcomes when making decisions. Furthermore, two of the comparisons - FH-QL vs single-state and FH-QL vs LO-AT - provide support for the alternative hypothesis, that subjects use previous outcomes information offered from the interface when making decisions in this binary task. The rest of the results do not provide enough evidence to support other alternative explanations.

Even if these findings point in the direction of subjects showing a higher degree of sophistication in their learning and decision-making processes, they also indicate that more

research is needed to clarify what type of information is taken into account by decision-makers when faced by similar binary decision tasks. Even if it was not possible to identify one single method the decision-makers use to assess their current status, it is clear from the results that historical information influences the reference point decision-makers use. This indication is of crucial importance for future developments. Future environment design for similar tasks in descriptive modelling studies should take into account the information provided to the subjects, because the current work indicates that it affects the DM's behaviour.

This descriptive investigation allowed for a comparison between models based on different reinforcement learning environments. Each of the tested scenarios represented a possible approach the subjects could use. The results indicate that in many instances subjects are well modelled with environments considering previous outcomes. The integration of such information into a reinforcement learning setup represents a novel approach into the study of decisions from experience, and an encouraging link with prospect theory. According to the results, prospect theory's shifting reference point does play a role for a number of subjects in this task. This subjective shifting phenomenon is in-line with the indications from the literature. In fact, subjects shift their preference depending on their previous experiences, as predicted by the "house-money effect" and the "break-even effect". These phenomena are described in chapter 2, section 2.1.2.

Hypothesis 1: State-space configuration.

FH-QL vs SS: $\chi^2(1, N = 139) = 4.496, p = 0.034 < 0.05$ significant;
FH-QL vs LO-AT: $\chi^2(1, N = 132) = 7.758, p = 0.005 < 0.05$ significant;
FH-QL vs FH-AT: $\chi^2(1, N = 142) = 3.409, p = 0.064 > 0.05$ not significant;
FH-QL vs LO-QL: $\chi^2(1, N = 152) = 0.947, p = 0.330 > 0.05$ not significant;
FH-AT vs SS: $\chi^2(1, N = 117) = 0.077, p = 0.782 > 0.05$ not significant;
FH-AT vs LO-QL: $\chi^2(1, N = 130) = 0.769, p = 0.380 > 0.05$ not significant;
FH-AT vs LO-AT: $\chi^2(1, N = 110) = 0.909, p = 0.340 > 0.05$ not significant;
LO-QL vs SS: $\chi^2(1, N = 127) = 1.331, p = 0.249 > 0.05$ not significant;
LO-AT vs SS: $\chi^2(1, N = 107) = 0.458, p = 0.499 > 0.05$ not significant.

All state-spaces: $\chi^2(4, N = 319) = 9.730, p = 0.045 < 0.05$ significant.

Hypothesis 2

This section presents the results of the tests relative to the second hypothesis. The hypothesis is concerned with the way subjects perceive and internalise the values of the available options. The prediction, expressed also in section 4.1.2, is that prospect theory's subjective value function would be the most descriptive among the three reward functions tested; the other two being the identity function (raw payoffs) and the hyperbolic tangent (fully saturating). In order to answer the question on whether PT's value function is the most descriptive account of subjects' payoffs perception for this task, a similar procedure to hypothesis 1 has been carried out. A subset of best models has been identified for each subject according to the AIC weights methodology. In a similar manner to the previous test the models considered are the ones with the lowest AIC and whose Akaike weights summed crosses the 95% confidence threshold. Using these model subsets as a starting point, the frequencies for each reward function have been estimated. A Pearson Chi-squared χ^2 (goodness of fit) test is performed on such frequencies. Fig. 5.6, provides a graphical representation of the distributions of the three reward function classes tested.

The first three-way Chi-squared test carried out resulted in $\chi^2(2, N = 319) = 8.418$ and $p = 0.004 < 0.05$. This test provides strong and statistically significant evidence against the null hypothesis, that there is no difference in the representativeness of the reward functions tested. The pairwise tests confirm the evidence against the hyperbolic tangent (Tanh) function when compared to PT value function. The same cannot be said for the comparison between raw and saturating payoff transformations.

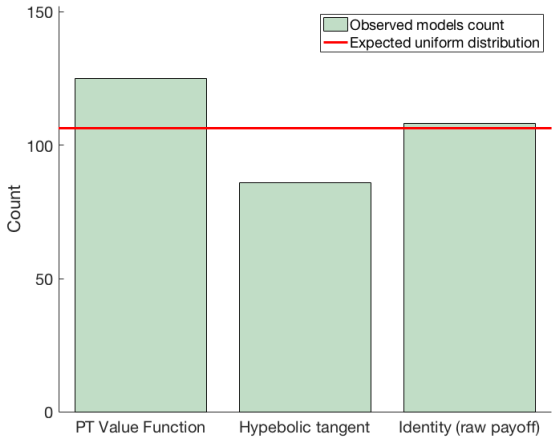

Fig. 5.6 The comparison of the subjective reward functions configurations. The bars depict the frequency for a specific configuration obtained in the same way as in the state-space test. The red line is the expected count for a uniform distribution in each comparison panel.

Moreover, even if the frequency of selected models featuring PT value function is higher than the ones with the identity function, this difference is not statistically significant and does not provide enough evidence to undoubtedly consider PT value function more representative than raw rewards.

These results provide an indication that decision-makers, in this particular task, do not perceive utility as a fully saturating function of payoffs. The same results cannot confirm that PT value function has a stronger descriptive power when compared to unaltered payoff values. One potential explanation for such lack of descriptive power is that the subjective value function fitted to the individual choices in this study followed the parametrisation with values from literature. These values were estimated from data obtained in decisions from description scenarios and could suffer from the discrepancy between decisions from experience and decisions from description. More work is required in this direction, which will be examined in chapter 6, when discussing the future developments.

Hypothesis 2: Reward function.

Tanh vs Identity: $\chi^2(1, N = 181) = 2.9227, p = 0.087 > 0.05$ not significant;
Tanh vs Identity: $\chi^2(1, N = 181) = 2.9227, p = 0.087 > 0.05$
PT vs Tanh: $\chi^2(1, N = 199) = 8.447, p = 0.003 < 0.05$ significant.

Three-way test:: $\chi^2(2, N = 319) = 8.418$ and $p = 0.004 < 0.05$ significant;

Hypothesis 3

This section presents the results for the analysis of the third hypothesis, focused on the study of the relationship between payoff variability (PV) and the subjects' speed of learning. The payoff variability is calculated as the standard deviation of the observed outcomes (eq. 4.1). The PV measures the variability of the outcomes experienced by a subject during the 200 trials. The learning speed is estimated with a free-parameter in the learning rule of the models fitted to the subject's choice data. It provides a measure of the quantity of information the subjects consider when learning about the option values, ultimately describing how fast they learn.

Because the subjects were subdivided in three substantially different conditions, this hypothesis is broken down into three tests, one for each condition, plus a cumulative test. This is done because each condition presents a payoff variability range different from the other conditions by design. The data is presented in Fig. 5.7, panel a) in the form of a scatter

plot with marginal distributions. The data points for each condition are colour-coded and are shown with different symbols. These points are identified as the pairs of outcome standard deviation experienced by a subject and the associated learning rate parameter α, which estimates the subject's speed of learning. The parameter values are combined into a single figure as the weighted average of the parameter estimates within each model selected by the Akaike weights criterion. This procedure allows inference based on a single estimate, which is the combination of the parameters estimated for each model within the 95% confidence set. An alternative approach, which is widely used in literature even if its limitations have been pointed out in multiple instances (Burnham and Anderson [2002]; Wagenmakers and Farrell [2004]), is to use the parameter estimate from the single best model (lowest AIC score); doing so would neglect the importance of the remaining models and their parameters estimates, which are likely to still be good candidates for describing the data. Subjects were fitted by a different number of models, as shown in Fig. 5.4, and in the tables A.2, A.6 and A.11. This does not affect the results because the model's parameter estimates were combined into a single value with the Akaike weights effectively balancing the amount of evidence for each of these estimates. The parameter which is then identified for a subject is an estimate that keeps into consideration the amount of evidence in favour of each estimate from the models subset. This is an important part of the methodology adopted which is used for this hypothesis as much as in the next ones.

In order to estimate the correlation of two quantities with the Pearson test of correlation the data for each variable is required to be normally distributed. Three tests of normality have been carried out on each variable: Kolmogorov-Smirnov, Anderson-Darling and Bera-Jarque tests. The Pearson test of correlation could not be adopted because all the normality tests results are negative, confirming the data being non-normally distributed. As the normality assumption does not hold true, the Spearman's test of rank correlation is used instead. This non-parametric test does not assume the data follows a particular distribution but relies instead on the ranks of the values, measuring the direction and the strength of the relationship between the two variables. It is important to remember that while Pearson's test establishes the linear correlation between two variables, Spearman's correlation instead focuses on monotonic relationships.

The null hypothesis for this relationship is that there is no correlation between the two quantities. The correlations between payoff variability with the parameters describing learning speed have been tested for each condition with the following results.

The results indicate that there is not enough evidence to reject the null hypothesis, when considering each condition separately. The Spearman's rank correlation test in condition 1 resulted in $\rho = 0.042$ with non significant p value $= 0.897 > 0.05$. The same test on

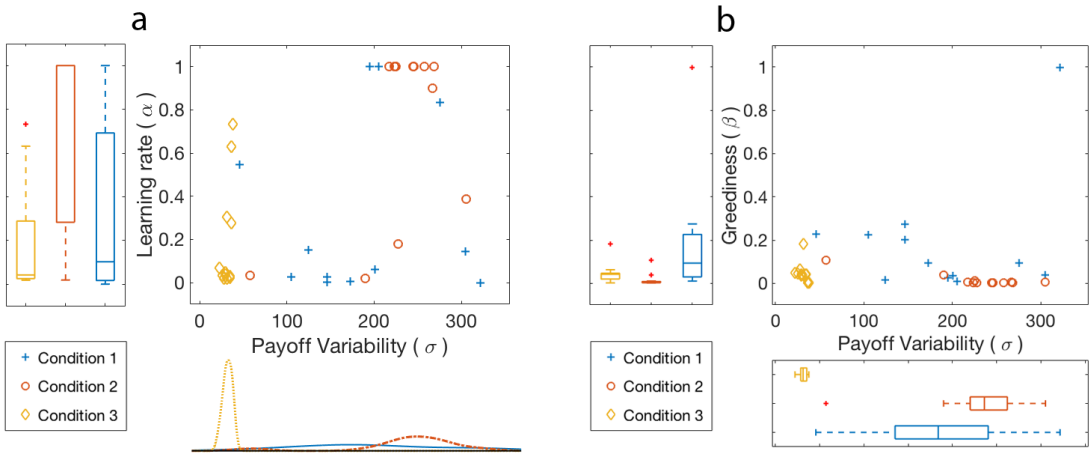

Fig. 5.7 For both panels, on the x-axis the Payoff Variability (subjective observed outcomes standard deviation). On the y-axis in panel a) the values of learning rate parameter estimated by model fitting procedure and Akaike weighted average of the best models; in panel b) the values of inverse temperature parameter (action-selection greediness) estimated in the same way as the learning rate. Data for each comparison is plotted on the same scatter and is colour and symbol coded for each condition. On the left hand side of each panel the box-plot of the marginal distribution for the parameter plotted. On the bottom of each panel the marginal distribution of the outcomes standard deviations values. As this information is presented twice, panel a) shows it in the form of kernel density plot while panel b) shows it as box-plot.

condition 2 produces $\rho = 0.281$ with non significant p value $= 0.377 > 0.05$. Finally, the test on the data for condition 3 yielded $\rho = 0.476$ with non significant p value $= 0.121 > 0.05$. It is worth noting however, that by pooling together the data across the three condition and analysing it in the same way, the results obtained are much different. The Spearman's rank correlation test across the three conditions is $\rho = 0.443$ with p value $= 0.007 < 0.05$, indicating moderate positive correlation with strong statistical significance.

According to the outcome of this test, subjects who experienced more variability were quicker in their learning. This indication is also noticeable from a graphical inspection of Fig. 5.7 panel a); in fact, the individuals who experienced low variability are clustered in the bottom left corner of the figure, while the subjects who encountered more variation appear grouped in the top right corner. This result is in contrast with previous suggestions from Erev and Barron [2005], that payoff variability has an impairing effect on learning. The datapoints in condition 2 and 3 appear clustered in distinctly opposite areas: Fig. 5.7 panel a), red circles in top-right corner and yellow diamonds in bottom-left corner. The reason why significance is achieved for the entire dataset is that, when pooling the data together, these datapoints lead to a dataset with a more marked ranking and orientation, as opposed to

condition data subsets. An attempt to reconcile the findings obtained so far with the results of the next investigation is provided at the end of the next section.

Hypothesis 3: correlation between payoff variability and speed of learning (α).

Condition 1: $\rho = 0.042$, p value $= 0.897 > 0.05$ not significant;
Condition 2: $\rho = 0.281$, p value $= 0.377 > 0.05$ not significant;
Condition 3: $\rho = 0.476$, p value $= 0.121 > 0.05$ not significant.

All conditions: $\rho = 0.443$ with p value $= 0.007 < 0.05$ significant.

Hypothesis 4

This hypothesis focuses on studying the relationship between payoff variability and the degree of greediness in decision-making behaviour exhibited by the subject in their interactions. In decision-making tasks' action-selection can be divided in exploitative and explorative. The former consists of choices favouring the option believed to be the best at a certain point in time, the latter selects alternatives that are believed not to be the best in order to acquire more knowledge about them, in case they are actually better than they are believed to be. Exploitative behaviour is also described as greedy because it is concerned with maximising the immediate reward and is based on the current level of information; on the other hand, extreme explorative behaviour results in random action-selection. The choice data for each subject lies somewhere on this continuum. Greediness (or randomness) is quantified by a free parameter in the probabilistic action-selection policy adopted in this thesis modelling. This parameter is called β and it is estimated as previously described for the learning-speed parameter, with the weighted average of the parameters identified by the subset of best models for each subject. The closer the values of β are to 0, the more random the behaviour, while high values indicate a greedier behaviour. The null hypothesis for this part of the investigation is that there is no correlation between the amount of payoff variability experienced by a subject and the corresponding greediness in choice behaviour. In other words, there is no correlation between the standard deviation of the outcomes attained by a subject and the greediness parameter estimated for that subject. Similarly to the previous analysis on payoff variability and learning speed, this hypothesis is broken down into three tests, since the subjects were subdivided in three conditions designed with different observable variability, plus a fourth cumulative test. The data about this hypothesis is presented in Fig. 5.7, panel b) and follows the same legend as in panel a). The datapoints identify the pairs of payoff

standard deviation and greediness estimate for each subject and are shown with different symbols and colours according to the experimental condition they belong to.

The Spearman rank correlation test in condition 1 indicates that there is no correlation between the two quantities with $\rho = -0.182$ and p value $= 0.572 > 0.05$ not significant. For condition 2 the test results in a negative correlation score $\rho = -0.566$ with non-significant p value $= 0.059 > 0.05$. The test on date in condition 3 produces even stronger negative correlation with $\rho = -0.657$, this time with a significant p value $= 0.024 < 0.05$. A zoomed in version of the scatter plot, for the datapoints from condition 3, is shown in Fig. 5.8.

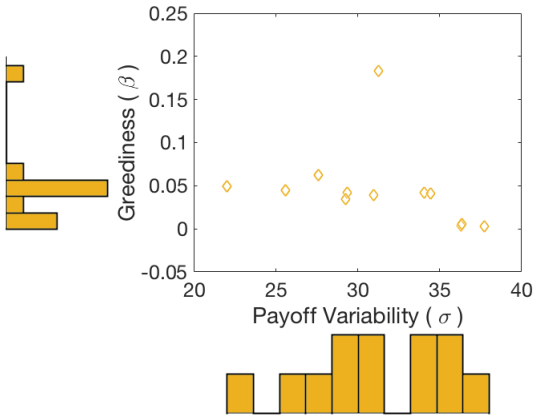

Fig. 5.8 Scatter plot with marginal distribution histograms of payoff variability against greediness parameter estimate, for subjects in condition 3.

The negative ρ indicates that subjects experiencing low payoff variability choose more greedily while for those subjects who experienced higher payoff variability behaviour was more exploratory. This relationship is not significant in the first two conditions and statistically significant in the third. A potential explanation of these results lies in the design of the third condition, portrayed in Fig. 5.9, Condition 3, where the two options' payoff distributions are quite distinct and only minimally overlap.

Among the subjects in condition 3, the ones who experienced more payoff variability were less greedy and more random in their action selection; this finding is in line with previous indications from Erev and Barron [2005]. The subjects in the conditions 1 and 2 have experienced a much greater payoff variability due to the structure of payoffs distributions in those conditions. Following previous suggestions from literature (Erev and Barron [2005]), more experienced payoff variability leads to an impairment in learning, slowing it down. The analysis carried out in this work does not support this, providing little evidence against it. Analysing the entirety of the scatter plot in Fig 5.7 panel a), it is worth noting that subjects who experienced more variability are generally better fitted by faster learning models (red

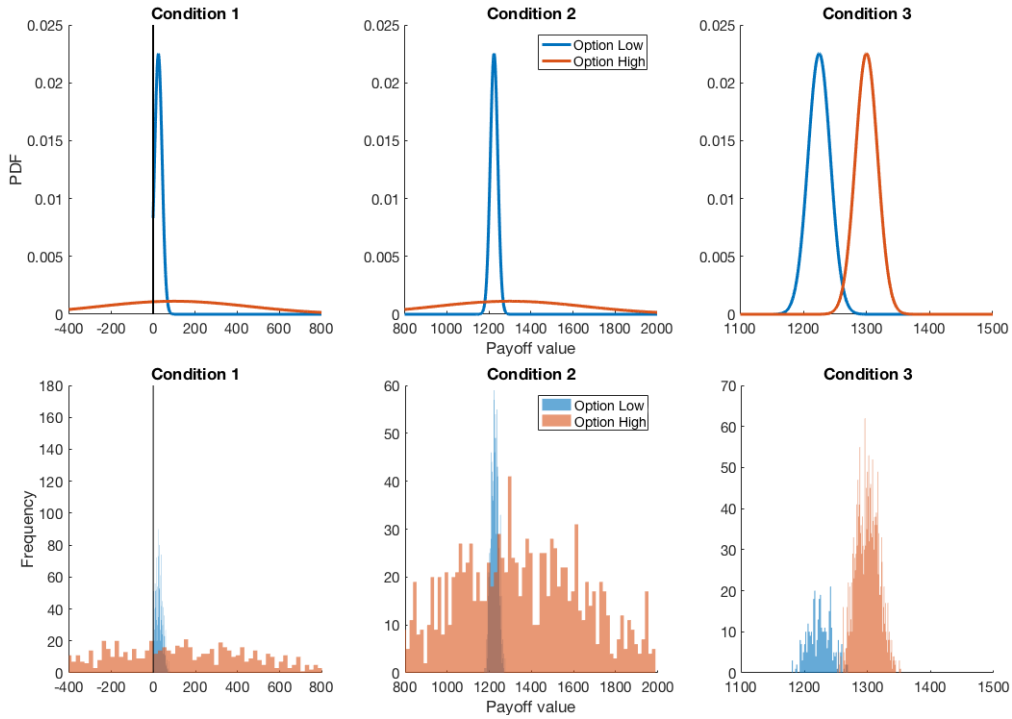

Fig. 5.9 Payoff probability density functions for each condition (top row) and observed payoffs distributions (bottom row). In blue the PDF and observed outcome frequency of the option yielding low expected payoffs, in red the high option. In condition 1 the outcomes of option low are truncated in 0 therefore never yielding negative rewards.

circles, top right corner), as opposed to subjects who experienced lower variability and are better fitted by models featuring slow learning rates (yellow diamonds, bottom left corner). Moving on to panel b) in the same figure, the subjects better fitted by fast learning models are also described with high randomness (red circles, bottom right corner). A possible explanation for these results is that those individuals who experienced high payoff variability tended to become more random in their decision-making, exploring their options and trying to achieve a better understanding of which option was the best. This effect, which confirms the indications in Erev and Barron [2005], has been proven to be significant in the subset of subjects facing the easiest of the three conditions. At the same time, these subjects appear to be quicker in learning their perceived best action, which not necessarily coincides with the true high return option. When considering the subjects from all conditions, the Spearman's test results in $\rho = -0.343$ with significant p value $= 0.041 < 0.05$, indicating a moderate negative correlation between subjects' experienced payoff variability and their exhibited greediness in action-selection strategy. The direction of this result agrees with the trend of the other test results. Moreover, together with the result for the last condition it provides evidence towards a negative relationship, indicating that subjects who experienced more

variability became less greedy, and more random. This finding corroborates the indications from literature (Erev and Barron [2005]).

Hypothesis 4: correlation between payoff variability and action-selection behaviour (β).

Condition 1: $\rho = -0.182$, p value $= 0.572 > 0.05$ not significant;
Condition 2: $\rho = -0.566$, p value $= 0.059 > 0.05$ not significant;
Condition 3: $\rho = -0.657$, p value $= 0.024 < 0.05$ significant.

All conditions: $\rho = -0.343$ with p value $= 0.041 < 0.05$ significant.

Hypothesis 5

The fifth and last hypothesis of this chapter regards the relationship between subjects' performance in the task and the degree of far-sightedness captured by the models fitted to the choice data. This last quantity is based on the estimation of the subjects' interest in future rewards when learning the task and it is encapsulated by a free-parameter called discount factor, or γ, in the models fitted to the subjects' choice data. In the reinforcement learning framework, and specifically in temporal difference learning, this parameter regulates the amount of future rewards considered during an agent's learning process. A value $\gamma = 0$ represents fully myopic behaviour with an agent being concerned only with obtaining immediate rewards. Values close to 0 indicate somewhat myopic tendencies, while values close to 1 describe far-sighted behaviour. In the context of the descriptive modelling performed in this work, this parameter provides a numerical representation of the myopia or far-sightedness of the subjects behavioural data.

Like in the previous hypotheses investigations, the estimation of this parameter is achieved by fitting the 15 models, identifying the most descriptive according to Akaike weights method and averaging these estimates according to the weights (section 4.1.15, eq. 4.19). Even if the models in the dataset do not all feature this parameter the estimation can still be performed. For those models based on average-tracking as learning rule, the discount factor parameter is not considered because this learning rule can be considered a special case of Q-learning where the discount factor parameter is set to 0, effectively reducing to a nested version as described in Burnham and Anderson [2002], pp 150-152.

This hypothesis test is structured in a similar way to the procedures adopted for the previous hypotheses. The model's parameter of interest for this comparison is the discount factor, while the other variable for the correlation test is the task performance, which can

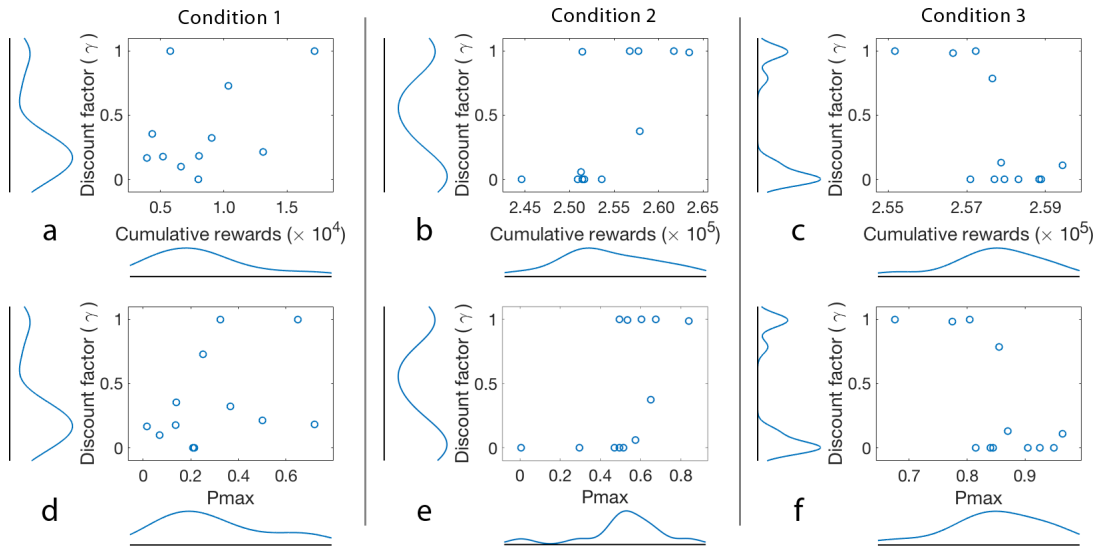

Fig. 5.10 On the y-axis, for all panels, the value of the discount factor parameter γ estimated by model fitting procedure and Akaike weighted average of the best models. Values of γ close to 1 indicate more far-sighted behaviour, while values close to 0 indicate more myopic behaviour. Each column depicts a condition. Top row shows the scatter plot of these values along with the cumulative rewards performance measure for each subject. The bottom row shows the scatter plot of discount factor values and corresponding Pmax (proportion of maximisation choices, section 4.1.16, eq. 3.1) for each subject. On the left hand side and at the bottom of each panel is pictured the marginal distribution of each variable in the form of kernel density estimates.

be measured in two different ways. From a subject's perspective, the performance can be interpreted and quantified as the final cumulative outcome, consisting of the sum of the 200 payoffs received at each time step. From the omniscient standpoint of the experiment designer, the performance can be identified as the proportion of maximisation choices (Pmax): a subject is performing well if she chooses the option with the highest expected return even if this yields poor short-term results. This measure has been used in literature (e.g. Barron and Erev [2003]; Erev and Barron [2005]; Erev et al. [2012]; Thaler et al. [1997]) and was also used to compare the predictive power of the models developed in this thesis against previous work (Erev and Barron [2005]). Because the structure of the experiment contemplated different ranges of payoffs in each condition, the correlation analysis is performed on each subset of subjects belonging to the three conditions. For completeness the correlation test is also performed on the aggregated subjects dataset. The null hypothesis is that there is no correlation between how myopic or far-sighted subjects are and their performance, being it quantified as the amount of money accrued or the portion of maximisation choices performed

over the course of the task. The alternative hypothesis is that there is a significant correlation between these quantities.

The scatter plot of the data for each condition is presented in Fig. 5.10. Top row panels a), b) and c) show the scatter plot and kernel density plots for the three conditions (first to third, left to right) when comparing discount factor γ and cumulative rewards performance. The figures on the bottom row, panels d), e) and f) show the comparisons for the three conditions, in the same order as in the previous three panels, of discount factor against proportion of maximisation performance.

Four Spearman's rank correlation tests have been carried out on the data subsets for each of the three conditions and on the entire dataset. Each datapoint represents the pair of estimated discount-factor parameter and performance, measured as sum of the outcomes received, for each subject. This first measure of performance can be considered as a subjective perception of success, because it is based on the information the subjects were able to see on the interface of the money-machine.

Furthermore, this was the quantity which translated the task performance into real-life monetary reward at the end of the experiment (Barron and Erev [2003]; Thaler et al. [1997]). The test in condition 1 resulted in $\rho = 0.315$ with non-significant p value $= 0.318 > 0.05$. The results for condition 2 are $\rho = 0.552$ with p value $= 0.067 > 0.05$, not significant. The third condition data resulted in a negative correlation with $\rho = -0.476$, again with non-significant p value $= 0.121 > 0.05$. If the data from the three conditions is pooled together and analysed in the same way, it results in $\rho = -0.008$ with a p value $= 0.965 > 0.05$. This final result indicates that the contrasting trends for the three conditions, positive for the first two and negative for the third, produces an uncorrelated and non significant dataset according to the Spearman's test. The results provided so far show that there is not enough evidence to support the alternative hypothesis in each of the three conditions and in the aggregated dataset, when using cumulative rewards as measure of the performance.

Because the performance can also be identified to be the amount of correct decisions a subject makes this hypothesis has been also tested with a measure of the performance based on this metric. This measure of the performance is not immediately accessible to the subjects, as they are not provided with a description of which option yields the best payoff on average. Subjects could find this information only by exploring the options and gauging information about their quality. Therefore, the Pmax represents a measure of performance from the experiment designer's point of view. These correlations tests have been evaluated with the same procedure as in the previous four tests. The same discount-factor estimates are used, but instead of the sum of payoffs obtained by the subject, the proportion of *high* choices has been used.

Four Spearman's correlation tests have been performed, one for each condition plus the test on the aggregated data. In the first condition the test resulted in $\rho = 0.508$ with non significant p value $= 0.091 > 0.05$ In the second condition a marginally significant, moderate positive correlation was found with $\rho = 0.581$ and p value $= 0.047 < 0.05$. The third condition test resulted in a non significant negative correlation, $\rho = -0.419$ and p value $= 0.177 > 0.05$. A final Spearman's test on the data aggregated from the three condition resulted in $\rho = -0.010$ with p value $= 0.955 > 0.05$. This final correlation test suffers from the same aggregation effect as in the previous tests. The aggregation of data from the conditions which presented opposite correlation trends resulted in a very weak correlation with the trends from the different conditions, substantially cancelling each other and producing the least significant result.

Most of these tests do not show enough evidence to reject the null hypothesis. Except for the second condition with Pmax as performance measure, all comparisons were non statistically significant at the 0.05 level. Therefore it can be concluded that for both measures of performance and for most of the conditions tested there is no correlation between myopic behaviour and task performance, the null hypothesis cannot be rejected. A possible explanation of these results is that for this type of tasks the time horizon is not long enough to influence subjects' decision-making too much, because the reward obtained from each action is obtained immediately. Therefore, as the correlation tests results suggest, there is no relationship between the degree of myopia subjects are best described with and their achievements measured either internally or externally.

In this regard it would be interesting to investigate a more realistic scenario in which an action's repercussion is not immediate but delayed. Investigating a task similar to the one examined in Thaler et al. [1997], could in fact produce more informative findings on the effect of myopic behaviour. The next chapter is focused on this type of investigation. Building on the conclusions drawn from the results of this chapter and extending them to a realistic scenario.

Hypothesis 5: correlation between myopic behaviour (γ) and task performance (cumulative rewards).

Condition 1: $\rho = 0.315$, p value $= 0.318 > 0.05$ not significant;
Condition 2: $\rho = 0.552$, p value $= 0.067 > 0.05$ not significant;
Condition 3: $\rho = -0.476$, p value $= 0.121 > 0.05$ not significant.

All conditions: $\rho = -0.008$ with p value $= 0.965 > 0.05$ not significant.

Hypothesis 5: correlation between myopic behaviour (γ) and task performance (Pmax).

Condition 1: $\rho = 0.508$, p value $= 0.091 > 0.05$ not significant;
Condition 2: $\rho = 0.581$, p value $= 0.047 < 0.05$ significant;
Condition 3: $\rho = -0.419$, p value $= 0.177 > 0.05$ not significant.

All conditions: $\rho = -0.010$ with p value $= 0.955 > 0.05$ not significant.

5.1.3 Discussion

The first part of the work presented in this thesis, described in chapter 4, analyses an experimental dataset developed by Barron and Erev (Barron and Erev [2003]; Erev and Barron [2005]) which replicated with some modifications the three experimental conditions previously adopted in Thaler et al. [1997]. This scenario represents an ideal starting point for this investigation, as it provides a controlled environment in which subjects were asked to make decisions and were allowed to observe the corresponding outcomes. These repeated, direct-experience partial-feedback interactions have been modelled with the reinforcement learning models proposed, following indications from the original papers. The attempt to connect this computational modelling framework with a theory of decision-making developed by Kahneman and Tversky [1979], which describes decision-making in risky situations, is part of the novelty of this work since no previous attempt to do this is known.

Prospect theory indicates that decision-makers shift their reference point according to previous experiences; this suggestion has been implemented through a two-state environment within the reinforcement learning framework. This point of view is also grounded in two behavioural phenomena regarding risk preferences, the "house-money" and "break-even" effects, according to which decision-makers change their attitude towards risk following previous gains or losses. The results of this investigation provided evidence that these subjects were best represented by models encompassing two-state environments based on previous outcomes, indicating a potential link between the RL modelling framework and prospect theory. Specifically, the comparison of the state-spaces proposed indicated that a two-states full-history model is more representative than either a two-states previous-outcomes model or a single-state trivial scenario. This indication offers an interesting starting point for future research efforts. When modelling subjects in this type of task, it is important to take

into account the way information is provided to the decision-makers, as this affects their perception and behaviour.

Using the same dataset and descriptive modelling approach, a set of three reward functions has been tested; the structures of these functions includes desirable features of decision-making identified in literature, such as decreasing sensitivity to extreme outcomes or loss-aversion. The results provide evidence towards prospect theory's subjective value function, a non-saturating decreasing function which also captures loss-aversion. Even if the results point in this direction, more research could improve descriptive studies of behaviour. In fact, the PT value function adopted in the present work has been parametrised according to the estimates from literature, in Tversky and Kahneman [1992]. These parameters have been estimated on decisions from description tasks and can potentially be improved by designing a similar experiment to identify more appropriate values for experience-based choice behaviour.

Three more research questions focused on the same data, with all three analysing the correlation between behaviour as described by the models' parameters and either the payoff variability experienced or the performance achieved. The payoff variability (PV) is a measure of the spread of outcomes experienced during the interactions. According to indications from the work in which the dataset was generated (Erev and Barron [2005]), a higher variability in observed outcomes leads to an impairment in learning associated with a more random behaviour. To better address this, the PV is compared to the values of parameter estimates, which describe speed of learning and randomness in action-selection for each subject.

In the first comparison, between PV and learning speed, the results indicate no correlation between these two quantities when considering the three experimental conditions separately. When pooling the data from these three conditions together, a significant moderate correlation is found. This could be the case because the Spearman correlation test adopted is based on the ranks of the datapoints; when pooling the data the arrangement of these datapoints follows a more monotonic distribution, leading to a significant result according to this test. This result indicates that those subjects who experienced more variability in the outcomes also behaved as if learning very fast. In Erev and Barron [2005] it was suggested the opposite, that when the payoff variability experienced by a subject increased, the learning was impaired. As the results obtained in the present analyses are at odds with the suggestions from previous work, more research is needed in this direction. A study encompassing two groups characterised by two distinct degrees of payoff variability could help shed light upon this open question.

The second comparison analyses the correlation between payoff variability and the greediness exhibited by the subjects. The parameter estimate for this behavioural feature indicates the strategy followed in choosing the actions; extreme values indicate random

choice in one end of the spectrum and greedy choice in the opposite end. The results from this comparison indicate that there is no correlation between the outcome variability observed by a subject and the action-selection policy followed, except for the third condition. In this last experimental setup, the underlying distributions for the two choices are very distinct. A small negative correlation has been found in this condition, indicating that subjects who experience higher variability tend to behave more randomly. The same tendency is observed when pooling the data from the three conditions together. These results provide some evidence confirming the indications from Erev and Barron [2005], that when outcomes are more spread people tend to become less efficient and act more erratically.

The last research question addressed in the first part of the work presented looks at the relationship between far-sightedness in the strategy adopted by the subjects and their performance in the task. Both an internal and an external measure of performance have been used to test the hypothesis that far-sighted subjects achieve better performance in this ambiguous binary decision task. The internal measure is based on the observed outcomes while the external is the amount of maximisation choices as a proportion of the total number of choices made. This idea is linked to the original paper (Thaler et al. [1997]) in which longer temporal arcs in feedback evaluation, from the subject's perspective, induced a higher proportion of maximisation choices. This concept was used to explain investors' choice as being biased towards short-sightedness (i.e. myopia). As the data analysed was generated with immediate feedback and had no feedback delays, a proxy for this was identified in a parameter estimate of the temporal-difference models proposed. The results indicate that there is no correlation between the degree to which subjects cared about future rewards and either measure of performance. These results suggest that myopic or far-sighted strategies do not have an influence on task performance. This could be the case because the time horizon in this task is too short, in fact the feedback is obtained at each interaction. Even if subjects considered future rewards within the modelling proposed, these do not play a role important enough in the experimental design of this specific task. A future investigation could be based on a task with a structure similar to the one considered in Thaler et al. [1997], with groups of subjects being exposed to different time-horizons. The present work provides an indication that for a task in which the outcome of the decisions is immediate, it is not far-sightedness that allows a decision-maker to achieve better returns.

An interesting extension to the modelling proposed for this task would be to integrate prospect theory's probability weighting function (Kahneman and Tversky [1979]; Tversky and Kahneman [1992]). This function is characterised by overweighting of small probabilities as well as underweighting larger ones. Combining these ideas with the action-selection policy

could help improve the descriptiveness of the modelling and more accurately describe the strategies adopted by the decision-makers.

Recent studies used fMRI neuro-imaging and neurophysiology to study the location of the signals captured by the quantitative models within the brain and how these interact with each other (O'Doherty et al. [2007]). Following the same procedures it would be interesting to pinpoint the differences between the subjects who experience different degrees of payoff variability and how their brains responds.

5.2 Quasi-field stock trading study results

Modelling stock-market investors as Reinforcement Learning agents

This section of the chapter is focused on the work published as first author and presented at the Institute of Electrical and Electronics Engineers, Evolving and Adaptive Intelligent Systems 2015 Conference (IEEE-EAIS 2015, Pastore et al. [2015]). One of the key motivations for this work is to test whether the descriptive modelling, found to be satisfactorily representative of choice behaviour in a laboratory study, can also capture behaviour in a quasi-field scenario. The term quasi-field indicates that, even if the subjects interact with an unconstrained scenario, these interactions are not happening in the real-world, where the outcomes of a subject's decisions involve gaining or losing the subject's personal money. The subjects analysed in this part of the work engaged in an online game consisting of a stock market trading simulation. Therefore, many of the constraints of a controlled experiment are not guaranteed to apply to this scenario. In this sense this task is closer to real-world interactions when compared to the previous binary decision task. The number of trials the subjects performed in this scenario is not fixed but depends on the subjects level of engagement. Moreover, the underlying distribution of decisions outcomes is not controlled by an experiment designer but derives from the real-world price fluctuations. Another reason to study subjects in this task is to try to understand people's attitude towards risk in a real-life decisions from experience scenario. The scope of this study can be collocated between computational psychology research and behavioural economics.

The next section presents the online game structure and the motivations for choosing it as a dataset. Subsequently, the hypotheses to be tested will be provided together with the methodology adopted for the tests. The models developed to test the assumptions in this work will be presented in the modelling section. The results of this part of the work will be presented in the next chapter, together with the results of the previous dataset analysis.

5.2.1 Transactions data

The dataset used for this analysis is composed of 46 players who produced a total of 1420 transactions. The data for these players was collected for the time-range between the 1st of January 2014 and the 31st of May 2014. Each player had the freedom to perform as many transactions as they wanted during these months, therefore some subjects have performed more transactions than others. Fig. 5.11 shows the timeline for players 1, 17, 37 and 41, as an example of the difference in behaviour and experience across participants. The graphs are scatterplots of the transactions, with the circles colour-coded according to the outcome of each transaction, green for gains and red for losses. The graphs can be read from left to right, on the x-axis are the transactions as they happened over time. The y-axis represents three discrete levels of risk associated with each transaction, deriving from the risk-based categorisation of the available stocks.

The choice made in this analysis, of describing players individually with computational models, follows the same rationale as the similar decision made in the previous chapter. The best fitting model is identified for each participant independently, with the aim of finding the most descriptive learning procedure and decision-making behaviour, for that particular participant. If an aggregate model is developed instead, by for example collating the data from all players' transactions, the effects of learning from different experiences and choices would average out, leading to a loss of individuality in the results. A similar approach is also proposed in Daw [2009], with the difference that the methodology described in this work aims at finding population parameters, by first estimating the most likely parameters for each subject and then considering some distribution (e.g. Gaussian) over these.

Hypothesis 1

The first hypothesis tests the assumption that Virtual Trader players grouped the 107 FTSE100 stocks, available to trade during the months considered, in discrete classes according to a measure of risk associated with each stock. To test this hypothesis, a classification based on a financial modelling measure of risk called beta coefficient β_F, or financial elasticity, was developed and compared to other 500 randomly generated arrangements. The proposed RL models were fitted to the choice data for each player and for each of these arrangements. This first investigation is not focused on testing whether RL is a good descriptive account of the decision-making players exhibited, which will be tested in the following hypotheses. In this first examination, models with the same complexity (same number of parameters) are compared against each other, with the intent of finding out whether the proposed risk-based classification is a good descriptive choice. This hypothesis is tested by comparing

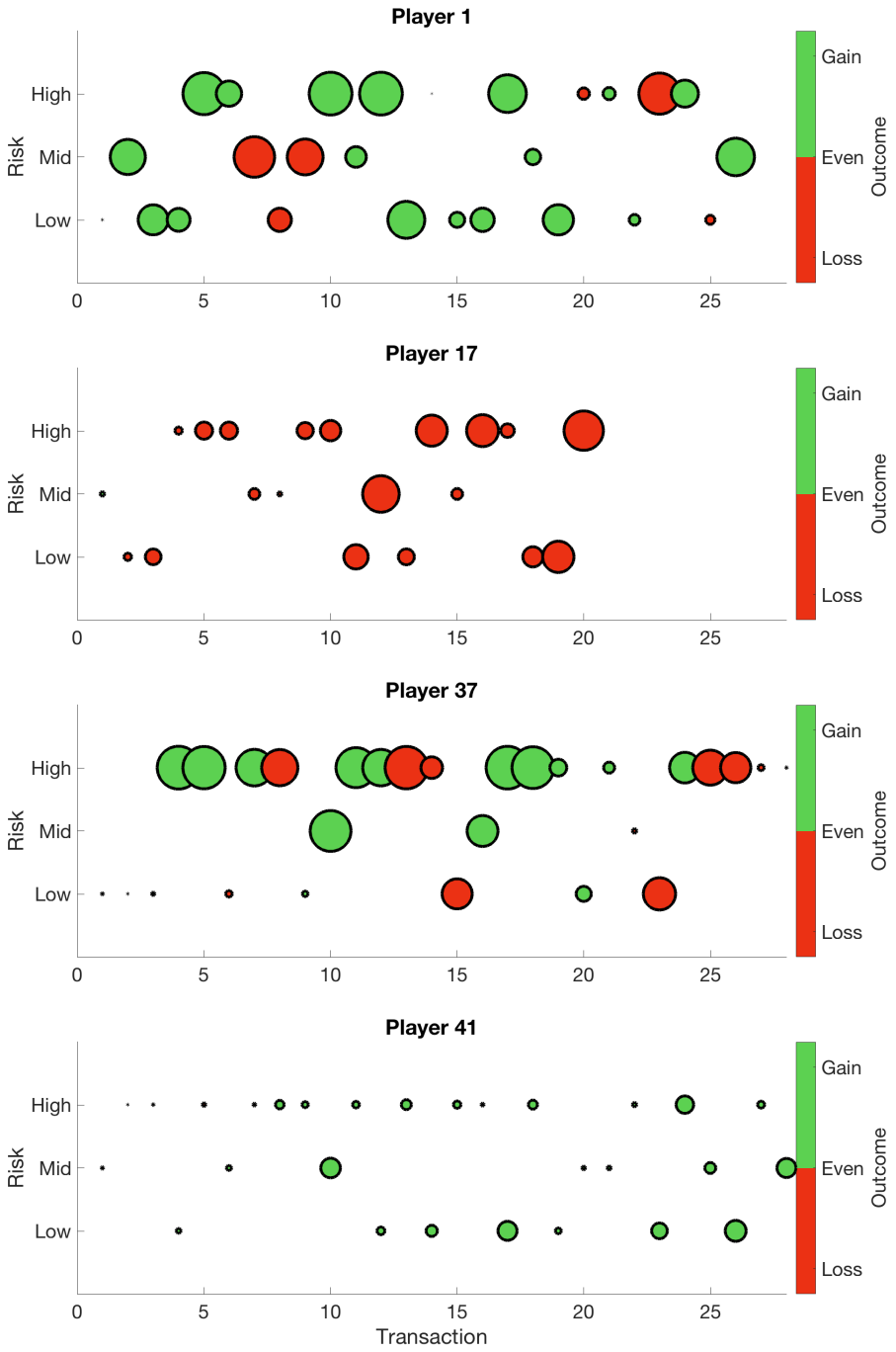

Fig. 5.11 Transaction timelines for players 1, 17, 37 and 41. From left to right, each circle represent the outcome of a transaction; the colour of a bubble indicates the outcome (green for gains and red for losses), the bubble's size shows the magnitude of the outcome. The y-axis represents the risk of the stock traded.

the AIC score of the model using the risk-based classification against the AIC score for each of the 500 randomly generated arrangements. Each of these comparisons results in

a binary value, which indicates whether the risk-based arrangement is better than the i-th scrambled classification. This method yields a binary vector of 500 comparison results, for each player. The Clopper-Pearson confidence interval method is subsequently applied to obtain a probability and confidence interval. The results of this test are shown in Fig. 5.12, where the x-axis presents the players arranged according to their ID; the y-axis displays the probability, ranging between 0 and 1, with the chance threshold depicted by the red line at the 0.5 level. The probability and relative 99% confidence interval, for a player to be likely described by the proposed classification is indicated by the errorbar; if a player's errorbar lies entirely above the chance threshold, that player is significantly likely to be adopting the risk-based classification proposed. This is the case for 31 players out of 46 (67.4%) indicating that these subjects are likely to use this risk-based classification of available stocks.

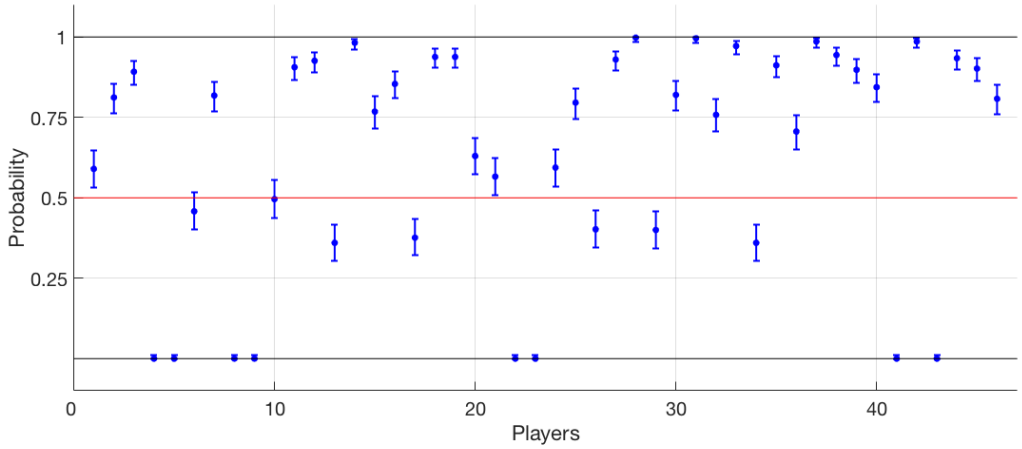

Fig. 5.12 Clopper-Pearson binomial confidence intervals for the comparison between risk-arranged against scrambled stocks in 3 categories, for all players in the dataset. The players are portrayed on the x-axis, the y-axis shows the probability that the risk-based stock arrangement is more descriptive than 500 alternatives. The errorbar represents the 99% confidence interval for each player, if the errorbar lies completely above chance threshold it can be concluded that the player is likely to perceive the tradable stocks as classified according to the proposed risk-based categorisation. 31 players out of 46 (67.4%) show confidence intervals which are significantly above chance.

A Pearson Chi-squared "goodness-of-fit" test provides the answer for this first hypothesis. The test resulted in $\chi^2(2, N = 46) = 5.565, p = 0.018 < 0.05$, shown in Fig. 5.13.

The results of these tests produce enough evidence in favour of the first hypothesis, that subjects classify the available tradable options according to a financial modelling measure of risk, which in this case was identified in the beta coefficient β_F, from the Capital Asset Pricing Model (CAPM; Sharpe [1964]).

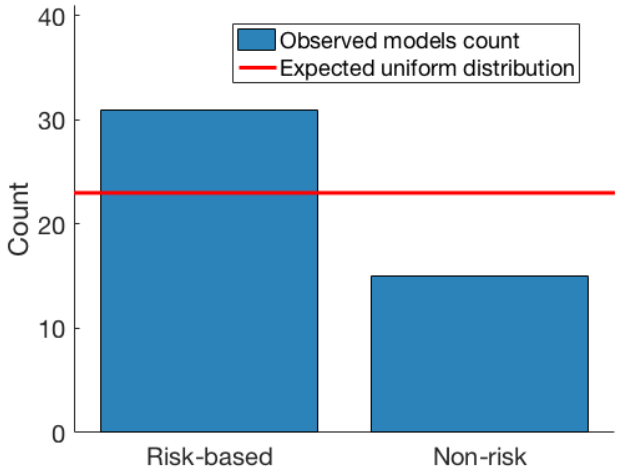

Fig. 5.13 The comparison of the risk-based stock classification against the randomly generated arrangements. The left bar represent the number of players whose confidence interval indicates they are likely to group stocks based on a measure of their risk. The right bar represents the remaining players, who cannot be assumed to adopt the risk-based classification. The red line is the expected count for a uniform distribution.

Hypothesis 1: Virtual Trader players use the risk of the stocks, estimated as beta coefficient to categorise them into discrete classes.

$\chi^2(2, N = 46) = 5.565, p = 0.018 < 0.05$ significant: hypothesis confirmed.

Hypothesis 2

The second hypothesis of this chapter is concerned with understanding whether the subjects who played the Virtual Trader online financial trading simulation behaved in a way which can be described as reinforcement learning. The reason for testing this assumption on this dataset is that the Virtual Trader online game can be seen as a repeated decision-making task, with a reward/punishment signal associated with each action taken. During the months when the data has been collected, the Virtual Trader game awarded a prize to the best achievers, converting the game objectives and rewards into concrete ones. The winners were established as the players whose cumulative assets achieved the highest value at the end of each month. Players were endowed with a virtual lump sum of GBP 100,000 at the moment of subscription and were free to trade the available FTSE100 stocks. The participants were ranked on their cumulative assets, the combination of portfolio holdings and cash owned by a player. The

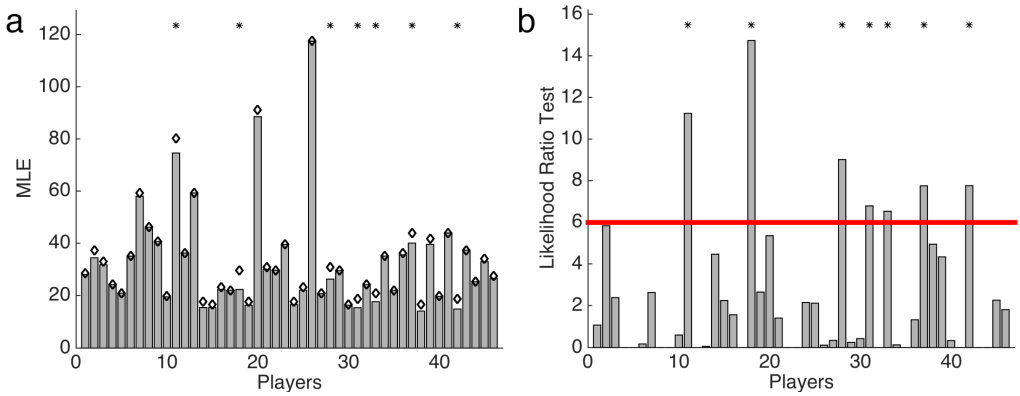

Fig. 5.14 Immediate rewards RL model versus Random model. Panel a) shows the maximum likelihood estimate of the myopic model (2 free parameters, learning rate α and soft-max inverse temperature β), compared to the random model. The players IDs are reported on the x-axis, with each bar representing one player; the bars indicate the MLE associated with each player; the hollow diamonds indicate the random model MLE for the same player. The y-axis shows the MLE scores. Since players made different numbers of transactions, the height of the bars varies greatly within the dataset. The asterisks identify those players who are better fitted by the myopic RL model as opposed to the random model. Panel b) shows the same comparison, with the difference that the results of the likelihood ratio test are shown on the y-axis. The horizontal line is the critical value threshold, above which the more complex model (in this case the myopic RL) is significantly better than the simpler one (in this comparison the random model). The value of this threshold depends on the degrees of freedom, which for the comparison portrayed is $d.o.f. = 2$ and LRT critical value is 5.991.

modelling developed to test this hypothesis, that players in this game behave in a myopic reinforcement learning way, used the cash component of the game to model the reward signal and to develop the state-space. More precisely, the monetary outcome of the "sell" transactions was assumed to be the reward signal for the agents modelled, and the sum of these signals was assumed to determine the current state of the world the agent is considered to be in. In order to provide an answer to the hypothesis, a simple RL model targeting immediate rewards was fit to each player's transactions set (i.e. a subject's choice dataset).

A series of gradient descent searches was performed to identify the parameters set with the highest likelihood. This procedure examines many different combination of values for the free-parameters, with the objective of minimising the output, the maximum likelihood estimate. The MLE score of the RL model is compared to the one of the random model with the likelihood ratio test. The MLE of a random model depends on the number of transactions made by the player and is calculated analytically as shown in 4.29.

The results of this comparison are presented in figure 5.14. Panel a) shows the players along the x-axis, and their associated MLE values on the y-axis. The values of MLE are

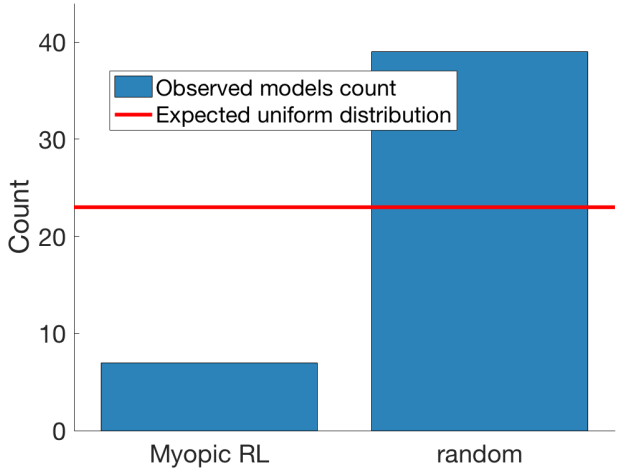

Fig. 5.15 The comparison of the immediate reward reinforcement learning against the random model. The bars depict the frequency of best fitting models for the 46 players; the left bar represents the frequency of myopic (immediate-rewards) reinforcement learning models, the right bar indicates the number of players best fitted by the random model. The red line is the expected count for a uniform distribution.

presented in the form of bars for the RL immediate-reward model and hollow diamonds for the random model. The height of bars and diamonds represents the goodness-of-fit of a model, with lower values indicating better fitting models. The asterisks identify the subjects whose RL model fits significantly better than the random model, according to the Likelihood ratio χ^2 test at the 95% confidence level (p value < 0.05). This is the case for 7 out of 46 players in the dataset, 15% of the total: players 11, 18, 28, 31, 33, 37 and 42. The comparison between the same two models is shown in panel b), with the y-axis now representing the likelihood ratio test values. The red horizontal line represents the threshold above which statistical significance is achieved; this critical value depends on the degrees of freedom. For this comparison these are $d.o.f. = 2$, leading to a critical value of 5.991.

The Pearson Chi-squared "goodness-of-fit" χ^2 test carried out on the frequencies of best fitting models is shown in Fig. 5.15 and resulted in $\chi^2(2, N = 46) = 22.261, p = 0.000 < 0.05$. These results indicate that the myopic RL model tested for this hypothesis is not a good descriptive account of behaviour for the Virtual Trader players analysed. The second hypothesis is therefore rejected.

One possible explanation for this model failing at describing the players is that, even if Virtual Trader is a simulation game and the players are unskilled investors, it is possible that some of them took into account information about the companies of the stocks traded before making decisions. This type of trading is based on "fundamental analysis", a method for the

estimation of securities value and price. An investor would read the financial statements of a company, investigate rumours and news from different sources and study the historical data before operating a trade. Another potential explanation for these negative results, is that the model proposed is based on a myopic goal. A more advanced model, which accounts for a player's interest for future rewards could better describe the choice data; this possibility is tested in the third hypothesis.

Hypothesis 2: Virtual Trader players behave following a myopic reinforcement learning pattern.

$\chi^2(2, N = 46) = 22.261, p = 0.000 < 0.05$ significant, against the RL model: hypothesis rejected.

Hypothesis 3

Literature suggests that unskilled investors' behave in a naive way, by making decisions influenced by their previous personal experience (Choi et al. [2009]; Huang [2012]). The third hypothesis tests whether the Virtual Trader players behaved naively when playing the game. To test this hypothesis the short-sighted RL model will be compared to a far-sighted counterpart, temporal difference learning. Moreover, the far-sighted model will also be compared to the random model, in order to assess whether the eventual improvement on the myopic version translates to a higher, population-wide descriptiveness. The ability of the Q-learning model to include future rewards in the learning procedure makes it a good candidate to capture far-sighted strategies in the Virtual Trader game dataset. As detailed in section 4.2.8, eq. 4.2.8, the Q-learning model offers a more complex learning rule than the myopic RL used to test the first and second hypothesis, and could provide a better description of the Virtual Trader players' behaviour. The comparisons between Q-learning against immediate-rewards RL and Q-learning against random model are carried out based on the likelihood ratio test methodology adopted for the previous hypothesis. The results of these tests are shown in Fig. 5.16.

Panels a) and c) of the figure show the comparison between the far-sighted model Q-learning, with the nested, myopic RL model. This first comparison, between far-sighted and myopic models, provides a statistically significant improvement in the fitting for 6 players (12, 14, 30, 39, 40 and 44), but this marginal improvement is not reflected in a better overall fitting performance for the entire players set. As shown in panels b) and d) in fact, the comparison between far-sighted model and baseline random model provides comparable

results to the ones obtained in the second hypothesis test. The TD-learning model fits 7 out of 46 players better than the random (11, 14, 18, 28, 31, 39, 42), representing 15% of the total dataset. These results are obtained with the likelihood ratio χ^2 test, at the 95% confidence level (p value < 0.05). A subset of these players was already captured by the myopic model (11, 18, 28, 31 and 42), with only 2 players not previously captured (14 and

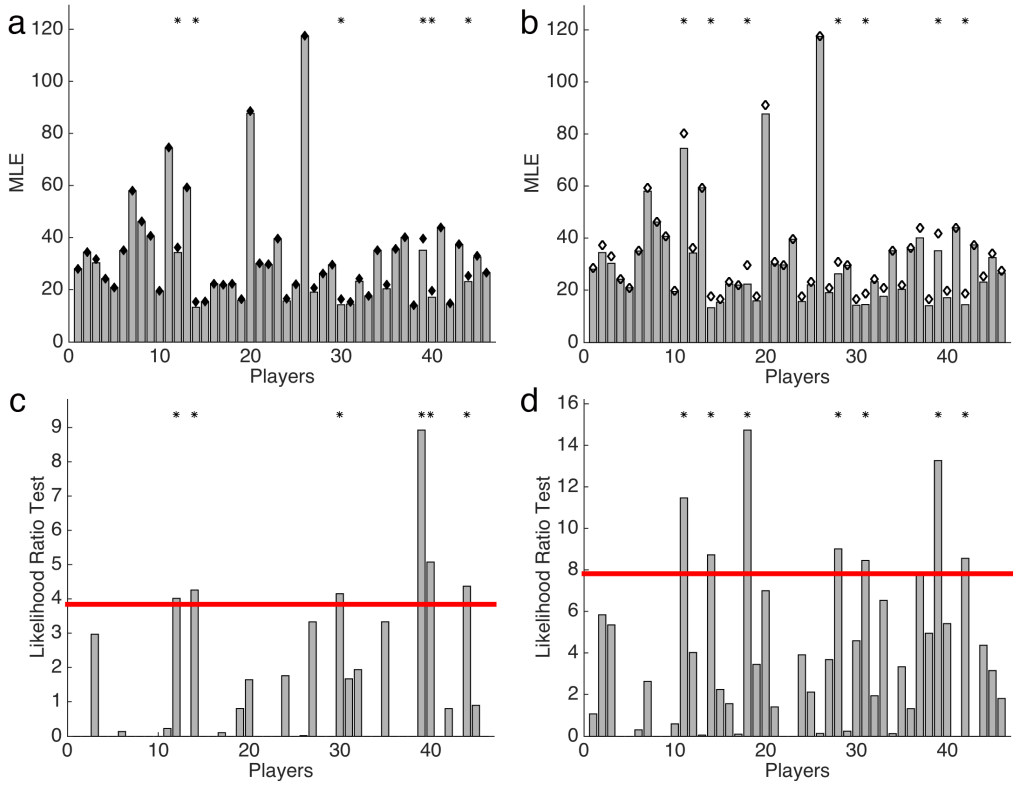

Fig. 5.16 Temporal-difference model (Q-learning) versus immediate rewards RL model and versus random model. Panels a) and b) show the comparison between Q-learning model (3 free parameters, capturing learning speed α, action-selection greediness β and far-sightedness γ) and the nested immediate-rewards RL model (2 free parameters, discount-factor parameter set to fully myopic $\gamma = 0$, nested version of Q-learning). Panel a) portrays players on the x-axis and MLE scores on the y-axis, each players' score is represented as a bar for the Q-learning model and as a filled diamond for the myopic RL model (immediate-rewards). The asterisks identify those players who are better fitted by a model capturing far-sightedness as opposed to a myopic model. Panel c) shows the same comparison but with the outcome of the likelihood ratio test on the y-axis, the LRT threshold for statistical significance is 3.841 ($d.o.f. = 1$). Panels b) and d) show the comparison of the far-sighted model with the baseline random model. Panel b) represents the MLE comparison, panel d) shows the LRT comparison with the threshold for significance being higher than all previous comparison, critical value 7.815 ($d.o.f. = 3$). As for the previous figures, the asterisks indicate the players for which the proposed model is more representative than the random baseline.

39). At the same time, due to increased complexity and its relative penalisation by the LRT, the far-sighted model fails to capture the behaviour for players 33 and 37, which were instead well represented by the immediate-rewards RL model.

According to these results, the third hypothesis is also rejected. The number of players well represented by a far-sighted model in fact, is not significantly higher than the baseline random model, or the myopic model. Conclusions on the entire population cannot be drawn, considering that the number of players fitted by either myopic or far-sighted reinforcement learning model is too small. Figure 5.17 panel a) shows the results of the three-way Pearson "goodness-of-fit" χ^2 test carried out between far-sighted model, myopic model and random baseline. The test resulted in $\chi^2(3, N = 46) = 46.7391, p = 0.000 < 0.05$. A similar comparison, which focused on the subset of players well described by the RL models proposed, is shown in Fig. 5.17, panel b). The Pearson "goodness-of-fit" χ^2 test for this comparison provides negative results with $\chi^2(2, N = 9) = 2.7778, p = 0.095 > 0.05$, not significant. This last result indicates that there is no evidence that players, if behaving in a reinforcement learning fashion, do so naively.

Only 9 players out of 46 are satisfactorily captured by either of the two RL models, with the myopic version capturing 7 and the far-sighted model merely capturing 2. Fig. 5.18 shows that for this subset of players, the assumption of risk-based stock classification is confirmed. This could indicate that the players who have not been captured by the RL

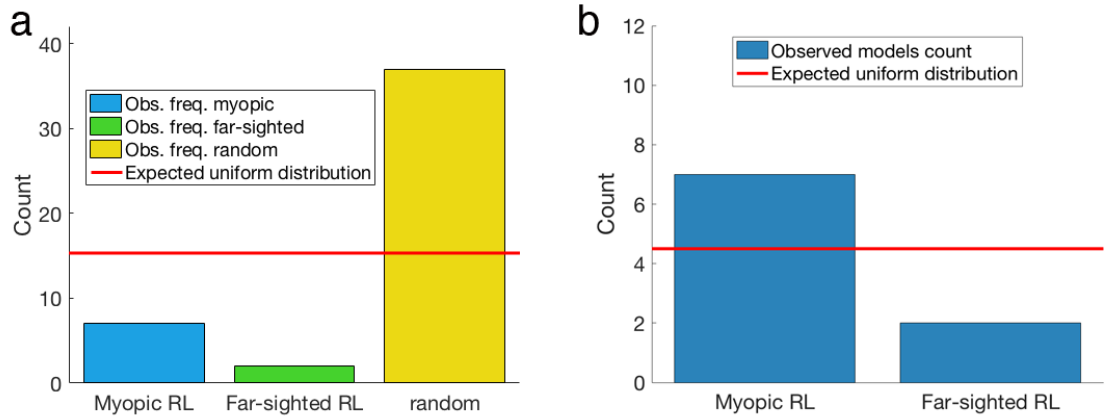

Fig. 5.17 Panel a) shows the comparison between the Q-learning far-sighted model, the immediate reward reinforcement learning and the random model. The bars depict the frequency of best fitting models for the 46 players; the bar represents the frequency of players best fitted by each model and are colour coded accordingly (blue for myopic RL, green for far-sighted RL and yellow for random model). The red line is the expected count for a uniform distribution. Panel b) shows the comparison between the two RL models tested, for the subset of players they captured.

models proposed could still act according to a reinforcement learning pattern, but that the action-space devised in this work is inappropriate. On the other hand, this result could also indicate that these players were fitted satisfactorily only by chance.

The results obtained from the hypotheses tested in this chapter indicate that players in the Virtual Trader dataset cannot be represented as reinforcement learning, neither as short- nor far-sighted agents. These results do not exclude that RL can still be a feasible way to represent these players, only that the particular modelling developed is not suitable for this task. Taking into account only the sales outcomes as reward signal is a potential explanation of this limitation. A potential way to increase the descriptiveness of the model developed, in order to capture Virtual Trader players behaviour, is to extend the definition of reward signal, for example, by estimating the value of the portfolio at each transaction, including the purchasing ones. These results also indicate that more research is needed to understand unskilled investors' behaviour, even when analysing a task that does not involve real-world monetary implications to the subjects.

This work aimed at investigating the indications originating from economics literature (Choi et al. [2009]; Huang [2012]) and from psychology (Erev and Roth [1998]). The results provided no evidence that the unskilled investors analysed consistently behave in a naive reinforcement learning way. As previously indicated, this could be due to limitations of

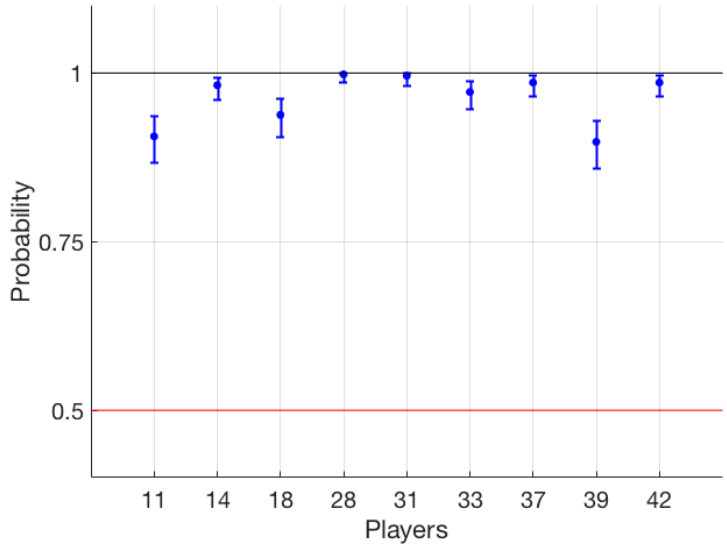

Fig. 5.18 Subset of Clopper-Pearson binomial confidence intervals for the comparison between risk-arranged against scrambled stocks in 3 categories, for the players best fit by a RL model (either myopic or far-sighted). All the 9 players for which either the myopic or the far-sighted RL model outperformed the random model present a statistically significant probability of using the stock classification into a discrete risk-based action-space.

the modelling developed or it could derive from the high complexity of the task analysed. Replicating this type of investigation within a controlled experiment setup might help clarify whether unskilled stock traders can be described as reinforcement learning agents. Future developments for this work and other interesting questions regarding reinforcement learning behaviour, in connection to decision-making in uncertain environments, will be discussed in the following chapter.

Hypothesis 3: Virtual Trader players behave following a far-sighted reinforcement learning pattern.

$\chi^2(3, N = 46) = 46.7391, p = 0.000 < 0.05$ significant, against both the RL models proposed: hypothesis rejected.

$\chi^2(2, N = 9) = 2.7778, p = 0.095 > 0.05$, not significant, no evidence in favour of either myopic or far-sighted.

5.2.2 Discussion

The second part of the work presented in this thesis focused on a dataset deriving from a quasi-field task. This dataset comprised transactions and outcomes deriving from the interactions of a group of an online stock trading game players. The main question was whether a descriptive modelling approach based on reinforcement learning models, similar to the one developed in the first part of this chapter, could also be useful in describing a more complex scenario. The first hypothesis of this investigation tested the assumption that the players of this online game grouped the available stocks as if belonging to discrete classes of risk; the result of this test indicate that this could indeed be the case. Although the results for the remaining hypotheses provide no support for the idea that reinforcement learning, either myopic or far-sighted, can be used to describe the learning and decision-making strategies underlying these players, it is necessary to take into account the fact that the model proposed is based on a reward signal derived from the sell transactions. This partial modelling could be inadequate because it does not consider the potential reward the players obtain by, for example, examining their portfolios at the time when purchases are made. In this regard, future development would be to include this reward signal. In a similar setup the prices of the stocks in a portfolio would be recorded, not only at the time of a sale, but also at the time of purchasing transactions. It is reasonable that players check the prices of the stocks they own at each available time. The results from this investigation do not provide enough

evidence to confute or support the indication that unskilled investors behave according to a naive reinforcement learning pattern, as suggested by Choi et al. [2009]. The aim of a future study would be to extend the models proposed here and simultaneously obtain a dataset including longer series of interactions, in comparison to the amount of available options.

Chapter 6

Discussion and Future Developments

Human learning and decision-making are extremely interesting and complex areas of research. They represent part of the overlapping interest of multiple research fields. Economics, psychology and neuroscience would all benefit greatly from an improved understanding of the processes involved in human decision-making behaviour. In this regard, computational models represent an interesting arrow to the bow of researchers. All these disciplines have produced attempts to formalise human behaviour into models. The development of computational models to characterise learning and decision-making has been of crucial importance. Neuroscientists, for example, use *in-silico*[1] models as substrates for hypothesis development and testing, before moving on to *in-vitro* or *in-vivo*. Computational models are of crucial importance for these researchers and have been used to study, for example, neurological diseases (Tomkins [2015]). Psychology, as well, would benefit from a better understanding of learning and decision-making, as this could help characterise the differences in healthy and detrimental behaviour in individuals (Bechara et al. [2000]). Economists have always tried to encapsulate decision-making into mathematical models, from both a normative and a descriptive point of view. Policy makers would benefit from the development of computational models of behaviour. In fact, by characterising human behaviour with mathematical formulations, researchers could develop complex simulations and computational models based on these insights (Rutkauskas and Ramanauskas [2009]) and use these as a platform where policy makers can conduct initial tests, before introducing new regulations. From these simple examples it is clear that more investigation on human behaviour, specifically on learning and decision-making, are of crucial importance for researchers from these disciplines. In this context, the work carried out in this thesis focused on a set of models characterising learning and decision-making, to better understand human behaviour in either an experi-

[1] in-silico modelling refers to the use of mathematical and computational models, in-vitro indicates the use of cell cultures, in-vivo means the study is performed on tissues of a living being

mental scenario or in a quasi-field study. The models tested, based on the reinforcement learning framework, previously received praise for their ability to capture human behaviour and represented a good starting point from where to begin the current investigation.

To collocate this work into the literature it is worth summarising the questions which prompted this research and the answers provided by the results of the analyses. The first of the results form the analysis of a two-choice decision-making task indicate that a state-space with two states based on the full history of previous outcomes accrued by a subject is a likely descriptive account of the subjects' perception. This result reinforces the notion of previous experiences playing a role in decision-making tasks, as indicated by the "house-money effect" and the "break-even effect". Such information can be used in designing the scenarios in which people operate choices. As an example, investors could be provided with selected information about their investments. This could help improve the final return on their investments. The second result points against a fully saturating reward transformation function. The weakness of this finding is in the fact that only a version of prospect theory value function has been tested, the one parametrised on previous decision from description tasks Therefore, more research is needed to discover which function could be playing a role in the internalisation of the rewards values in binary choice decision-making tasks. An interesting result, that provides evidence against previous work findings (Erev and Barron [2005]), in the sense that subjects who experienced more variable outcomes tended to learn faster. More research is needed in order to shed light on this specific dilemma. This finding is particularly interesting because it challenges the previous notion that more uncertainty impairs learning. In a real life scenario this might mean that people who are less certain about the potential outcome of their actions might quickly pick one of the options available and might eventually end up worse off. The next result, instead, provides additional evidence to the indication from Erev and Barron [2005], that more variability in observed payoffs leads to more randomness in choice behaviour. This finding, paired with the previous one, indicates that payoff variability plays a major role in decision-making tasks. Future experimental design should take this into account and carefully identify the appropriate level of uncertainty to avoid potentially confusing the subjects. Moreover, policy makers should take these last two findings into account in the regulation of financial or insurance advisory, as it is clear that decision-makers are heavily influenced by noisy past outcomes. The last test for the first dataset analysed provides no indication on the relationship between myopic behaviour and task performance. This might be due to the short time horizon of the task. Therefore, a more articulated controlled experiment is needed to adequately test for this potential relationship.

In the second dataset analysis it emerged that the players of the online stock trading simulation could have used risk as a measure to categorise the number of available options

into a more manageable number. This has implications in real life, evidently because it can be deduced that investors might be coarsely pooling together options with similar risk, by using only a technical calculation and neglecting a more fundamental hazard, such as the nature of a stock or the cyclical risk of the sector to which it belongs. The other indications from this analysis are inconclusive, as the number of subjects best approximated by the proposed RL models is too low. This provides two indications: the first is that modelling only one of the two components of the task, in this case the sales a subject makes, is not enough to fully capture the decision-making behaviour; the second indication is that such complex task might not be one in which subjects operate using a reinforcement learning approach. Future work should focus on testing more advanced RL models, including ones that leverage function approximation for the reward function, including multi-order polynomials (Ylöstalo [2006]) and Gaussian radial basis functions (Kober et al. [2013]; Park and Sandberg [1991]; Sutton and Barto [2018]), as well as devising a controlled experiment in which to test for their descriptive power. A stock market trading simulation might be too complex for this purpose and a binary decision task too trivial, therefore a controlled experiment that keeps into account the findings of this work could be designed to gain knowledge to help answer the currently unanswered questions. The suggested directions of future research would help shed light on choice behaviour, providing indications for policy makers, financial regulators and governments. The current global situation offers many instances where to apply the information gained from research in the fields of behavioural economics and neuroscience. Financial education is expensive and has been found to be ineffective Choi et al. [2002, 2005]; Cole and Shastry [2008]. Therefore, the key for better investor behaviour could be selecting and reducing the frequency of the information available (Thaler et al. [1997]), and limiting the reinforcements (Choi et al. [2009]).

People will always choose as they please. Nonetheless, institutional bodies and choice architects have the power to nudge them towards the right option. They can do so by harnessing the knowledge deriving from research efforts on decision-making. By doing so large negative events, such as the recent sub-prime mortgage crisis, could be avoided, ultimately leading to a more stable and prosperous economy.

Alvin Pastore
London, January 2019

References

Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*.

Ahmed, S. and Gutkin, B. (2011). *Computational Neuroscience of Drug Addiction*, volume 10. Springer Science & Business MEdia.

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldview of AIC and BIC. *Ecology*, 95(March):631–636.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alexander, W. H. and Brown, J. W. (2010). Hyperbolically discounted temporal difference learning. *Neural computation*, 22(6):1511–27.

Andreassen, P. B. and Kraus, S. J. (1990). Judgmental extrapolition and the salience of change. *Journal of Forecasting*, 9:347–372.

Arkes, H. and Ayton, P. (1999). The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin*, 125(5):591–600.

Arkes, H. R. and Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1):124–140.

Atkeson, C. G., Moore, A. W., and Schaal, S. (1997). Locally weighted learning for control. In *Lazy learning*, pages 75—-113.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Lent, R., Herculano-Houzel, S., and Others, A. (2009). Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-Up Primate Brain. *Journal of Comparative Neurology*, 513(5):532–541.

Azfar, O. (1999). Rationalizing hyperbolic discounting. *Journal of Economic Behavior & Organization*, 38(2):245–252.

Bagnell, J. A. and Schneider, J. G. (2001). Autonomous helicopter control using reinforcement learning policy search methods. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, volume 2, pages 1615—-1620.

Bagnell, J Andrew and Kakade, Sham M and Schneider, Jeff G and Ng, A. Y. (2004). Policy search by dynamic programming. In *Advances in neural information processing systems*, pages 831—-838.

Barber, B. M., Lee, Y. T., Liu, Y. J., and Odean, T. (2007). Is the aggregate investor reluctant to realise losses? Evidence from Taiwan. *European Financial Management*, 13(3):423–447.

Barber, B. M., Lee, Y.-T., Liu, Y.-J., and Odean, T. (2014). Do Day Traders Rationally Learn About Their Ability? *Working Paper*.

Barber, B. M. and Odean, T. (2013). *The Behavior of Individual Investors*, volume 2. Elsevier B.V.

Barber, B. M., Odean, T., and Zhu, N. (2009). Systematic noise. *Journal of Financial Markets*, 12(4):547–569.

Barraclough, D. J., Conroy, M. L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature neuroscience*, 7(4):404–410.

Barron, G. and Erev, I. (2003). Small Feedback-based Decisions and Their Limited Correspondence to Description-based Decisions. *Journal of Behavioral Decision Making*, 16(3):215–233.

Bechara, A., Tranel, D., and Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, 123(11):2189–2202.

Becker, S. W. and Brownson, F. O. (1964). What Price Ambiguity? or the Role of Ambiguity in Decision-Making. *Journal of Political Economy*, 72(1):62–73.

Benartzi, S. and Thaler, R. (1995). Myopic loss aversion and the equity premium puzzle. *The quarterly journal of Economics*, 110.1(1):73–92.

Benartzi, S. and Thaler, R. H. (2007). Heuristics and Biases in Retirement Savings Behavior. *Journal of Economic Perspectives*, 21(3):81–104.

Benhabib, J. and Bisin, A. (2005). Modeling internal commitment mechanisms and self-control : A neuroeconomics approach to consumption – saving decisions. *Games and Economic Behavior*, 52:460–492.

Beninga, S. (2000). *Financial Modeling*.

Bernheim, B. D. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *The American Economic Review*, 94(5):1558–1590.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36.

Bertoluzzo, F. and Corazza, M. (2007). Making Financial Trading by Recurrent Reinforcement Learning. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 619–626.

Bertoluzzo, F. and Corazza, M. (2012). Testing Different Reinforcement Learning Configurations for Financial Trading: Introduction and Applications. *Procedia Economics and Finance*, 3(0):68–77.

Bickel, W. and Marsch, L. (2001). Toward a Behavioral Economic Understanding of Drug Dependence: Delay Counting Process. *Addiction*, 96(February 2000):73–86.

Bitzer, S., Howard, M., and Vijayakumar, S. (2010). Using dimensionality reduction to exploit constraints in reinforcement learning. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 3219—-3225.

Black, F., Jensen, M. C., and Scholes, M. (1972). *The Capital Asset Pricing Model: Some Empirical Tests*, volume 81.

Bogacz, R. and Gurney, K. (2007). The Basal Ganglia and Cortex Implement Optimal Decision Making Between Alternative Actions. *Neural Computation*, 19(2):442–477.

Bogacz, R. and Larsen, T. (2011). Integration of Reinforcement Learning and Optimal Decision-Making Theories of the Basal Ganglia. *Neural Computation*, 23(4):817–851.

Braziunas, D. (2003). POMDP solution methods.

Britten, K., Newsome, W., Shadlen, M. N., Celebrini, S., and Movshon, J. (1996). A relationship between behavoiral choice and the visual responses of neurons in macaque MT. *VIsual Neuroscience*, 13(1):87–100.

Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. a. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 12(12):4745–4765.

Brocas, B. I. and Carrillo, J. D. (2008). The Brain as a Hierarchical Organization. *American Economic Review*, 98(4):1312–1346.

Brown, P., Chappel, N., Da Silva Rosa, R., and Walter, T. (2006). The Reach of the Disposition Effect: Large Sample Evidence Across Investor Classes. *International Review of Finance*, 6(1-2):43.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed)*, volume 172. Springer, New York.

Burnham, K. P. and Anderson, R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304.

Busemeyer, J. R. and Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3):432–459.

Camerer, C. (1999). Behavioral economics: Reunifying psychology and economics. *Proceedings of the National Academy of Sciences*, 96(19):10575–10577.

Camerer, C. and Ho, T.-H. (1999). Experience-weighted attraction in games. *Econometrica*, 67(4):827–874.

Cameron, I. G. M., Watanabe, M., Pari, G., and Munoz, D. P. (2010). Executive impairment in Parkinson's disease: Response automaticity and task switching. *Neuropsychologia*, 48(7):1948–1957.

Campbell, J. Y. (2003). Consumption-Based Asset Pricing. In *Handbook of the Economics of Finance*, volume 1, pages 803–887. Elsevier.

Cappè, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516—-1541.

Cassandra, A. R. (1998). A survey of POMDP applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*, page 1724.

Chen, Y., Mabu, S., Hirasawa, K., and Hu, J. (2007). Trading rules on stock markets using genetic network programming with sarsa learning. *Proceedings of the 9th annual conference on Genetic and evolutionary computation GECCO 07*, 12:1503.

Chevalier, J. and Ellison, G. (1997). Risk Taking by Mutual Funds as a Response to Incentives. *The Journal of Political Economy*, 105(6):1167–1200.

Choi, J., Laibson, D., Madrian, B., and Metrick, A. (2002). Defined contribution pensions: Plan rules, participant choices, and the path of least resistance. *Tax Policy and the Economy*, 16(January):67–114.

Choi, J. and Mertens, T. (2006). Extrapolative Expectations and the Equity Premium. *Yale University*.

Choi, J. J., Laibson, D., and Madrian, B. C. (2005). $100 bills on the sidewalk: Suboptimal investment in 401 (k) plans. *Review of Economics and Statistics*, 93(3):748—-763.

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2009). Reinforcement learning and savings behavior. *Journal of Finance*, 64(6):2515–2534.

Chopra, N., Lakonishok, J., and Ritter, J. R. (1992). Measuring abnormal performance: do stocks overreact? *Journal of Financial Economics*, 31(2):235–168.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

Coates, A., Abbeel, P., and Ng, A. Y. (2009). Apprenticeship learning for helicopter control. *Communications of the ACM*, 52(7):97—-105.

Cohen, J. D., McClure, S. M., and Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1481):933–942.

Cole, S. and Shastry, G. K. (2008). *If You are So Smart, why Aren't You Rich?: The Effects of Education, Financial Literacy and Cognitive Ability on Financial Market Participation*.

Conn, K. and Peters, R. A. (2007). Reinforcement learning with a supervisor for a mobile robot in a real-world environment. In *Computational Intelligence in Robotics and Automation, 2007. CIRA 2007. International Symposium on*.

Curley, S. P. and Yates, J. F. (1985). The center and range of the probability interval as factors affecting ambiguity preferences. *Organizational Behavior and Human Decision Processes*, 36(2):273–287.

Curley, S. P., Yates, J. F., and Abrams, R. A. (1986). Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes*, 38(2):230–256.

Darwin, C. and De Beer, S. G. (1956). The origin of species by means of natural selection: or, the preservation of favoured races in the struggle for life. Technical report, Oxford University Press.

David, F. N. (1962). *Games, gods and gambling : the origins and history of probability and statistical ideas from the earliest times to the Newtonian era*. Hafner Publishing Company, London, England.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications*. PhD thesis, Carnegie Mellon University.

Daw, N. D. (2009). Trial-by-trial data analysis using computational models. In *Decision Making, Affect, and Learning: Attention and Performance XXIII*, pages 3–38. Oxford University Press Oxford.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–11.

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

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, affective & behavioral neuroscience*, 8(4):429–53.

Dayan, P. and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2):185–96.

Dayan, P., Niv, Y., Seymour, B., and Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural networks : the official journal of the International Neural Network Society*, 19(8):1153–60.

Dayan, Peter and Hinton, G. E. (1997). Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278.

De, S., Gondhi, N., and Pochiraju, B. (2010). Does Sign Matter More than Size? An Investigation into the Source of Investor Overconfidence.

De Bondt, W. F. D. and Thaler, R. H. (1990). Do Security Analysts Overreact ? *The American Economic Review*, 80(2):52–57.

De Bondt, W. F. M. (1991). What Do Economists Know about the Stock Market?

De Bondt, W. F. M. and Thaler, R. H. (1985). Does the stock market overreact. *Journal of Finance*, 40(3):793–805.

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D., and Fiez, J. A. (2000). Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum. *Journal of Neurophysiology*, 84(6):3072–3077.

Dickinson, A. and Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1):1–18.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10:732–739.

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP Journal*, 1(1):30.

Drachman, D. A. (2005). Do we have brain to spare ? *Neurology*, 64(12):2004–2005.

Duan, Y., Liu, Q., and Xu, X. (2007). Application of reinforcement learning in robot soccer. *Engineering Applications of Artificial Intelligence*, 20(7):936—-950.

Dutt, V., Arló-Costa, H., Helzner, J., and Gonzalez, C. (2013). The Description-Experience Gap in Risky and Ambiguous Gambles. *Journal of Behavioral Decision Making*, 327(October 2013):316–327.

Edwards, A. (1972). *Likelihood*. Cambridge University Press, Cambridge.

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychological Review*, 69(2):109.

Ellsberg, D. (1960). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, 19(May 2013):643–669.

Ellsberg, D. (1963). [Risk, Ambiguity, and the Savage Axioms]: Reply. *The Quarterly Journal of Economics*, 77(2):336–342.

Erev, I. and Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, 112(4):912–931.

Erev, I., Ert, E., and Yechiam, E. (2008). Loss aversion, diminishing Sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making*, 21(5):575–597.

Erev, I., Haruvy, E., Kagel, J. H., and Roth, A. E. (2012). Learning and the Economics of Small Decisions. *The Handbook of Experimental Economics*, pages 1–124.

Erev, I., Haruvy, E., and Yechiam, E. (2003). A strategic problem in small decisions: Implications to skill learning.

Erev, I. and Roth, A. (1998). Predicting how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review*, 88(4):848–881.

Erev, I. and Roth, A. E. (2014). Maximization, learning, and economic behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 111(January):10818–10825.

Estes, W. (1976a). Some functions of memory in probability learning and choice behavior. *Psychology of Learning and Motivation*, 10:1–45.

Estes, W. K. (1976b). The Cognitive Side of Probability Learning. *Psychological Review*, 83(1):37–64.

Everitt, B. J. and Wolf, M. E. (2002). Psychomotor stimulant addiction: a neural systems perspective. *The Journal of Neuroscience*, 22(9):3312–3320.

Fama, E. F. and French, K. R. (1988). Permanent and Temporary Components of Stock Prices. *Journal of Political Economy*, 96(2):246.

Fioravante, D. and Regehr, W. G. (2011). Short-term forms of presynaptic plasticity. *Current Opinion in Neurobiology*, 21(2):269–274.

Fisher, K. J. and Statman, M. (2000). Investor sentiment and stock returns. *Financial Analysts Journal*, 56(2):16–23.

Frackowiak, R. and Markram, H. (2015). The future of human cerebral cartography : a novel approach. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370.

Frazzini, A. (2006). The disposition effect and underreaction to news. *Journal of Finance*, 61(4):2017–2046.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time Discounting and time preference: a critical review. *Journal of Economic Literature*, XL:351–401.

Frey, R., Rieskamp, J., and Hertwig, R. (2015). Sell in May and Go Away ? Learning and Risk Taking in Nonmonotonic Decision Problems. *Journal of Experimental Psychology*, 41(1):193–208.

Friston, K. J. (2011). Functional and Effective Connectivity : A Review. *Brain Connectivity*, 1(1):13–36.

Garivier, A. and Cappè, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359—-376.

Gaskett, C., Fletcher, L., and Zelinsky, A. (2000). Reinforcement learning for a vision based mobile robot. In *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)(Cat. No. 00CH37113)*, pages 403—-409.

Gelfand, E. A. and Dey, D. K. (1994). Bayesian Model Choice : Asymptotics and Exact Calculations. *J. of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Glimcher, P. W. and Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. *Science (New York, N.Y.)*, 306(5695):447–52.

Gneezy, U. and Potters, J. (1997). An Experiment on Risk Taking and Evaluation Periods. *The Quarterly Journal of Economics*, 112(2):631–645.

Gold, J. I. and Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli.

Gold, J. I. and Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of Neuroscienceeuroscience*, 30:535–574.

Goldberg, David E and Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine Learning*, 3(2):95—-99.

Grant, S., London, E. D., Newlin, D. B., Villemagne, V. L., Liu, X., Contoreggi, C., Phillips, R. L., Kimes, A. S., and Margolin, A. (1996). Activation of memory circuits during cue-elicited cocaine craving. *Proceedings of the National Academy of Sciences*, 93(21):12040–12045.

Grinblatt, M. and Keloharju, M. (2001). What Makes Investors Trade? *The Journal Of Finance*, 56 (2)(2):549–578.

Guenter, F., Hersch, M., Calinon, S., and Billard, A. (2007). Reinforcement learning for imitating constrained reaching movements. *Advanced Robotics*, 21(13):1521—-1544.

Gullapalli, V., Franklin, J. A., and Benbrahim, H. (1994). Acquiring robot skills via reinforcement learning. *IEEE Control Systems*, 14(1):13—-24.

Hai, A., Shappir, J., and Spira, M. E. (2010). In-cell recordings by extracellular microelectrodes. *Nature Methods*, 7(3):200–202.

Hailu, G. and Sommer, G. (1998). Integrating symbolic knowledge in reinforcement learning. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, pages 1491—-1496.

Haith, A. M., Reppert, T. R., and Shadmehr, R. (2012). Evidence for hyperbolic temporal discounting of reward in control of movements. *The Journal of Neuroscience*, 32(34):11727–36.

Haruvy, E. and Erev, I. (2002). On the Application and Interpretation of Learning Models. In *Experimental business research*, number 2002, pages 285–300.

Haruvy, E., Erev, I., and Sonsino, D. (2001). The Medium Prizes Paradox: Evidence from a Simulated Casino. *Journal of Risk and Uncertainty*, 22(3):251–261.

Hau, R., Pleskac, T. J., and Hertwig, R. (2010). Decisions From Experience and Statistical Probabilities: Why They Trigger Different Choices Than a Priori Probabilities. *The Journal of Behavioral Decision Making*.

Hau, R., Pleskac, T. J., Kiefer, J., and Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21(5):493–518.

Heath, C., Huddart, S., and Lang, M. (1999). Psychological Factors and Stock Option Exercise. *The Quarterly Journal of Economics*, 114(May):601–627.

Hebb, D. O. (1949). *The organisation of behaviour: a neuropsychological theory*. Wiley.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2006). The role of information sampling in risky choice. In *Information Sampling and Adaptive Cognition*, pages 72–91. Cambridge University Press New York, NY, New York.

Hertwig, R., Barron, G., Weber, E. U., Erev, I., Hertwing, R., Barron, G., Weber, E. U., Erev, I., Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions From Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8):534–539.

Hertwig, R. and Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12):517–523.

Hertwig, R. and Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115(2):225–237.

Hoffrage, U., Weber, A., Hertwig, R., and Chase, V. M. (2003). How to keep children safe in traffic: find the daredevils early. *Journal of experimental psychology. Applied*, 9(4):249–260.

Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature neuroscience*, 1(4):304–9.

Houk, J. C. and Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. *Cerebral cortex (New York, N.Y. : 1991)*, 5(2):95–110.

Hsu, Y. L. and Chow, E. H. (2013). The house money effect on investment risk taking: Evidence from Taiwan. *Pacific Basin Finance Journal*, 21(1):1102–1115.

Huang, X. (2012). Industry Investment Experience and Stock Selection. *Available at SSRN 1786271*, -(November).

Huelsenbeck, J. P. and Crandall, K. a. (1997). Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood. *Annual Review of Ecology and Systematics*, 28(1):437–466.

Huettel, S. A., Song, A. W., and McCarthy, G. (2005). Decisions under Uncertainty: Probabilistic Context Influences Activation of Prefrontal and Parietal Cortices. *Journal of Neuroscience*, 25(13):3304–3311.

Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J., and Dayan, P. (2016). The modulation of savouring by prediction error and its effects on choice. *eLife*, 5(APRIL2016):1–24.

Inta, D., Meyer-Lindenberg, A., and Gass, P. (2011). Alterations in postnatal neurogenesis and dopamine dysregulation in schizophrenia: A hypothesis. *Schizophrenia Bulletin*, 37(4):674–680.

Ippolito, R. A. (1992). Consumer Reaction to Measures of Poor Quality: Evidence from the Mutual Fund Industry. *The Journal of Law & Economics*, 35(1):45–70.

Jensen, N. E. (1967). An Introduction to Bernoullian Utility Theory: I. Utility Functions. *The Swedish Journal of Economics*, 69(3):163–183.

Jessup, R. K., Bishara, A. J., and Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science*, 19(10):1015–1022.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, 15(4-6):535–547.

Kable, J. W. and Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12):1625–1633.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99—-134.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin UK.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis Of Decision Under Risk. *Econometrica*, 47(2):263–291.

Kahneman, D. and Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39(4):341.

Kalakrishnan, M., Righetti, L., Pastor, P., and Schaal, S. (2011). Learning force control policies for compliant manipulation. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 4639—-4644.

Kass, R. and Raftery, A. (1995). Bayes Factors. *J. Amer. Statist. Assoc.*, 90(430):773–795.

Katz, L. (1964). Effects of Differential Monetary Gain and Loss on Sequential Two-Choice Behavior. *Journal of Experimental Psychology*, 68(3):245–249.

Kaustia, M. and Knüpfer, S. (2008). Do investors learn from personal experience ? Evidence from IPO subscriptions. *The Journal of Finance*, 63(6):2679–2702.

Keren, G. (1991). Calibration and Probability Judgments: Conceptual and Methodological Issues. *Acta Psychologica*, 77:217–273.

Keren, G. and Wagenaar, W. A. (1987). Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(3):387–391.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Houghton mifflin company.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., and Glover, G. (2005). Distributed Neural Representation of Expected Value. *Journal of Neuroscience*, 25(19):4806–4812.

Kobayashi, S. and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *The Journal of Neuroscience*, 28(31):7837–7846.

Kober, J., Andrew Bagnell, J., and Peters, J. (2013). Reinforcement Learning in Robotics: A Survey. *The International Journal of Robotics Research*, 32(11):1238–1274.

Kober, J., Mohler, B., and Peters, J. (2008). Learning perceptual coupling for motor primitives. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 834—-839.

Kober, J. and Peters, J. (2009). Policy Search for Motor Primitives in Robotics. In *Advances in neural information processing systems*, pages 849–856.

Kohl, Nate and Stone, P. (2004). Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, pages 2619—-2624.

Kollar, Thomas and Roy, N. (2008). Trajectory optimization using reinforcement learning for map exploration. *The International Journal of Robotics Research*, 27(2):175–196.

Kuhnen, C. M. and Knutson, B. (2005). The neural basis of financial risk taking. *Neuron*, 47(5):763–770.

Kullmann, D. M. and Lamsa, K. P. (2007). Long-term synaptic plasticity in hippocampal interneurons. *Nature Reviews Neuroscience*, 8(9):687–699.

Kurth-Nelson, Z. and Redish, A. D. (2010). A reinforcement learning model of precommitment in decision making. *Frontiers in behavioral neuroscience*, 4(December):184.

Lakonishok, J., Shleifer, A., and Vishny, R. W. (1992). The Structure and Performance of the Money Management Industry. *Brookings Papers on Economic Activity*, 1992(1992):Brookings Papers on Economic Activity.

Lau, B. and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, 84(3):555–79.

Lee, J. (2001). Stock price prediction using reinforcement learning. *Industrial Electronics. Proceedings. ISIE 2001. IEEE International Symposium on*, 1:690–695.

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., and Read, J. P. (2003a). The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology*, 11(1):26–33.

Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., and Pedulla, C. M. (2003b). Evaluation of the Balloon Analogue Risk Task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescence*, 26(4):475–479.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., and Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, 8(2):75–84.

Lichtman, J. W., Livet, J., and Sanes, J. R. (2008). A technicolour approach to the connectome. *Nature Reviews Neuroscience*, 9:417–422.

Lichtman, J. W. and Sanes, J. R. (2008). Ome sweet ome : what can the genome tell us about the connectome ? *Current Opinion in Neurobiology*, 18:346–353.

Loewenstein, G. and Donoghue, T. O. (2005). Animal Spirits : Affective and Deliberative Processes in Economic Behavior.

Loewenstein, G., Rick, S., and Cohen, J. D. (2008). Neuroeconomics. *Annual review of psychology*, 59:647–72.

Lohrenz, T., McCabe, K., Camerer, C. F., and Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences of the United States of America*, 104(22):9493–8.

Lopes, L. L. (1981). Notes, Comments, and New Findings Decision Making in the Short Run. *Journal of Experimental Psychology: Human Learning and Memory*, 7(5):377–385.

Luhmann, C. C. (2013). Discounting of delayed rewards is not hyperbolic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4):1274–9.

Maas, L. C., Lukas, S. E., Kaufman, M. J., Weiss, R. D., Daniels, S. L., Rogers, V. W., Kukes, T. J., and Renshaw, P. F. (1998). Functional magnetic resonance imaging of human brain activation during cue-induced cocaine craving. *Am.J Psychiatry*, 155(February):124–126.

Mahlon, D. and Thomas, W. (2009). Update on models of basal ganglia function and dysfunction. *Parkinsonism & Related Disorders*, 15:S237—-S240.

Malmendier, U. and Nagel, S. (2011). Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?*. *The Quarterly Journal of Economics*.

Markowitz, H. (1952). The Utility of Wealth.

Mehlhorn, K., Ben-Asher, N., Dutt, V., and Gonzalez, C. (2013). Observed Variability and Values Matter: Toward a Better Understanding of Information Search and Decisions from Experience. *Journal of Behavioral Decision Making*, 339(November 2013):328–339.

Mehra, R. (2003). The equity premium: Why is it a puzzle? *Financial Analysts Journal*, 59(1):54–69.

Mehra, R. and Prescott, E. (1985). The equity premium: A puzzle. *Journal of monetary Economics*, 15:145–161.

Merton, R. C. (1973). An Intertemporal Capital Asset Pricing Model. *Econometrica: Journal of the Econometric SocietySociety*, 41(5):867–887.

Miller, B. R., Walker, A. G., Shah, A. S., Barton, S. J., and Rebec, G. V. (2008). Dysregulated information processing by medium spiny neurons in striatum of freely behaving mouse models of Huntington's disease. *Journal of Neurophysiology*, 100:2205–2216.

Minderer, M., Liu, W., Sumanovski, L. T., Sebastian, K., Helmchen, F., and Margolis, D. J. (2012). Chronic imaging of cortical sensory map dynamics using a genetically encoded calcium indicator. *The Journal of physiology*, 590(1):99–107.

Mirenowicz, J. and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, 72(2):1024–1027.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 16(5):1936–1947.

Montague, P. R., Hyman, S. E., and Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, 431(October):760–767.

Moody, J. and Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889.

Moore, K. L., Dalley, A. F., and Agur, A. M. (2013). *Clinically oriented anatomy*.

Myers, J. L. and Sadler, E. (1960). Effects of range of payoffs as a variable in risk taking. *Journal of Experimental Psychology*, 60(5):306–309.

Nevmyvaka, Y., Feng, Y., and Kearns, M. (2006). Reinforcement learning for optimized trade execution. *Proceedings of the 23rd international conference on Machine learning ICML 06*, 17(1):673–680.

Newell, B. R. and Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, 14(6):1133–1139.

Neyman, J. (1935). On the problem of confidence intervals. *The annals of mathematical statistics*, 6(3):111–116.

Ng, A. Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., and Berger, Eric Liang, E. (2006). Autonomous inverted helicopter flight via reinforcement learning. In *Experimental Robotics IX*.

Nilsson, H., Rieskamp, J., and Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, 55(1):84—-93.

Niv, Y., Edlund, J., Dayan, P., and O'Doherty, J. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562.

Niv, Y., Joel, D., Meilijson, I., and Ruppin, E. (2002). Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviors. *Adaptive Behavior*, 10(1):5–24.

O, J., Lee, J., Lee, J., and Zhang, B. (2006). Adaptive stock trading with dynamic asset allocation using reinforcement learning. *Information Sciences*, 176(15):2121–2147.

Odean, T. (1998). Are Investors Reluctant to Realize Their Losses? *The Journal of finance*, 53(5):1775–1798.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental. *Science*, 304:452–454.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain RID D-9230-2011. *Neuron*, 38(2):329–337.

O'Doherty, J. P., Hampton, A., and Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104:35–53.

Offerman, T. and Sonnemans, J. (2004). What's Causing Overreaction? An Experimental Investigation of Recency and the Hot-Hand Effect. *Small*, 106(3):533–553.

Park, J. and Sandberg, I. (1991). Universal Approximation using Radial-Basis-Function Networks. *Neural Computation*, 3(2):246–257.

Pastore, A., Esposito, U., and Vasilaki, E. (2015). Modelling stock-market investors as Reinforcement Learning agents. In *2015 IEEE International Conference on Evolving and Adaptive Intelligent Systems, EAIS 2015*.

Paulus, M. P., Rogalsky, C., Simmons, A., Feinstein, J. S., and Stein, M. B. (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *NeuroImage*, 19(4):1439–1448.

Pavlov, I. P. (2010). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, 17(3):136 – 141.

Perin, R., Berger, T. K., and Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. *Proceedings of the National Academy of Sciences*, 108(13):5419–5424.

Peters, J. and Schaal, S. (2008). Learning to control in operational space. *The International Journal of Robotics Research*, 27(2):197–212.

Peters, J., Vijayakumar, S., and Schaal, S. (2004). Linear quadratic regulation as benchmark for policy gradient methods. *Los Angeles, CA: USC Technical Report*.

Platt, M. L. and Huettel, S. a. (2008). Risky business: the neuroeconomics of decision making under uncertainty. *Nature Neuroscience*, 11(4):398–403.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of experimental psychology. Learning, memory, and cognition*, 34(1):167–85.

Plonsky, O., Teodorescu, K., and Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review*, 122(4):621–647.

Post, T., Assem, M. J. V. D., Baltussen, G., and Thaler, R. H. (2008). Deal or No Deal ? Decision Making under Risk in a Large-Payoff Game. *The American Economic Review*, 98(1):38–71.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163.

Rakow, T., Demes, K. A., and Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, 106(2):168–179.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science (New York, N.Y.)*, 253(5023):980–986.

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience*, 7(12):967–975.

Redgrave, P., Prescott, T. J., and Gurney, K. (1999). the Basal Ganglia: a Vertebrate Solution To the Selection Problem? *Neuroscience*, 89(4):1009–1023.

Reiner, A., Albin, R. L., Anderson, K. D., D'Amato, C. J., Penney, J. B., and Young, A. B. (1988). Differential loss of striatal projection neurons in Huntington disease. *Proceedings of the National Academy of Sciences*, 85(15):5733–5737.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II Current Research and Theory*, pages 64–99. Appleton-Century-Crofts, New York.

Rivlin, T. J. (1969). *An introduction to the approximation of functions*.

Roberts, John W and Moret, Lionel and Zhang, Jun and Tedrake, R. (2010). Motor learning at intermediate Reynolds number: experiments with policy gradient on the flapping flight of a rigid wing. In *From motor learning to interaction learning in robots*, pages 293—-309.

Rubinstein, Reuven Y and Kroese, D. P. (2004). The cross-entropy method: A unified approach to Monte Carlo simulation, randomized optimization and machine learning. *Information Science & Statistics, Springer Verlag, NY*.

Rumelhart, D. E., Mcclelland, J. L., , and Others, P. R. G. (1987). *Parallel distributed processing*, volume 2. MIT Press, Cambridge, MA.

Rutkauskas, A. V. and Ramanauskas, T. (2009). Building an artificial stock market populated by reinforcement-learning agents. *Journal of Business Economics and Management*, 10(4):329–341.

Sanfey, A. G., Loewenstein, G., McClure, S. M., and Cohen, J. D. (2006). Neuroeconomics: cross-currents in research on decision-making. *Trends in cognitive sciences*, 10(3):108–16.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science*, 300(5626):1755–1758.

Sato, Masa-aki and Nakamura, Yutaka and Ishii, S. (2002). Reinforcement learning for biped locomotion. In *International Conference on Artificial Neural Networks*, pages 777—-782.

Schoenbaum, G. and Setlow, B. (2005). Cocaine makes actions insensitive to outcomes but not extinction: Implications for altered orbitofrontal-amygdalar function. *Cerebral Cortex*, 15(8):1162–1169.

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80(1):1–27.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306):1593–9.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Sejnowski, T. J. and Rosenberg, C. R. (1987). Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, 1:145–168.

Seung, H. S. (2009). Perspective Reading the Book of Memory : Sparse Sampling versus Dense Mapping of Connectomes. *Neuron*, 62(1):17–29.

Shadlen, M. N., Britten, K. H., Newsome, W. T., and Movshon, J. a. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci*, 16(4):1486–1510.

Shadlen, M. N. and Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proceedings of the National Academy of Sciences of the United States of America*, 93(January):628–633.

Shapira, Z. and Venezia, I. (2001). Patterns of behavior of professionally managed and independent investors. *Journal of Banking and Finance*, 25(8):1573–1587.

Sharpe, W. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance*, XIX(3):425–442.

Shefrin, H. and Statman, M. (1985). The disposition to sell winners too early and ride losers too long: theory and evidence. *The Journal of Finance*, 15(3).

Sherman, M. S. and Guillery, R. W. (2006). *Exploring the Thalamus and Its Role in Cortical Function*. Cambridge University Press.

Sherman, S. M. (2007). The Thalamus Is More Than Just A Relay. *Current Opinion in Neurobiology*, 17(4):417–422.

Sherman, S. M. and Guillery, R. W. (2001). *Exploring the Thalamus*. Academic Press.

Siegel, J. J. (1998). *Stocks for the Long Run: The Definitive Guide to Financial Market Returns and Long-Term Investment Strategies*. McGraw-Hill, New York.

Sirri, E. R. and Tufano, P. (1998). Costly Search and Mutual Fund Flows. *The Journal of Finance*, 53(5):1589–1622.

Slovic, P. and Tversky, A. (1974). Who Accepts Savag's Axiom. *Oregon Research Institute*.

Song, S., Sjostrom, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly Nonrandom Features of Synaptic Connectivity in Local Cortical Circuits. *Plos Biology*, 3(3).

Sporns, O., Tononi, G., and Kötter, R. (2005). The Human Connectome : A Structural Description of the Human Brain. *PLoS Comput Biol*, 1(4).

Starmer, C. (2000). Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk.

Stocco, A., Lebiere, C., and Anderson, J. R. (2011). Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination. *Psychological Review*, 117(2):541–574.

Story, G. W., Vlaev, I., Seymour, B., Darzi, A., and Dolan, R. J. (2014). Does temporal discounting explain unhealthy behavior? A systematic review and reinforcement learning perspective. *Frontiers in behavioral neuroscience*, 8(March):76.

Strahilevitz, M., Barber, B., and Odean, T. (2011). Once burned, twice shy: How naïve learning, counterfactuals, and regret affect the repurchase of stocks previously sold. *Journal of Marketing Research*, XLVIII:102–120.

Strens, Malcolm JA and Moore, A. W. (2001). Direct Policy Search using Paired Statistical Tests. In *ICML*, pages 545—-552.

Sulzer, D. (2005). The complex regulation of dopamine output: A review of current themes. *Clinical Neuroscience Research*, 5(2-4):117–121.

Suri, R. E. and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Exp. Brain Res.*, 121(3):350–354.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An introduction*. MIT Press, Cambridge, MA.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcment learning*.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. *Advances in neural information processing systems*.

Symonds, M. R. E. and Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, 65(1):13–21.

Thaler, R. and Johnson, E. J. (1990). Gambling with the House Money and Trying to Break Even : The Effects of Prior Outcomes on Risky Choice. *Management science*, 36(6):643–660.

Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1:39–60.

Thaler, R. H. (1981). Some Empirical Evidence on Dynamic Inconsistency.

Thaler, R. H. (1985). Mental Accounting and Consumer Choice. *Marketing Science*, 4(3):199–214.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, 12(3):183–206.

Thaler, R. H., Tversky, A., Kahneman, D., and Schwartz, A. (1997). The Effect of Myopia and Loss Aversion on Risk Taking : An Experimental Test. *The Quarterly Journal of Economics*, 112(2):647–661.

Thomson, A. M. (2000). Facilitation , augmentation and potentiation at central synapses. *Trends in Neurosciences*, 23(7):305–312.

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, 2:i.

Thrun, S. (1995). An approach to learning mobile robot navigation. *Robotics and Autonomous systems*, 15(4):301—-319.

Tobler, P. N., Fletcher, P. C., Bullmore, E. T., and Schultz, W. (2007). Learning-Related Human Brain Activations Reflecting Individual Finances. *Neuron*, 54(1):167–175.

Toga, A. W., Clark, K. A., Thompson, P. M., Shattuck, D. W., and Horn, J. D. V. (2012). Mapping the Human Connectome. *Neurosurgery*, 71(1):1–5.

Tom, S. M., Fox, C. R., Trepel, C., and Poldrack, R. a. (2007). The neural basis of loss aversion in decision-making under risk. *Science (New York, N.Y.)*, 315(5811):515–518.

Tomkins, A. (2015). *Action selection in the striatum : Implications for Huntington ' s disease*. PhD thesis, The University Of Sheffield.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty : heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Tversky, A. and Kahneman, D. (1992). Advances in Prospect-Theory - Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Van den Heuvel, D. M. and Pasterkamp, R. J. (2008). Getting connected in the dopamine system. *Progress in Neurobiology*, 85(1):75–93.

Vernade, C., Cappè, O., and Perchet, V. (2017). Stochastic Bandit Models for Delayed Conversions. *arXiv preprint arXiv:1706.09186*.

Vlassis, Nikos and Toussaint, Marc and Kontes, Georgios and Piperidis, S. (2009). Learning model-free robot control by a Monte Carlo EM algorithm. *Autonomous Robots*, 27(2):123–130.

von Neumann, J. and Morgenstern, O. (1947). Theory of Games and Economic Behavior.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228–243.

Wagenmakers, E.-j. and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, 11(1):192–196.

Wallsten, T. S., Pleskac, T. J., and Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. *Psychological Review,*, 112(4):862–880.

Wang, G. J., Volkow, N. D., Fowler, J. S., Cervany, P., Hitzemann, R. J., Pappas, N. R., Wong, C. T., and Felder, C. (1999). Regional brain metabolic activation during craving elicited by recall of previous drug experiences. *Life Sciences*, 64(9):775–784.

Watkins, C. J. C. H. (1989). *Chris Watkins' PhD thesis*. PhD thesis, University of Cambridge.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3-4):279–292.

Weber, M. and Camerer, C. F. (1998). The Disposition Effect in Securities Trading: An Experimental Analysis. *Journal of Economic Behavior & Organization*, 33(2):167–184.

Weber, M. and Welfens, F. (2011). The Follow-On Purchase and Repurchase Behavior of Individual Investors: An Experimental Investigation. *Die Betriebswirtschaft*, 71(January):139–154.

Wedeen, V. J., Rosene, D. L., Wang, R., Dai, G., Mortazavi, F., Hagmann, P., Kaas, J. H., and Tseng, W.-Y. I. (2012). The Geometric Structure of the Brain Fiber Pathways. *Science*, 335(6076):1628–1634.

Wedell, D. H. and Böckenholt, U. (1990). Moderation of Preference Reversals in the Long Run. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):429–438.

Wickens, J. R., Horvitz, J. C., Costa, R. M., and Killcross, S. (2007). Dopaminergic mechanisms in actions and habits. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(31):8181–8183.

Wickersham, I. R., Lyon, D. C., Barnard, R. J. O., Mori, T., Finke, S., Conzelmann, K.-k., Young, J. A. T., and Callaway, E. M. (2007). Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron*, 53(5):639–647.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229—-256.

Yechiam, E. and Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, 19(1):1–16.

Ylöstalo, J. (2006). Function approximation using polynomials. *IEEE Signal Processing Magazine*, 23(5):99–102.

Zeckhauser, R. (1993). Hot Hands in Mutual Funds: Short-Run Persistence of Relative Performance, 1974-1988. *The Journal of Finance*, 48(1):93.

Zhang, F., Aravanis, A. M., Adamantidis, A., de Lecea, L., and Deisseroth, K. (2007). Circuit-breakers: optical technologies for probing neural signals and systems. *Nature Reviews Neuroscience*, 8(8):577–581.

Zion, U. B., Erev, I., Haruvy, E., and Shavit, T. (2010). Adaptive behavior leads to underdiversification. *Journal of Economic Psychology*, 31(6):985–995.

Appendix A

The following tables present the summaries of models scores, for each subject in the three conditions of the experiment analysed in chapter 4. The first column, denoted with "Config", is composed of three acronyms identifying the combination of state-space, learning rule and reward function defining a model. The legend for these acronyms is reported in table A.1. The second column shows the number of parameters for a model. The remaining columns show measures of model fitness and representativeness. Specifically, column three reports the log-likelihood score, column four shows the AIC score, column five is the AIC difference of a model compared to the best model in the set. The last column represents the Akaike weight for each model, based on its AIC score. Lines in bold font identify the models which have been selected as representative of the behavioural data for a specific subject, according to the methodology described in 4, section 4.1.15. At the end of the tables for each condition there is a graphical depiction of these results in the form of bar charts, with a detailed description in the caption.

Table A.1 Abbreviations for model components

State space	Learning Rule	Reward Function
SS: Single state **FH**: Full history **LO**: Latest outcome	**AT**: Average Tracking **QL**: Q-learning	**ID**: Identity **TH**: Hyperbolic tangent **PT**: Prospect theory's value function

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	1.000	0.003	0.000	-123.161	252.321	6.089	0.022
FH	**QL**	**TH**	**3**	**1.000**	**0.058**	**0.000**	**-122.692**	**251.384**	**5.152**	**0.035**
SS	AT	TH	2	1.000	0.038	0.000	-132.259	268.517	22.285	0.000
SS	AT	ID	2	1.000	0.002	0.000	-132.347	268.694	22.461	0.000
LO	AT	PT	2	0.056	0.010	0.000	-133.679	271.358	25.125	0.000
LO	AT	TH	2	1.000	0.019	0.000	-137.078	278.157	31.924	0.000
LO	QL	PT	3	0.919	0.002	1.000	-126.542	259.085	12.852	0.001
FH	**QL**	**PT**	**3**	**1.000**	**0.003**	**1.000**	**-120.499**	**246.998**	**0.766**	**0.317**
LO	QL	ID	3	0.817	0.002	1.000	-126.766	259.532	13.299	0.001
FH	**AT**	**PT**	**2**	**1.000**	**0.003**	**0.000**	**-121.116**	**246.233**	**0.000**	**0.466**
FH	**AT**	**ID**	**2**	**1.000**	**0.003**	**0.000**	**-123.161**	**250.321**	**4.088**	**0.060**
FH	**AT**	**TH**	**2**	**1.000**	**0.058**	**0.000**	**-122.692**	**249.384**	**3.151**	**0.096**
LO	QL	TH	3	0.813	0.038	1.000	-126.090	258.179	11.947	0.001
SS	AT	PT	2	0.062	0.014	0.000	-128.736	261.472	15.239	0.000
LO	AT	ID	2	1.000	0.001	0.000	-137.044	278.087	31.855	0.000

Table A.2 Models summary: problem 1 - subject 1

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.002**	**1.000**	**-119.590**	**245.180**	**1.249**	**0.247**
FH	**QL**	**TH**	**3**	**1.000**	**0.050**	**1.000**	**-118.965**	**243.931**	**0.000**	**0.462**
SS	AT	TH	2	0.051	0.023	0.000	-138.543	281.086	37.156	0.000
SS	AT	ID	2	0.054	0.001	0.000	-138.606	281.212	37.282	0.000
LO	AT	PT	2	0.001	1.000	0.000	-126.767	257.533	13.603	0.001
LO	AT	TH	2	0.000	0.000	0.000	-138.629	281.259	37.328	0.000
LO	QL	PT	3	0.001	0.553	1.000	-126.743	259.487	15.556	0.000
FH	**QL**	**PT**	**3**	**1.000**	**0.003**	**1.000**	**-119.492**	**244.985**	**1.054**	**0.273**
LO	QL	ID	3	0.151	0.006	1.000	-128.015	262.030	18.100	0.000
FH	AT	PT	2	0.026	0.024	0.000	-125.359	254.717	10.787	0.002
FH	AT	ID	2	0.052	0.005	0.000	-137.356	278.712	34.782	0.000
FH	AT	TH	2	0.047	0.133	0.000	-136.879	277.759	33.828	0.000
LO	QL	TH	3	0.143	0.140	1.000	-127.686	261.371	17.441	0.000
SS	AT	PT	2	0.024	0.026	0.000	-123.408	250.816	6.886	0.015
LO	AT	ID	2	0.000	0.000	0.000	-138.629	281.259	37.328	0.000

Table A.3 Models summary: problem 1 - subject 2

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.000**	**0.983**	**1.000**	**-112.463**	**230.927**	**1.713**	**0.177**
FH	QL	TH	3	0.011	1.000	1.000	-114.405	234.810	5.597	0.025
SS	AT	TH	2	0.008	1.000	0.000	-119.799	243.598	14.385	0.000
SS	**AT**	**ID**	**2**	**0.000**	**1.000**	**0.000**	**-112.700**	**229.401**	**0.188**	**0.380**
LO	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	52.046	0.000
LO	AT	TH	2	0.006	1.000	0.000	-126.115	256.231	27.017	0.000
LO	QL	PT	3	1.000	0.001	1.000	-137.141	280.282	51.068	0.000
FH	QL	PT	3	1.000	0.002	1.000	-132.682	271.364	42.151	0.000
LO	QL	ID	3	0.000	0.799	1.000	-123.288	252.576	23.363	0.000
FH	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	52.046	0.000
FH	**AT**	**ID**	**2**	**0.000**	**1.000**	**0.000**	**-112.607**	**229.213**	**0.000**	**0.417**
FH	AT	TH	2	0.009	1.000	0.000	-121.019	246.038	16.824	0.000
LO	QL	TH	3	0.007	1.000	1.000	-125.261	256.522	27.309	0.000
SS	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	52.046	0.000
LO	AT	ID	2	0.000	1.000	0.000	-123.317	250.634	21.421	0.000

Table A.4 Models summary: problem 1 - subject 3

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.073	0.023	0.677	-82.772	171.544	8.553	0.005
FH	QL	TH	3	0.076	0.464	0.672	-82.210	170.420	7.429	0.008
SS	AT	TH	2	0.047	0.878	0.000	-84.399	172.798	9.807	0.002
SS	AT	ID	2	0.043	0.044	0.000	-84.966	173.932	10.941	0.001
LO	**AT**	**PT**	**2**	**0.154**	**0.016**	**0.000**	**-81.633**	**167.266**	**4.275**	**0.040**
LO	AT	TH	2	0.049	0.899	0.000	-85.179	174.358	11.367	0.001
LO	**QL**	**PT**	**3**	**0.181**	**0.012**	**0.818**	**-79.961**	**165.922**	**2.931**	**0.078**
FH	**QL**	**PT**	**3**	**0.152**	**0.013**	**0.829**	**-78.495**	**162.991**	**0.000**	**0.336**
LO	QL	ID	3	0.105	0.017	0.790	-83.945	173.889	10.898	0.001
FH	**AT**	**PT**	**2**	**0.141**	**0.016**	**0.000**	**-79.850**	**163.700**	**0.710**	**0.236**
FH	AT	ID	2	0.047	0.040	0.000	-83.713	171.426	8.436	0.005
FH	AT	TH	2	0.050	0.787	0.000	-83.168	170.337	7.346	0.009
LO	QL	TH	3	0.107	0.351	0.776	-83.329	172.659	9.668	0.003
SS	**AT**	**PT**	**2**	**0.145**	**0.017**	**0.000**	**-79.697**	**163.394**	**0.403**	**0.275**
LO	AT	ID	2	0.045	0.046	0.000	-85.810	175.620	12.629	0.001

Table A.5 Models summary: problem 1 - subject 4

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.001	1.000	1.000	-128.453	262.907	54.218	0.000
FH	QL	TH	3	0.022	1.000	1.000	-121.058	248.117	39.428	0.000
SS	AT	TH	2	0.873	0.014	0.000	-137.683	279.365	70.676	0.000
SS	AT	ID	2	1.000	0.000	0.000	-138.103	280.206	71.517	0.000
LO	AT	PT	2	0.004	0.140	0.000	-109.989	223.979	15.290	0.000
LO	AT	TH	2	0.003	1.000	0.000	-137.364	278.729	70.040	0.000
LO	QL	PT	3	0.276	0.007	1.000	-105.719	217.437	8.748	0.007
FH	**QL**	**PT**	**3**	**0.007**	**0.091**	**0.000**	**-102.344**	**210.689**	**2.000**	**0.195**
LO	QL	ID	3	0.315	0.004	1.000	-111.161	228.322	19.633	0.000
FH	**AT**	**PT**	**2**	**0.007**	**0.091**	**0.000**	**-102.344**	**208.689**	**0.000**	**0.529**
FH	AT	ID	2	0.001	1.000	0.000	-131.249	266.497	57.808	0.000
FH	AT	TH	2	0.004	1.000	0.000	-134.978	273.956	65.267	0.000
LO	QL	TH	3	0.271	0.108	1.000	-110.226	226.451	17.762	0.000
SS	**AT**	**PT**	**2**	**0.007**	**0.103**	**0.000**	**-103.023**	**210.046**	**1.357**	**0.269**
LO	AT	ID	2	0.000	1.000	0.000	-138.231	280.461	71.772	0.000

Table A.6 Models summary: problem 1 - subject 5

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.001**	**0.755**	**-133.763**	**273.526**	**1.348**	**0.092**
FH	**QL**	**TH**	**3**	**1.000**	**0.030**	**0.758**	**-133.526**	**273.052**	**0.874**	**0.117**
SS	**AT**	**TH**	**2**	**0.864**	**0.030**	**0.000**	**-134.089**	**272.178**	**0.000**	**0.181**
SS	**AT**	**ID**	**2**	**0.891**	**0.001**	**0.000**	**-134.217**	**272.435**	**0.256**	**0.159**
LO	AT	PT	2	0.502	0.001	0.000	-137.953	279.907	7.728	0.004
LO	**AT**	**TH**	**2**	**0.003**	**1.000**	**0.000**	**-135.450**	**274.900**	**2.722**	**0.046**
LO	QL	PT	3	1.000	0.001	0.942	-136.532	279.065	6.886	0.006
FH	**QL**	**PT**	**3**	**1.000**	**0.002**	**0.803**	**-134.742**	**275.485**	**3.306**	**0.035**
LO	QL	ID	3	0.000	0.700	0.000	-135.489	276.978	4.800	0.016
FH	**AT**	**PT**	**2**	**1.000**	**0.001**	**0.000**	**-135.836**	**275.673**	**3.494**	**0.032**
FH	**AT**	**ID**	**2**	**0.769**	**0.001**	**0.000**	**-134.950**	**273.899**	**1.721**	**0.076**
FH	**AT**	**TH**	**2**	**0.892**	**0.027**	**0.000**	**-134.808**	**273.615**	**1.437**	**0.088**
LO	**QL**	**TH**	**3**	**1.000**	**0.019**	**0.858**	**-135.019**	**276.037**	**3.859**	**0.026**
SS	**AT**	**PT**	**2**	**1.000**	**0.001**	**0.000**	**-134.940**	**273.880**	**1.702**	**0.077**
LO	**AT**	**ID**	**2**	**0.000**	**0.700**	**0.000**	**-135.489**	**274.978**	**2.800**	**0.045**

Table A.7 Models summary: problem 1 - subject 6

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.015**	**0.110**	**0.071**	**-86.036**	**178.073**	**4.747**	**0.053**
FH	**QL**	**TH**	**3**	**0.035**	**1.000**	**0.662**	**-85.459**	**176.918**	**3.592**	**0.094**
SS	AT	TH	2	0.021	1.000	0.000	-97.501	199.001	25.675	0.000
SS	**AT**	**ID**	**2**	**0.011**	**0.138**	**0.000**	**-84.663**	**173.326**	**0.000**	**0.565**
LO	AT	PT	2	0.037	0.037	0.000	-94.714	193.427	20.101	0.000
LO	AT	TH	2	0.022	1.000	0.000	-101.821	207.642	34.316	0.000
LO	QL	PT	3	0.138	0.012	0.940	-90.950	187.899	14.573	0.000
FH	**QL**	**PT**	**3**	**0.129**	**0.015**	**0.924**	**-85.249**	**176.497**	**3.171**	**0.116**
LO	QL	ID	3	0.011	0.136	0.000	-90.801	187.602	14.275	0.000
FH	AT	PT	2	0.066	0.028	0.000	-88.960	181.921	8.594	0.008
FH	**AT**	**ID**	**2**	**0.014**	**0.117**	**0.000**	**-86.037**	**176.074**	**2.747**	**0.143**
FH	AT	TH	2	0.027	1.000	0.000	-93.699	191.398	18.072	0.000
LO	QL	TH	3	0.032	1.000	0.668	-90.514	187.028	13.702	0.001
SS	AT	PT	2	0.033	0.041	0.000	-88.064	180.127	6.801	0.019
LO	AT	ID	2	0.011	0.136	0.000	-90.801	185.602	12.275	0.001

Table A.8 Models summary: problem 1 - subject 7

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	1.000	0.002	1.000	-130.838	267.676	32.803	0.000
FH	QL	TH	3	1.000	0.036	1.000	-130.077	266.155	31.282	0.000
SS	AT	TH	2	0.000	0.000	0.000	-138.629	281.259	46.386	0.000
SS	AT	ID	2	0.000	0.000	0.000	-138.629	281.259	46.386	0.000
LO	**AT**	**PT**	**2**	**0.022**	**0.062**	**0.000**	**-116.894**	**237.788**	**2.915**	**0.152**
LO	AT	TH	2	0.000	0.000	0.000	-138.629	281.259	46.386	0.000
LO	**QL**	**PT**	**3**	**0.083**	**0.015**	**1.000**	**-114.436**	**234.873**	**0.000**	**0.653**
FH	**QL**	**PT**	**3**	**0.042**	**0.026**	**1.000**	**-116.623**	**239.246**	**4.373**	**0.073**
LO	QL	ID	3	1.000	0.000	1.000	-134.157	274.315	39.442	0.000
FH	**AT**	**PT**	**2**	**0.009**	**0.112**	**0.000**	**-117.807**	**239.614**	**4.741**	**0.061**
FH	AT	ID	2	0.000	0.000	0.000	-138.629	281.259	46.386	0.000
FH	AT	TH	2	0.000	0.000	0.000	-138.629	281.259	46.386	0.000
LO	QL	TH	3	0.219	0.106	1.000	-126.516	259.032	24.159	0.000
SS	**AT**	**PT**	**2**	**0.009**	**0.112**	**0.000**	**-117.807**	**239.614**	**4.741**	**0.061**
LO	AT	ID	2	0.000	0.000	0.000	-138.629	281.259	46.386	0.000

Table A.9 Models summary: problem 1 - subject 8

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.005	0.511	0.000	-57.368	120.735	6.575	0.013
FH	QL	TH	3	0.052	1.000	0.915	-57.398	120.796	6.635	0.013
SS	AT	TH	2	0.011	1.000	0.000	-109.298	222.596	108.436	0.000
SS	**AT**	**ID**	**2**	**0.005**	**0.511**	**0.000**	**-57.368**	**118.735**	**4.575**	**0.036**
LO	**AT**	**PT**	**2**	**0.019**	**0.221**	**0.000**	**-58.044**	**120.087**	**5.927**	**0.018**
LO	AT	TH	2	0.011	1.000	0.000	-111.203	226.406	112.245	0.000
LO	**QL**	**PT**	**3**	**0.306**	**0.014**	**0.957**	**-56.690**	**119.379**	**5.219**	**0.026**
FH	**QL**	**PT**	**3**	**0.035**	**0.115**	**0.527**	**-55.054**	**116.109**	**1.948**	**0.133**
LO	QL	ID	3	0.155	0.018	0.950	-58.399	122.797	8.637	0.005
FH	**AT**	**PT**	**2**	**0.018**	**0.223**	**0.000**	**-55.080**	**114.161**	**0.000**	**0.351**
FH	**AT**	**ID**	**2**	**0.005**	**0.511**	**0.000**	**-57.368**	**118.735**	**4.575**	**0.036**
FH	AT	TH	2	0.011	1.000	0.000	-109.298	222.596	108.436	0.000
LO	QL	TH	3	0.162	0.371	0.941	-57.094	120.187	6.027	0.017
SS	**AT**	**PT**	**2**	**0.018**	**0.223**	**0.000**	**-55.080**	**114.161**	**0.000**	**0.351**
LO	AT	ID	2	0.005	0.532	0.000	-60.325	124.649	10.489	0.002

Table A.10 Models summary: problem 1 - subject 9

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.000	1.000	0.000	-128.227	262.454	9.963	0.003
FH	QL	TH	3	0.002	1.000	1.000	-128.381	262.762	10.271	0.003
SS	AT	TH	2	0.002	1.000	0.000	-128.371	260.742	8.252	0.008
SS	AT	ID	2	0.000	0.700	0.000	-128.172	260.344	7.854	0.010
LO	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	28.769	0.000
LO	AT	TH	2	0.017	0.309	0.000	-129.781	263.562	11.072	0.002
LO	QL	PT	3	1.000	0.001	1.000	-132.700	271.401	18.910	0.000
FH	QL	PT	3	0.000	0.000	0.719	-138.629	283.259	30.769	0.000
LO	**QL**	**ID**	**3**	**0.144**	**0.004**	**1.000**	**-123.245**	**252.490**	**0.000**	**0.509**
FH	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	28.769	0.000
FH	AT	ID	2	0.000	0.700	0.000	-128.229	260.458	7.967	0.009
FH	AT	TH	2	0.002	1.000	0.000	-128.425	260.850	8.359	0.008
LO	**QL**	**TH**	**3**	**0.145**	**0.081**	**1.000**	**-123.381**	**252.762**	**0.272**	**0.445**
SS	AT	PT	2	0.000	0.000	0.000	-138.629	281.259	28.769	0.000
LO	AT	ID	2	0.018	0.013	0.000	-129.909	263.818	11.328	0.002

Table A.11 Models summary: problem 1 - subject 10

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.650**	**0.018**	**0.628**	**-15.789**	**37.577**	**2.131**	**0.065**
FH	**QL**	**TH**	**3**	**0.654**	**0.381**	**0.618**	**-15.725**	**37.450**	**2.004**	**0.070**
SS	**AT**	**TH**	**2**	**0.508**	**0.521**	**0.000**	**-15.723**	**35.447**	**0.000**	**0.190**
SS	**AT**	**ID**	**2**	**0.503**	**0.025**	**0.000**	**-15.802**	**35.603**	**0.157**	**0.175**
LO	AT	PT	2	0.517	0.023	0.000	-18.486	40.972	5.525	0.012
LO	**AT**	**TH**	**2**	**0.531**	**0.509**	**0.000**	**-17.327**	**38.653**	**3.206**	**0.038**
LO	QL	PT	3	0.712	0.015	0.835	-17.147	40.293	4.847	0.017
FH	**QL**	**PT**	**3**	**0.630**	**0.017**	**0.737**	**-16.582**	**39.163**	**3.717**	**0.030**
LO	**QL**	**ID**	**3**	**0.714**	**0.016**	**0.730**	**-16.294**	**38.587**	**3.141**	**0.039**
FH	**AT**	**PT**	**2**	**0.494**	**0.024**	**0.000**	**-17.417**	**38.834**	**3.387**	**0.035**
FH	**AT**	**ID**	**2**	**0.502**	**0.025**	**0.000**	**-16.490**	**36.980**	**1.534**	**0.088**
FH	**AT**	**TH**	**2**	**0.507**	**0.522**	**0.000**	**-16.412**	**36.824**	**1.377**	**0.095**
LO	**QL**	**TH**	**3**	**0.717**	**0.338**	**0.718**	**-16.225**	**38.449**	**3.003**	**0.042**
SS	**AT**	**PT**	**2**	**0.494**	**0.024**	**0.000**	**-16.729**	**37.458**	**2.011**	**0.069**
LO	**AT**	**ID**	**2**	**0.526**	**0.025**	**0.000**	**-17.420**	**38.839**	**3.393**	**0.035**

Table A.12 Models summary: problem 1 - subject 11

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.136	0.012	1.000	-122.596	251.191	56.106	0.000
FH	QL	TH	3	0.126	0.265	1.000	-119.979	245.959	50.873	0.000
SS	AT	TH	2	1.000	0.003	0.000	-138.593	281.187	86.101	0.000
SS	AT	ID	2	1.000	0.000	0.000	-138.627	281.255	86.169	0.000
LO	AT	PT	2	0.007	0.189	0.000	-99.014	202.029	6.943	0.013
LO	AT	TH	2	1.000	0.003	0.000	-138.590	281.179	86.094	0.000
LO	QL	PT	3	0.064	0.022	1.000	-97.949	201.897	6.812	0.014
FH	**QL**	**PT**	**3**	**0.004**	**0.261**	**0.000**	**-95.543**	**197.086**	**2.001**	**0.151**
LO	QL	ID	3	0.182	0.008	1.000	-108.076	222.151	27.066	0.000
FH	**AT**	**PT**	**2**	**0.003**	**0.276**	**0.000**	**-95.543**	**195.085**	**0.000**	**0.411**
FH	AT	ID	2	1.000	0.000	0.000	-138.627	281.255	86.169	0.000
FH	AT	TH	2	1.000	0.003	0.000	-138.593	281.187	86.101	0.000
LO	QL	TH	3	0.169	0.182	1.000	-106.727	219.454	24.368	0.000
SS	**AT**	**PT**	**2**	**0.003**	**0.276**	**0.000**	**-95.543**	**195.085**	**0.000**	**0.411**
LO	AT	ID	2	1.000	0.000	0.000	-138.627	281.254	86.169	0.000

Table A.13 Models summary: problem 1 - subject 12

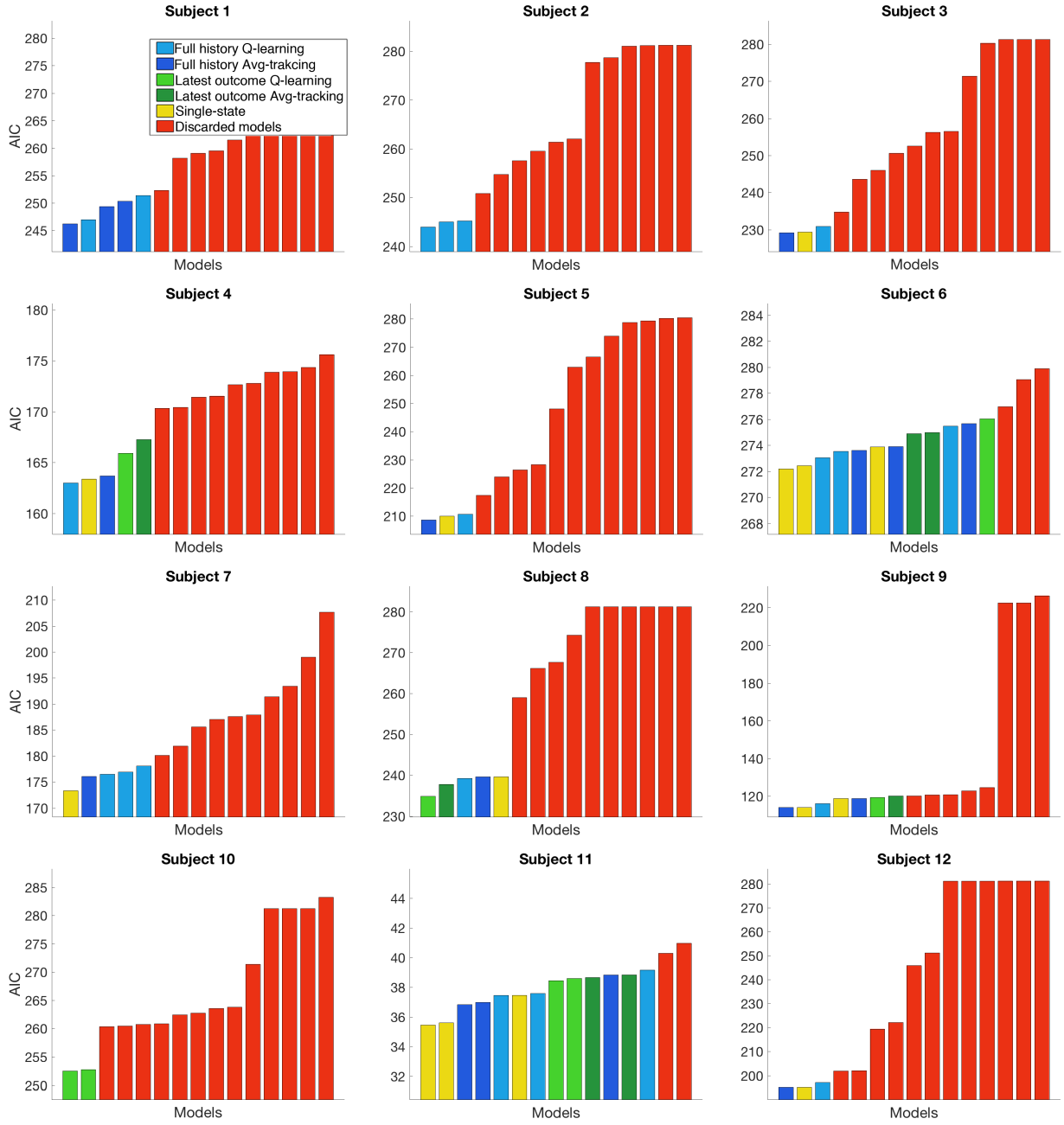

Fig. A.1 Comparison of AIC scores of the 15 models fitted to the subjects in condition 1. Each bar represents a model and its height represents the associated AIC score. The color of the bar indicates the type of model according to the combination of state-space and learning rule adopted. Yellow bars represent the single-state model. Light and dark green represent latest-outcome models combined with Q-learning or average-tracking respectively. Light and dark blue represent full-history models combined with Q-learning or average-tracking respectively. The red bars represent the discarded models.

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.004**	**0.000**	**-113.394**	**232.789**	**2.000**	**0.071**
FH	QL	TH	3	1.000	0.153	0.000	-123.560	253.121	22.332	0.000
SS	AT	TH	2	1.000	0.153	0.000	-123.560	251.121	20.332	0.000
SS	**AT**	**ID**	**2**	**1.000**	**0.004**	**0.000**	**-113.394**	**230.789**	**0.000**	**0.192**
LO	**AT**	**PT**	**2**	**1.000**	**0.010**	**0.000**	**-114.329**	**232.658**	**1.870**	**0.075**
LO	AT	TH	2	1.000	0.153	0.000	-123.560	251.121	20.332	0.000
LO	**QL**	**PT**	**3**	**1.000**	**0.010**	**0.000**	**-114.329**	**234.659**	**3.870**	**0.028**
FH	**QL**	**PT**	**3**	**1.000**	**0.010**	**0.000**	**-114.329**	**234.659**	**3.870**	**0.028**
LO	**QL**	**ID**	**3**	**1.000**	**0.004**	**0.000**	**-113.394**	**232.789**	**2.000**	**0.071**
FH	**AT**	**PT**	**2**	**1.000**	**0.010**	**0.000**	**-114.329**	**232.658**	**1.870**	**0.075**
FH	**AT**	**ID**	**2**	**1.000**	**0.004**	**0.000**	**-113.394**	**230.789**	**0.000**	**0.192**
FH	AT	TH	2	1.000	0.153	0.000	-123.560	251.121	20.332	0.000
LO	QL	TH	3	1.000	0.153	0.000	-123.560	253.121	22.332	0.000
SS	**AT**	**PT**	**2**	**1.000**	**0.010**	**0.000**	**-114.329**	**232.658**	**1.870**	**0.075**
LO	**AT**	**ID**	**2**	**1.000**	**0.004**	**0.000**	**-113.394**	**230.789**	**0.000**	**0.192**

Table A.14 Models summary: problem 2 - subject 1

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.175**	**0.000**	**1.000**	**-132.732**	**271.463**	**0.124**	**0.159**
FH	**QL**	**TH**	**3**	**0.191**	**0.011**	**1.000**	**-132.669**	**271.339**	**0.000**	**0.169**
SS	AT	TH	2	0.044	0.057	0.000	-138.090	280.180	8.841	0.002
SS	AT	ID	2	0.033	0.002	0.000	-137.812	279.624	8.285	0.003
LO	AT	PT	2	0.035	0.005	0.000	-137.826	279.652	8.314	0.003
LO	AT	TH	2	0.044	0.057	0.000	-138.090	280.180	8.841	0.002
LO	**QL**	**PT**	**3**	**0.174**	**0.001**	**1.000**	**-132.712**	**271.425**	**0.086**	**0.162**
FH	**QL**	**PT**	**3**	**0.174**	**0.001**	**1.000**	**-132.712**	**271.425**	**0.086**	**0.162**
LO	**QL**	**ID**	**3**	**0.175**	**0.000**	**1.000**	**-132.732**	**271.463**	**0.124**	**0.159**
FH	AT	PT	2	0.035	0.005	0.000	-137.826	279.652	8.314	0.003
FH	AT	ID	2	0.033	0.002	0.000	-137.812	279.624	8.285	0.003
FH	AT	TH	2	0.044	0.057	0.000	-138.090	280.180	8.841	0.002
LO	**QL**	**TH**	**3**	**0.191**	**0.011**	**1.000**	**-132.669**	**271.339**	**0.000**	**0.169**
SS	AT	PT	2	0.035	0.005	0.000	-137.826	279.652	8.314	0.003
LO	AT	ID	2	0.033	0.002	0.000	-137.812	279.624	8.285	0.003

Table A.15 Models summary: problem 2 - subject 2

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.003**	**0.483**	**-104.249**	**214.498**	**0.000**	**0.297**
FH	QL	TH	3	0.693	0.006	1.000	-112.449	230.898	16.400	0.000
SS	AT	TH	2	1.000	0.100	0.000	-116.176	236.353	21.855	0.000
SS	**AT**	**ID**	**2**	**0.551**	**0.004**	**0.000**	**-106.896**	**217.792**	**3.294**	**0.057**
LO	**AT**	**PT**	**2**	**0.581**	**0.011**	**0.000**	**-108.105**	**220.209**	**5.712**	**0.017**
LO	AT	TH	2	1.000	0.100	0.000	-116.176	236.353	21.855	0.000
LO	**QL**	**PT**	**3**	**1.000**	**0.008**	**0.468**	**-105.428**	**216.855**	**2.358**	**0.091**
FH	**QL**	**PT**	**3**	**1.000**	**0.008**	**0.468**	**-105.428**	**216.855**	**2.358**	**0.091**
LO	**QL**	**ID**	**3**	**1.000**	**0.003**	**0.483**	**-104.249**	**214.498**	**0.000**	**0.297**
FH	**AT**	**PT**	**2**	**0.581**	**0.011**	**0.000**	**-108.105**	**220.209**	**5.712**	**0.017**
FH	**AT**	**ID**	**2**	**0.551**	**0.004**	**0.000**	**-106.896**	**217.792**	**3.294**	**0.057**
FH	AT	TH	2	1.000	0.100	0.000	-116.176	236.353	21.855	0.000
LO	QL	TH	3	0.693	0.006	1.000	-112.449	230.898	16.400	0.000
SS	**AT**	**PT**	**2**	**0.581**	**0.011**	**0.000**	**-108.105**	**220.209**	**5.712**	**0.017**
LO	**AT**	**ID**	**2**	**0.551**	**0.004**	**0.000**	**-106.896**	**217.792**	**3.294**	**0.057**

Table A.16 Models summary: problem 2 - subject 3

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.002**	**0.000**	**-125.669**	**257.337**	**2.000**	**0.048**
FH	QL	TH	3	1.000	0.114	0.000	-127.480	260.960	5.623	0.008
SS	**AT**	**TH**	**2**	**1.000**	**0.114**	**0.000**	**-127.480**	**258.959**	**3.623**	**0.022**
SS	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-125.668**	**255.337**	**0.000**	**0.132**
LO	**AT**	**PT**	**2**	**1.000**	**0.006**	**0.000**	**-125.808**	**255.616**	**0.279**	**0.115**
LO	**AT**	**TH**	**2**	**1.000**	**0.114**	**0.000**	**-127.480**	**258.959**	**3.623**	**0.022**
LO	**QL**	**PT**	**3**	**1.000**	**0.006**	**0.000**	**-125.808**	**257.616**	**2.279**	**0.042**
FH	**QL**	**PT**	**3**	**1.000**	**0.006**	**0.000**	**-125.808**	**257.616**	**2.279**	**0.042**
LO	**QL**	**ID**	**3**	**1.000**	**0.002**	**0.000**	**-125.669**	**257.337**	**2.000**	**0.048**
FH	**AT**	**PT**	**2**	**1.000**	**0.006**	**0.000**	**-125.808**	**255.616**	**0.279**	**0.115**
FH	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-125.668**	**255.337**	**0.000**	**0.132**
FH	**AT**	**TH**	**2**	**1.000**	**0.114**	**0.000**	**-127.480**	**258.959**	**3.623**	**0.022**
LO	QL	TH	3	1.000	0.114	0.000	-127.480	260.960	5.623	0.008
SS	**AT**	**PT**	**2**	**1.000**	**0.006**	**0.000**	**-125.808**	**255.616**	**0.279**	**0.115**
LO	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-125.668**	**255.337**	**0.000**	**0.132**

Table A.17 Models summary: problem 2 - subject 4

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.002**	**0.284**	**-131.246**	**268.491**	**1.485**	**0.069**
FH	QL	TH	3	0.205	0.014	0.976	-132.640	271.280	4.274	0.017
SS	AT	TH	2	1.000	0.048	0.000	-135.432	274.864	7.858	0.003
SS	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-131.503**	**267.006**	**0.000**	**0.146**
LO	**AT**	**PT**	**2**	**1.000**	**0.004**	**0.000**	**-131.891**	**267.783**	**0.777**	**0.099**
LO	AT	TH	2	1.000	0.048	0.000	-135.432	274.864	7.858	0.003
LO	**QL**	**PT**	**3**	**1.000**	**0.004**	**0.216**	**-131.752**	**269.503**	**2.498**	**0.042**
FH	**QL**	**PT**	**3**	**1.000**	**0.004**	**0.216**	**-131.752**	**269.503**	**2.498**	**0.042**
LO	**QL**	**ID**	**3**	**1.000**	**0.002**	**0.284**	**-131.246**	**268.491**	**1.485**	**0.069**
FH	**AT**	**PT**	**2**	**1.000**	**0.004**	**0.000**	**-131.891**	**267.783**	**0.777**	**0.099**
FH	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-131.503**	**267.006**	**0.000**	**0.146**
FH	AT	TH	2	1.000	0.048	0.000	-135.432	274.864	7.858	0.003
LO	QL	TH	3	0.205	0.014	0.976	-132.640	271.280	4.274	0.017
SS	**AT**	**PT**	**2**	**1.000**	**0.004**	**0.000**	**-131.891**	**267.783**	**0.777**	**0.099**
LO	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-131.503**	**267.006**	**0.000**	**0.146**

Table A.18 Models summary: problem 2 - subject 5

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.004	0.007	0.700	-79.628	165.257	12.027	0.001
FH	**QL**	**TH**	**3**	**0.406**	**0.010**	**0.988**	**-73.895**	**153.790**	**0.560**	**0.212**
SS	AT	TH	2	0.001	0.700	0.000	-80.008	164.015	10.786	0.001
SS	AT	ID	2	0.000	0.100	0.000	-79.640	163.281	10.051	0.002
LO	AT	PT	2	0.001	0.100	0.000	-79.700	163.400	10.171	0.002
LO	AT	TH	2	0.001	0.700	0.000	-80.008	164.015	10.786	0.001
LO	**QL**	**PT**	**3**	**0.371**	**0.001**	**0.986**	**-73.615**	**153.230**	**0.000**	**0.280**
FH	**QL**	**PT**	**3**	**0.371**	**0.001**	**0.986**	**-73.615**	**153.230**	**0.000**	**0.280**
LO	QL	ID	3	0.004	0.007	0.700	-79.628	165.257	12.027	0.001
FH	AT	PT	2	0.001	0.100	0.000	-79.700	163.400	10.171	0.002
FH	AT	ID	2	0.000	0.100	0.000	-79.640	163.281	10.051	0.002
FH	AT	TH	2	0.001	0.700	0.000	-80.008	164.015	10.786	0.001
LO	**QL**	**TH**	**3**	**0.406**	**0.010**	**0.988**	**-73.895**	**153.790**	**0.560**	**0.212**
SS	AT	PT	2	0.001	0.100	0.000	-79.700	163.400	10.171	0.002
LO	AT	ID	2	0.000	0.100	0.000	-79.640	163.281	10.051	0.002

Table A.19 Models summary: problem 2 - subject 6

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.448	0.001	0.962	-110.715	227.430	59.470	0.000
FH	**QL**	**TH**	**3**	**1.000**	**0.005**	**1.000**	**-81.051**	**168.103**	**0.143**	**0.241**
SS	AT	TH	2	0.001	1.000	0.000	-113.413	230.826	62.867	0.000
SS	AT	ID	2	0.000	0.700	0.000	-113.454	230.907	62.948	0.000
LO	AT	PT	2	0.000	0.700	0.000	-113.390	230.779	62.820	0.000
LO	AT	TH	2	0.001	1.000	0.000	-113.413	230.826	62.867	0.000
LO	**QL**	**PT**	**3**	**1.000**	**0.000**	**1.000**	**-80.980**	**167.959**	**0.000**	**0.259**
FH	**QL**	**PT**	**3**	**1.000**	**0.000**	**1.000**	**-80.980**	**167.959**	**0.000**	**0.259**
LO	QL	ID	3	0.448	0.001	0.962	-110.715	227.430	59.470	0.000
FH	AT	PT	2	0.000	0.700	0.000	-113.390	230.779	62.820	0.000
FH	AT	ID	2	0.000	0.700	0.000	-113.454	230.907	62.948	0.000
FH	AT	TH	2	0.001	1.000	0.000	-113.413	230.826	62.867	0.000
LO	**QL**	**TH**	**3**	**1.000**	**0.005**	**1.000**	**-81.051**	**168.103**	**0.143**	**0.241**
SS	AT	PT	2	0.000	0.700	0.000	-113.390	230.779	62.820	0.000
LO	AT	ID	2	0.000	0.700	0.000	-113.454	230.907	62.948	0.000

Table A.20 Models summary: problem 2 - subject 7

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.018**	**0.003**	**0.000**	**-123.097**	**252.195**	**2.678**	**0.028**
FH	**QL**	**TH**	**3**	**0.021**	**0.087**	**0.000**	**-122.758**	**251.517**	**2.000**	**0.040**
SS	**AT**	**TH**	**2**	**0.021**	**0.087**	**0.000**	**-122.758**	**249.517**	**0.000**	**0.108**
SS	**AT**	**ID**	**2**	**0.018**	**0.003**	**0.000**	**-123.097**	**250.195**	**0.678**	**0.077**
LO	**AT**	**PT**	**2**	**0.019**	**0.008**	**0.000**	**-123.038**	**250.076**	**0.559**	**0.082**
LO	**AT**	**TH**	**2**	**0.021**	**0.087**	**0.000**	**-122.758**	**249.517**	**0.000**	**0.108**
LO	**QL**	**PT**	**3**	**0.019**	**0.008**	**0.000**	**-123.038**	**252.076**	**2.559**	**0.030**
FH	**QL**	**PT**	**3**	**0.019**	**0.008**	**0.000**	**-123.038**	**252.076**	**2.559**	**0.030**
LO	**QL**	**ID**	**3**	**0.018**	**0.003**	**0.000**	**-123.097**	**252.195**	**2.678**	**0.028**
FH	**AT**	**PT**	**2**	**0.019**	**0.008**	**0.000**	**-123.038**	**250.076**	**0.559**	**0.082**
FH	**AT**	**ID**	**2**	**0.018**	**0.003**	**0.000**	**-123.097**	**250.195**	**0.678**	**0.077**
FH	**AT**	**TH**	**2**	**0.021**	**0.087**	**0.000**	**-122.758**	**249.517**	**0.000**	**0.108**
LO	**QL**	**TH**	**3**	**0.021**	**0.087**	**0.000**	**-122.758**	**251.517**	**2.000**	**0.040**
SS	**AT**	**PT**	**2**	**0.019**	**0.008**	**0.000**	**-123.038**	**250.076**	**0.559**	**0.082**
LO	**AT**	**ID**	**2**	**0.018**	**0.003**	**0.000**	**-123.097**	**250.195**	**0.678**	**0.077**

Table A.21 Models summary: problem 2 - subject 8

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.033**	**0.010**	**0.000**	**-5.557**	**17.115**	**2.000**	**0.033**
FH	**QL**	**TH**	**3**	**0.032**	**0.295**	**0.000**	**-5.575**	**17.149**	**2.034**	**0.033**
SS	**AT**	**TH**	**2**	**0.032**	**0.295**	**0.000**	**-5.575**	**15.149**	**0.034**	**0.088**
SS	**AT**	**ID**	**2**	**0.033**	**0.010**	**0.000**	**-5.557**	**15.115**	**0.000**	**0.090**
LO	**AT**	**PT**	**2**	**0.033**	**0.023**	**0.000**	**-5.563**	**15.126**	**0.011**	**0.089**
LO	**AT**	**TH**	**2**	**0.032**	**0.295**	**0.000**	**-5.575**	**15.149**	**0.034**	**0.088**
LO	**QL**	**PT**	**3**	**0.033**	**0.023**	**0.000**	**-5.563**	**17.126**	**2.011**	**0.033**
FH	**QL**	**PT**	**3**	**0.033**	**0.023**	**0.000**	**-5.563**	**17.126**	**2.011**	**0.033**
LO	**QL**	**ID**	**3**	**0.033**	**0.010**	**0.000**	**-5.557**	**17.115**	**2.000**	**0.033**
FH	**AT**	**PT**	**2**	**0.033**	**0.023**	**0.000**	**-5.563**	**15.126**	**0.011**	**0.089**
FH	**AT**	**ID**	**2**	**0.033**	**0.010**	**0.000**	**-5.557**	**15.115**	**0.000**	**0.090**
FH	**AT**	**TH**	**2**	**0.032**	**0.295**	**0.000**	**-5.575**	**15.149**	**0.034**	**0.088**
LO	**QL**	**TH**	**3**	**0.032**	**0.295**	**0.000**	**-5.575**	**17.149**	**2.034**	**0.033**
SS	**AT**	**PT**	**2**	**0.033**	**0.023**	**0.000**	**-5.563**	**15.126**	**0.011**	**0.089**
LO	**AT**	**ID**	**2**	**0.033**	**0.010**	**0.000**	**-5.557**	**15.115**	**0.000**	**0.090**

Table A.22 Models summary: problem 2 - subject 9

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.002**	**0.000**	**-125.498**	**256.997**	**2.000**	**0.060**
FH	QL	TH	3	1.000	0.104	0.000	-130.575	267.151	12.154	0.000
SS	AT	TH	2	1.000	0.104	0.000	-130.575	265.150	10.154	0.001
SS	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-125.498**	**254.997**	**0.000**	**0.164**
LO	**AT**	**PT**	**2**	**1.000**	**0.006**	**0.000**	**-125.962**	**255.924**	**0.928**	**0.103**
LO	AT	TH	2	1.000	0.104	0.000	-130.575	265.150	10.154	0.001
LO	**QL**	**PT**	**3**	**1.000**	**0.006**	**0.000**	**-125.962**	**257.924**	**2.928**	**0.038**
FH	**QL**	**PT**	**3**	**1.000**	**0.006**	**0.000**	**-125.962**	**257.924**	**2.928**	**0.038**
LO	**QL**	**ID**	**3**	**1.000**	**0.002**	**0.000**	**-125.498**	**256.997**	**2.000**	**0.060**
FH	**AT**	**PT**	**2**	**1.000**	**0.006**	**0.000**	**-125.962**	**255.924**	**0.928**	**0.103**
FH	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-125.498**	**254.997**	**0.000**	**0.164**
FH	AT	TH	2	1.000	0.104	0.000	-130.575	265.150	10.154	0.001
LO	QL	TH	3	1.000	0.104	0.000	-130.575	267.151	12.154	0.000
SS	**AT**	**PT**	**2**	**1.000**	**0.006**	**0.000**	**-125.962**	**255.924**	**0.928**	**0.103**
LO	**AT**	**ID**	**2**	**1.000**	**0.002**	**0.000**	**-125.498**	**254.997**	**0.000**	**0.164**

Table A.23 Models summary: problem 2 - subject 10

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**1.000**	**0.000**	**0.995**	**-60.749**	**127.498**	**0.000**	**0.177**
FH	**QL**	**TH**	**3**	**1.000**	**0.005**	**0.994**	**-60.886**	**127.771**	**0.273**	**0.154**
SS	AT	TH	2	0.064	0.077	0.000	-133.507	271.015	143.517	0.000
SS	AT	ID	2	0.072	0.006	0.000	-121.284	246.567	119.069	0.000
LO	AT	PT	2	0.075	0.013	0.000	-123.839	251.678	124.180	0.000
LO	AT	TH	2	0.064	0.077	0.000	-133.507	271.015	143.517	0.000
LO	**QL**	**PT**	**3**	**1.000**	**0.000**	**0.994**	**-60.790**	**127.580**	**0.082**	**0.169**
FH	**QL**	**PT**	**3**	**1.000**	**0.000**	**0.994**	**-60.790**	**127.580**	**0.082**	**0.169**
LO	**QL**	**ID**	**3**	**1.000**	**0.000**	**0.995**	**-60.749**	**127.498**	**0.000**	**0.177**
FH	AT	PT	2	0.075	0.013	0.000	-123.839	251.678	124.180	0.000
FH	AT	ID	2	0.072	0.006	0.000	-121.284	246.567	119.069	0.000
FH	AT	TH	2	0.064	0.077	0.000	-133.507	271.015	143.517	0.000
LO	**QL**	**TH**	**3**	**1.000**	**0.005**	**0.994**	**-60.886**	**127.771**	**0.273**	**0.154**
SS	AT	PT	2	0.075	0.013	0.000	-123.839	251.678	124.180	0.000
LO	AT	ID	2	0.072	0.006	0.000	-121.284	246.567	119.069	0.000

Table A.24 Models summary: problem 2 - subject 11

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	1.000	0.002	0.537	-127.270	260.540	26.593	0.000
FH	**QL**	**TH**	**3**	**1.000**	**0.004**	**1.000**	**-114.168**	**234.335**	**0.389**	**0.226**
SS	AT	TH	2	0.935	0.066	0.000	-134.573	273.146	39.199	0.000
SS	AT	ID	2	0.554	0.002	0.000	-130.269	264.537	30.591	0.000
LO	AT	PT	2	0.574	0.006	0.000	-130.696	265.391	31.444	0.000
LO	AT	TH	2	0.935	0.066	0.000	-134.573	273.146	39.199	0.000
LO	**QL**	**PT**	**3**	**1.000**	**0.000**	**0.999**	**-113.973**	**233.947**	**0.000**	**0.274**
FH	**QL**	**PT**	**3**	**1.000**	**0.000**	**0.999**	**-113.973**	**233.947**	**0.000**	**0.274**
LO	QL	ID	3	1.000	0.002	0.537	-127.270	260.540	26.593	0.000
FH	AT	PT	2	0.574	0.006	0.000	-130.696	265.391	31.444	0.000
FH	AT	ID	2	0.554	0.002	0.000	-130.269	264.537	30.591	0.000
FH	AT	TH	2	0.935	0.066	0.000	-134.573	273.146	39.199	0.000
LO	**QL**	**TH**	**3**	**1.000**	**0.004**	**1.000**	**-114.168**	**234.335**	**0.389**	**0.226**
SS	AT	PT	2	0.574	0.006	0.000	-130.696	265.391	31.444	0.000
LO	AT	ID	2	0.554	0.002	0.000	-130.269	264.537	30.591	0.000

Table A.25 Models summary: problem 2 - subject 12

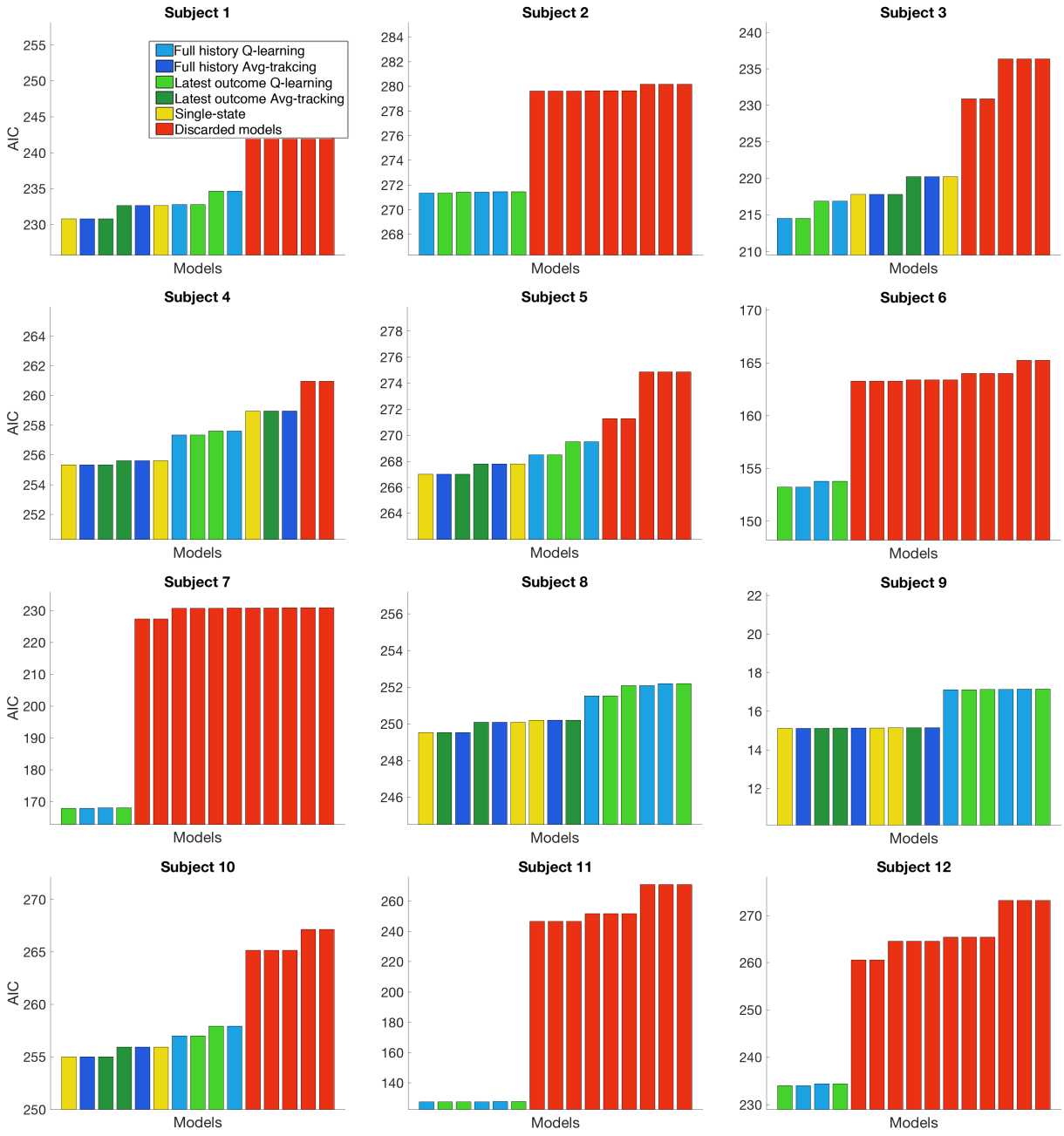

Fig. A.2 Comparison of AIC scores of the 15 models fitted to the subjects in condition 2. Charts and legend as in Fig. A.1.

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.022**	**0.004**	**0.000**	**-80.679**	**167.358**	**2.000**	**0.035**
FH	**QL**	**TH**	**3**	**0.022**	**0.125**	**0.000**	**-80.874**	**167.748**	**2.389**	**0.029**
SS	**AT**	**TH**	**2**	**0.022**	**0.125**	**0.000**	**-80.874**	**165.748**	**0.389**	**0.079**
SS	**AT**	**ID**	**2**	**0.022**	**0.004**	**0.000**	**-80.679**	**165.358**	**0.000**	**0.096**
LO	**AT**	**PT**	**2**	**0.022**	**0.010**	**0.000**	**-80.716**	**165.432**	**0.074**	**0.093**
LO	**AT**	**TH**	**2**	**0.022**	**0.125**	**0.000**	**-80.874**	**165.748**	**0.389**	**0.079**
LO	**QL**	**PT**	**3**	**0.022**	**0.010**	**0.000**	**-80.716**	**167.432**	**2.074**	**0.034**
FH	**QL**	**PT**	**3**	**0.022**	**0.010**	**0.000**	**-80.716**	**167.432**	**2.074**	**0.034**
LO	**QL**	**ID**	**3**	**0.022**	**0.004**	**0.000**	**-80.679**	**167.358**	**2.000**	**0.035**
FH	**AT**	**PT**	**2**	**0.022**	**0.010**	**0.000**	**-80.716**	**165.432**	**0.074**	**0.093**
FH	**AT**	**ID**	**2**	**0.022**	**0.004**	**0.000**	**-80.679**	**165.358**	**0.000**	**0.096**
FH	**AT**	**TH**	**2**	**0.022**	**0.125**	**0.000**	**-80.874**	**165.748**	**0.389**	**0.079**
LO	**QL**	**TH**	**3**	**0.022**	**0.125**	**0.000**	**-80.874**	**167.748**	**2.389**	**0.029**
SS	**AT**	**PT**	**2**	**0.022**	**0.010**	**0.000**	**-80.716**	**165.432**	**0.074**	**0.093**
LO	**AT**	**ID**	**2**	**0.022**	**0.004**	**0.000**	**-80.679**	**165.358**	**0.000**	**0.096**

Table A.26 Models summary: problem 3 - subject 1

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.275**	**0.000**	**0.980**	**-79.116**	**164.233**	**0.000**	**0.185**
FH	**QL**	**TH**	**3**	**0.276**	**0.011**	**0.982**	**-79.418**	**164.837**	**0.604**	**0.137**
SS	AT	TH	2	0.004	0.264	0.000	-98.975	201.949	37.717	0.000
SS	AT	ID	2	0.981	0.020	0.000	-84.201	172.401	8.168	0.003
LO	AT	PT	2	0.989	0.049	0.000	-87.473	178.946	14.713	0.000
LO	AT	TH	2	0.004	0.264	0.000	-98.975	201.949	37.717	0.000
LO	**QL**	**PT**	**3**	**0.275**	**0.001**	**0.980**	**-79.177**	**164.354**	**0.121**	**0.174**
FH	**QL**	**PT**	**3**	**0.275**	**0.001**	**0.980**	**-79.177**	**164.354**	**0.121**	**0.174**
LO	**QL**	**ID**	**3**	**0.275**	**0.000**	**0.980**	**-79.116**	**164.233**	**0.000**	**0.185**
FH	AT	PT	2	0.989	0.049	0.000	-87.473	178.946	14.713	0.000
FH	AT	ID	2	0.981	0.020	0.000	-84.201	172.401	8.168	0.003
FH	AT	TH	2	0.004	0.264	0.000	-98.975	201.949	37.717	0.000
LO	**QL**	**TH**	**3**	**0.276**	**0.011**	**0.982**	**-79.418**	**164.837**	**0.604**	**0.137**
SS	AT	PT	2	0.989	0.049	0.000	-87.473	178.946	14.713	0.000
LO	AT	ID	2	0.981	0.020	0.000	-84.201	172.401	8.168	0.003

Table A.27 Models summary: problem 3 - subject 2

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	QL	ID	3	0.131	0.001	0.788	-87.289	180.579	13.561	0.001
FH	**QL**	**TH**	**3**	**0.631**	**0.005**	**1.000**	**-80.509**	**167.018**	**0.000**	**0.499**
SS	AT	TH	2	0.029	0.122	0.000	-93.694	191.387	24.369	0.000
SS	AT	ID	2	0.033	0.004	0.000	-92.111	188.223	21.205	0.000
LO	AT	PT	2	0.032	0.010	0.000	-92.436	188.872	21.854	0.000
LO	AT	TH	2	0.029	0.122	0.000	-93.694	191.387	24.369	0.000
LO	QL	PT	3	0.130	0.003	0.795	-87.460	180.919	13.901	0.000
FH	QL	PT	3	0.130	0.003	0.795	-87.460	180.919	13.901	0.000
LO	QL	ID	3	0.131	0.001	0.788	-87.289	180.579	13.561	0.001
FH	AT	PT	2	0.032	0.010	0.000	-92.436	188.872	21.854	0.000
FH	AT	ID	2	0.033	0.004	0.000	-92.111	188.223	21.205	0.000
FH	AT	TH	2	0.029	0.122	0.000	-93.694	191.387	24.369	0.000
LO	**QL**	**TH**	**3**	**0.631**	**0.005**	**1.000**	**-80.509**	**167.018**	**0.000**	**0.499**
SS	AT	PT	2	0.032	0.010	0.000	-92.436	188.872	21.854	0.000
LO	AT	ID	2	0.033	0.004	0.000	-92.111	188.223	21.205	0.000

Table A.28 Models summary: problem 3 - subject 3

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.730**	**0.000**	**1.000**	**-92.384**	**190.769**	**0.000**	**0.174**
FH	**QL**	**TH**	**3**	**0.737**	**0.006**	**1.000**	**-92.489**	**190.978**	**0.209**	**0.156**
SS	AT	TH	2	0.001	1.000	0.000	-114.626	233.252	42.484	0.000
SS	AT	ID	2	0.000	0.700	0.000	-111.480	226.960	36.191	0.000
LO	AT	PT	2	0.000	0.700	0.000	-112.002	228.004	37.235	0.000
LO	AT	TH	2	0.001	1.000	0.000	-114.626	233.252	42.484	0.000
LO	**QL**	**PT**	**3**	**0.731**	**0.000**	**1.000**	**-92.405**	**190.810**	**0.041**	**0.170**
FH	**QL**	**PT**	**3**	**0.731**	**0.000**	**1.000**	**-92.405**	**190.810**	**0.041**	**0.170**
LO	**QL**	**ID**	**3**	**0.730**	**0.000**	**1.000**	**-92.384**	**190.769**	**0.000**	**0.174**
FH	AT	PT	2	0.000	0.700	0.000	-112.002	228.004	37.235	0.000
FH	AT	ID	2	0.000	0.700	0.000	-111.480	226.960	36.191	0.000
FH	AT	TH	2	0.001	1.000	0.000	-114.626	233.252	42.484	0.000
LO	**QL**	**TH**	**3**	**0.737**	**0.006**	**1.000**	**-92.489**	**190.978**	**0.209**	**0.156**
SS	AT	PT	2	0.000	0.700	0.000	-112.002	228.004	37.235	0.000
LO	AT	ID	2	0.000	0.700	0.000	-111.480	226.960	36.191	0.000

Table A.29 Models summary: problem 3 - subject 4

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.102**	**0.003**	**0.462**	**-21.770**	**49.541**	**1.890**	**0.037**
FH	**QL**	**TH**	**3**	**0.127**	**0.072**	**0.584**	**-21.913**	**49.827**	**2.176**	**0.032**
SS	**AT**	**TH**	**2**	**0.055**	**0.163**	**0.000**	**-22.005**	**48.010**	**0.359**	**0.079**
SS	**AT**	**ID**	**2**	**0.056**	**0.005**	**0.000**	**-21.825**	**47.651**	**0.000**	**0.094**
LO	**AT**	**PT**	**2**	**0.056**	**0.013**	**0.000**	**-21.861**	**47.721**	**0.071**	**0.091**
LO	**AT**	**TH**	**2**	**0.055**	**0.163**	**0.000**	**-22.005**	**48.010**	**0.359**	**0.079**
LO	**QL**	**PT**	**3**	**0.114**	**0.006**	**0.522**	**-21.797**	**49.594**	**1.944**	**0.036**
FH	**QL**	**PT**	**3**	**0.114**	**0.006**	**0.522**	**-21.797**	**49.594**	**1.944**	**0.036**
LO	**QL**	**ID**	**3**	**0.102**	**0.003**	**0.462**	**-21.770**	**49.541**	**1.890**	**0.037**
FH	**AT**	**PT**	**2**	**0.056**	**0.013**	**0.000**	**-21.861**	**47.721**	**0.071**	**0.091**
FH	**AT**	**ID**	**2**	**0.056**	**0.005**	**0.000**	**-21.825**	**47.651**	**0.000**	**0.094**
FH	**AT**	**TH**	**2**	**0.055**	**0.163**	**0.000**	**-22.005**	**48.010**	**0.359**	**0.079**
LO	**QL**	**TH**	**3**	**0.127**	**0.072**	**0.584**	**-21.913**	**49.827**	**2.176**	**0.032**
SS	**AT**	**PT**	**2**	**0.056**	**0.013**	**0.000**	**-21.861**	**47.721**	**0.071**	**0.091**
LO	**AT**	**ID**	**2**	**0.056**	**0.005**	**0.000**	**-21.825**	**47.651**	**0.000**	**0.094**

Table A.30 Models summary: problem 3 - subject 5

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.376**	**0.000**	**1.000**	**-46.050**	**98.099**	**0.000**	**0.130**
FH	**QL**	**TH**	**3**	**0.404**	**0.013**	**1.000**	**-46.131**	**98.262**	**0.162**	**0.120**
SS	AT	TH	2	0.004	1.000	0.000	-49.277	102.554	4.454	0.014
SS	**AT**	**ID**	**2**	**0.000**	**0.700**	**0.000**	**-48.336**	**100.672**	**2.573**	**0.036**
LO	**AT**	**PT**	**2**	**0.000**	**1.000**	**0.000**	**-48.440**	**100.880**	**2.780**	**0.032**
LO	AT	TH	2	0.004	1.000	0.000	-49.277	102.554	4.454	0.014
LO	**QL**	**PT**	**3**	**0.376**	**0.001**	**1.000**	**-46.067**	**98.134**	**0.035**	**0.127**
FH	**QL**	**PT**	**3**	**0.376**	**0.001**	**1.000**	**-46.067**	**98.134**	**0.035**	**0.127**
LO	**QL**	**ID**	**3**	**0.376**	**0.000**	**1.000**	**-46.050**	**98.099**	**0.000**	**0.130**
FH	**AT**	**PT**	**2**	**0.000**	**1.000**	**0.000**	**-48.440**	**100.880**	**2.780**	**0.032**
FH	**AT**	**ID**	**2**	**0.000**	**0.700**	**0.000**	**-48.336**	**100.672**	**2.573**	**0.036**
FH	AT	TH	2	0.004	1.000	0.000	-49.277	102.554	4.454	0.014
LO	**QL**	**TH**	**3**	**0.404**	**0.013**	**1.000**	**-46.131**	**98.262**	**0.162**	**0.120**
SS	**AT**	**PT**	**2**	**0.000**	**1.000**	**0.000**	**-48.440**	**100.880**	**2.780**	**0.032**
LO	**AT**	**ID**	**2**	**0.000**	**0.700**	**0.000**	**-48.336**	**100.672**	**2.573**	**0.036**

Table A.31 Models summary: problem 3 - subject 6

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.020**	**0.004**	**0.000**	**-82.101**	**170.203**	**2.000**	**0.034**
FH	**QL**	**TH**	**3**	**0.020**	**0.112**	**0.000**	**-82.230**	**170.460**	**2.257**	**0.030**
SS	**AT**	**TH**	**2**	**0.020**	**0.112**	**0.000**	**-82.230**	**168.460**	**0.257**	**0.082**
SS	**AT**	**ID**	**2**	**0.020**	**0.004**	**0.000**	**-82.101**	**168.203**	**0.000**	**0.094**
LO	**AT**	**PT**	**2**	**0.020**	**0.009**	**0.000**	**-82.126**	**168.253**	**0.050**	**0.091**
LO	**AT**	**TH**	**2**	**0.020**	**0.112**	**0.000**	**-82.230**	**168.460**	**0.257**	**0.082**
LO	**QL**	**PT**	**3**	**0.020**	**0.009**	**0.000**	**-82.126**	**170.253**	**2.050**	**0.034**
FH	**QL**	**PT**	**3**	**0.020**	**0.009**	**0.000**	**-82.126**	**170.253**	**2.050**	**0.034**
LO	**QL**	**ID**	**3**	**0.020**	**0.004**	**0.000**	**-82.101**	**170.203**	**2.000**	**0.034**
FH	**AT**	**PT**	**2**	**0.020**	**0.009**	**0.000**	**-82.126**	**168.253**	**0.050**	**0.091**
FH	**AT**	**ID**	**2**	**0.020**	**0.004**	**0.000**	**-82.101**	**168.203**	**0.000**	**0.094**
FH	**AT**	**TH**	**2**	**0.020**	**0.112**	**0.000**	**-82.230**	**168.460**	**0.257**	**0.082**
LO	**QL**	**TH**	**3**	**0.020**	**0.112**	**0.000**	**-82.230**	**170.460**	**2.257**	**0.030**
SS	**AT**	**PT**	**2**	**0.020**	**0.009**	**0.000**	**-82.126**	**168.253**	**0.050**	**0.091**
LO	**AT**	**ID**	**2**	**0.020**	**0.004**	**0.000**	**-82.101**	**168.203**	**0.000**	**0.094**

Table A.32 Models summary: problem 3 - subject 7

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.032**	**0.004**	**0.000**	**-36.863**	**79.726**	**2.000**	**0.033**
FH	**QL**	**TH**	**3**	**0.032**	**0.123**	**0.000**	**-36.902**	**79.803**	**2.078**	**0.032**
SS	**AT**	**TH**	**2**	**0.032**	**0.123**	**0.000**	**-36.902**	**77.803**	**0.078**	**0.087**
SS	**AT**	**ID**	**2**	**0.032**	**0.004**	**0.000**	**-36.863**	**77.726**	**0.000**	**0.091**
LO	**AT**	**PT**	**2**	**0.032**	**0.010**	**0.000**	**-36.871**	**77.741**	**0.015**	**0.090**
LO	**AT**	**TH**	**2**	**0.032**	**0.123**	**0.000**	**-36.902**	**77.803**	**0.078**	**0.087**
LO	**QL**	**PT**	**3**	**0.032**	**0.010**	**0.000**	**-36.871**	**79.741**	**2.015**	**0.033**
FH	**QL**	**PT**	**3**	**0.032**	**0.010**	**0.000**	**-36.871**	**79.741**	**2.015**	**0.033**
LO	**QL**	**ID**	**3**	**0.032**	**0.004**	**0.000**	**-36.863**	**79.726**	**2.000**	**0.033**
FH	**AT**	**PT**	**2**	**0.032**	**0.010**	**0.000**	**-36.871**	**77.741**	**0.015**	**0.090**
FH	**AT**	**ID**	**2**	**0.032**	**0.004**	**0.000**	**-36.863**	**77.726**	**0.000**	**0.091**
FH	**AT**	**TH**	**2**	**0.032**	**0.123**	**0.000**	**-36.902**	**77.803**	**0.078**	**0.087**
LO	**QL**	**TH**	**3**	**0.032**	**0.123**	**0.000**	**-36.902**	**79.803**	**2.078**	**0.032**
SS	**AT**	**PT**	**2**	**0.032**	**0.010**	**0.000**	**-36.871**	**77.741**	**0.015**	**0.090**
LO	**AT**	**ID**	**2**	**0.032**	**0.004**	**0.000**	**-36.863**	**77.726**	**0.000**	**0.091**

Table A.33 Models summary: problem 3 - subject 8

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.018**	**0.006**	**0.000**	**-39.635**	**85.269**	**2.000**	**0.034**
FH	**QL**	**TH**	**3**	**0.019**	**0.176**	**0.000**	**-39.688**	**85.376**	**2.107**	**0.032**
SS	**AT**	**TH**	**2**	**0.019**	**0.176**	**0.000**	**-39.688**	**83.376**	**0.107**	**0.086**
SS	**AT**	**ID**	**2**	**0.018**	**0.006**	**0.000**	**-39.635**	**83.269**	**0.000**	**0.091**
LO	**AT**	**PT**	**2**	**0.018**	**0.014**	**0.000**	**-39.645**	**83.290**	**0.021**	**0.090**
LO	**AT**	**TH**	**2**	**0.019**	**0.176**	**0.000**	**-39.688**	**83.376**	**0.107**	**0.086**
LO	**QL**	**PT**	**3**	**0.018**	**0.014**	**0.000**	**-39.645**	**85.290**	**2.021**	**0.033**
FH	**QL**	**PT**	**3**	**0.018**	**0.014**	**0.000**	**-39.645**	**85.290**	**2.021**	**0.033**
LO	**QL**	**ID**	**3**	**0.018**	**0.006**	**0.000**	**-39.635**	**85.269**	**2.000**	**0.034**
FH	**AT**	**PT**	**2**	**0.018**	**0.014**	**0.000**	**-39.645**	**83.290**	**0.021**	**0.090**
FH	**AT**	**ID**	**2**	**0.018**	**0.006**	**0.000**	**-39.635**	**83.269**	**0.000**	**0.091**
FH	**AT**	**TH**	**2**	**0.019**	**0.176**	**0.000**	**-39.688**	**83.376**	**0.107**	**0.086**
LO	**QL**	**TH**	**3**	**0.019**	**0.176**	**0.000**	**-39.688**	**85.376**	**2.107**	**0.032**
SS	**AT**	**PT**	**2**	**0.018**	**0.014**	**0.000**	**-39.645**	**83.290**	**0.021**	**0.090**
LO	**AT**	**ID**	**2**	**0.018**	**0.006**	**0.000**	**-39.635**	**83.269**	**0.000**	**0.091**

Table A.34 Models summary: problem 3 - subject 9

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.030**	**0.004**	**0.000**	**-88.942**	**183.884**	**2.000**	**0.036**
FH	**QL**	**TH**	**3**	**0.029**	**0.133**	**0.000**	**-89.217**	**184.434**	**2.550**	**0.028**
SS	**AT**	**TH**	**2**	**0.029**	**0.133**	**0.000**	**-89.217**	**182.434**	**0.550**	**0.075**
SS	**AT**	**ID**	**2**	**0.030**	**0.004**	**0.000**	**-88.942**	**181.884**	**0.000**	**0.099**
LO	**AT**	**PT**	**2**	**0.030**	**0.010**	**0.000**	**-88.993**	**181.986**	**0.101**	**0.094**
LO	**AT**	**TH**	**2**	**0.029**	**0.133**	**0.000**	**-89.217**	**182.434**	**0.550**	**0.075**
LO	**QL**	**PT**	**3**	**0.030**	**0.010**	**0.000**	**-88.993**	**183.986**	**2.101**	**0.035**
FH	**QL**	**PT**	**3**	**0.030**	**0.010**	**0.000**	**-88.993**	**183.986**	**2.101**	**0.035**
LO	**QL**	**ID**	**3**	**0.030**	**0.004**	**0.000**	**-88.942**	**183.884**	**2.000**	**0.036**
FH	**AT**	**PT**	**2**	**0.030**	**0.010**	**0.000**	**-88.993**	**181.986**	**0.101**	**0.094**
FH	**AT**	**ID**	**2**	**0.030**	**0.004**	**0.000**	**-88.942**	**181.884**	**0.000**	**0.099**
FH	**AT**	**TH**	**2**	**0.029**	**0.133**	**0.000**	**-89.217**	**182.434**	**0.550**	**0.075**
LO	**QL**	**TH**	**3**	**0.029**	**0.133**	**0.000**	**-89.217**	**184.434**	**2.550**	**0.028**
SS	**AT**	**PT**	**2**	**0.030**	**0.010**	**0.000**	**-88.993**	**181.986**	**0.101**	**0.094**
LO	**AT**	**ID**	**2**	**0.030**	**0.004**	**0.000**	**-88.942**	**181.884**	**0.000**	**0.099**

Table A.35 Models summary: problem 3 - subject 10

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.061**	**0.003**	**0.522**	**-59.640**	**125.279**	**1.688**	**0.045**
FH	**QL**	**TH**	**3**	**0.075**	**0.075**	**0.627**	**-60.049**	**126.098**	**2.507**	**0.030**
SS	**AT**	**TH**	**2**	**0.029**	**0.178**	**0.000**	**-60.397**	**124.794**	**1.203**	**0.057**
SS	**AT**	**ID**	**2**	**0.030**	**0.006**	**0.000**	**-59.795**	**123.591**	**0.000**	**0.105**
LO	**AT**	**PT**	**2**	**0.030**	**0.014**	**0.000**	**-59.908**	**123.817**	**0.226**	**0.094**
LO	**AT**	**TH**	**2**	**0.029**	**0.178**	**0.000**	**-60.397**	**124.794**	**1.203**	**0.057**
LO	**QL**	**PT**	**3**	**0.064**	**0.007**	**0.550**	**-59.720**	**125.441**	**1.850**	**0.042**
FH	**QL**	**PT**	**3**	**0.064**	**0.007**	**0.550**	**-59.720**	**125.441**	**1.850**	**0.042**
LO	**QL**	**ID**	**3**	**0.061**	**0.003**	**0.522**	**-59.640**	**125.279**	**1.688**	**0.045**
FH	**AT**	**PT**	**2**	**0.030**	**0.014**	**0.000**	**-59.908**	**123.817**	**0.226**	**0.094**
FH	**AT**	**ID**	**2**	**0.030**	**0.006**	**0.000**	**-59.795**	**123.591**	**0.000**	**0.105**
FH	**AT**	**TH**	**2**	**0.029**	**0.178**	**0.000**	**-60.397**	**124.794**	**1.203**	**0.057**
LO	**QL**	**TH**	**3**	**0.075**	**0.075**	**0.627**	**-60.049**	**126.098**	**2.507**	**0.030**
SS	**AT**	**PT**	**2**	**0.030**	**0.014**	**0.000**	**-59.908**	**123.817**	**0.226**	**0.094**
LO	**AT**	**ID**	**2**	**0.030**	**0.006**	**0.000**	**-59.795**	**123.591**	**0.000**	**0.105**

Table A.36 Models summary: problem 3 - subject 11

Config			d.o.f.$_i$	α_i	β_i	γ_i	$\log(L_i)$	AIC$_i$	Δ_i(AIC)	w_i(AIC)
FH	**QL**	**ID**	**3**	**0.047**	**0.003**	**0.000**	**-63.226**	**132.452**	**2.000**	**0.036**
FH	**QL**	**TH**	**3**	**0.045**	**0.106**	**0.000**	**-63.469**	**132.938**	**2.486**	**0.028**
SS	**AT**	**TH**	**2**	**0.045**	**0.106**	**0.000**	**-63.469**	**130.938**	**0.486**	**0.077**
SS	**AT**	**ID**	**2**	**0.047**	**0.003**	**0.000**	**-63.226**	**130.452**	**0.000**	**0.098**
LO	**AT**	**PT**	**2**	**0.047**	**0.008**	**0.000**	**-63.274**	**130.549**	**0.096**	**0.093**
LO	**AT**	**TH**	**2**	**0.045**	**0.106**	**0.000**	**-63.469**	**130.938**	**0.486**	**0.077**
LO	**QL**	**PT**	**3**	**0.047**	**0.008**	**0.000**	**-63.274**	**132.549**	**2.096**	**0.034**
FH	**QL**	**PT**	**3**	**0.047**	**0.008**	**0.000**	**-63.274**	**132.549**	**2.096**	**0.034**
LO	**QL**	**ID**	**3**	**0.047**	**0.003**	**0.000**	**-63.226**	**132.452**	**2.000**	**0.036**
FH	**AT**	**PT**	**2**	**0.047**	**0.008**	**0.000**	**-63.274**	**130.549**	**0.096**	**0.093**
FH	**AT**	**ID**	**2**	**0.047**	**0.003**	**0.000**	**-63.226**	**130.452**	**0.000**	**0.098**
FH	**AT**	**TH**	**2**	**0.045**	**0.106**	**0.000**	**-63.469**	**130.938**	**0.486**	**0.077**
LO	**QL**	**TH**	**3**	**0.045**	**0.106**	**0.000**	**-63.469**	**132.938**	**2.486**	**0.028**
SS	**AT**	**PT**	**2**	**0.047**	**0.008**	**0.000**	**-63.274**	**130.549**	**0.096**	**0.093**
LO	**AT**	**ID**	**2**	**0.047**	**0.003**	**0.000**	**-63.226**	**130.452**	**0.000**	**0.098**

Table A.37 Models summary: problem 3 - subject 12

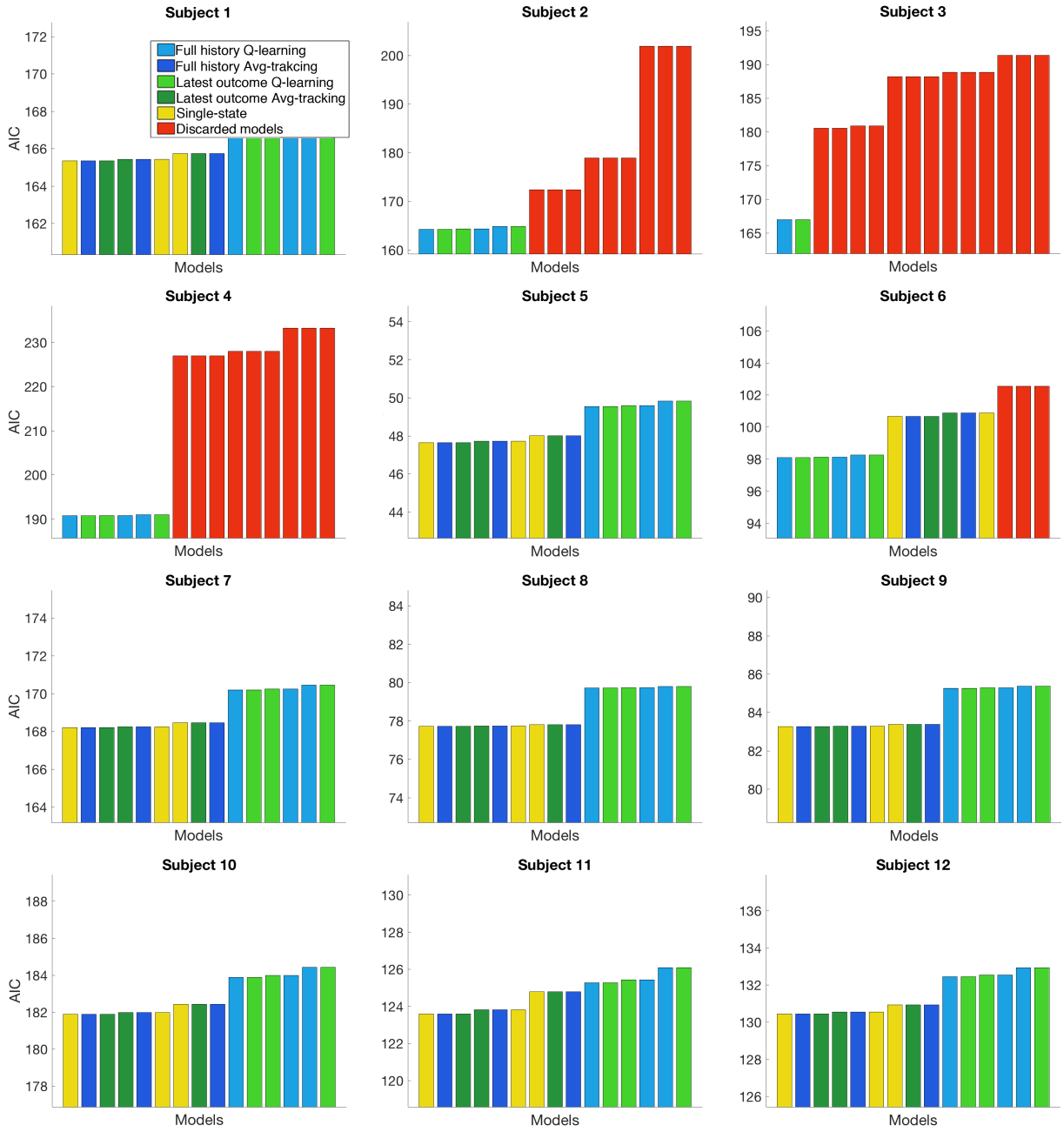

Fig. A.3 Comparison of AIC scores of the 15 models fitted to the subjects in condition 3. Charts and legend as in Fig. A.1.