Understanding Human Choices as Computationally Rational Processes

Haiyang Chen

School of Computer Science

College of Engineering and Physical Sciences

University of Birmingham

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

March 2022

What I cannot create, I do not understand. – Richard P. Feynman

Abstract

Risky choice involves deciding between gambles that can differ in the probability and value of outcomes. This thesis exposes the cognitive processes that underpin risky choice in humans. The approach taken involves the use of deep neural networks and reinforcement learning to discover policies that are adaptive to distributions of risky choice problems. Risky choice has been extensively studied for hundreds of years and in the modern era many phenomena have been reported. Sometimes these phenomena are explained away as "irrationalities" or "biases". This thesis uses computational methods to demonstrate that apparently irrational risky choice can be ecological rational and sometimes rational given cognitive bounds. Moreover it does so for a broader range of risky choice problems than has so far been investigated. These include both contextual choice problems and the fourfold pattern of risky choice. The implications for future work are discussed.

The results show that (1) context effects (including attraction, compromise and similarity) can emerge from an optimal (rational) "classifier" that chooses the option with the highest expected value; (2) the new model could predict context effects, as for people, when the representation format encourages attribute comparisons; (3) the new model approximates a bounded optimal cognitive policy and makes quantitative predictions that correspond well to evidence about human contextual choice; (4) an alternative explanation that a wide range of risky choice phenomena emerge from boundedly optimal adaptation of a decision making agent to processing constraints. In each study, the model is not pre-programmed to process all information but learns to process only that information that helps it maximize utility. We argue that the models provide evidence that apparently irrational risky choices are emergent consequences of processes that prefer higher value (rational) policies or classifiers.

My thesis is that a number of models offer novel and rational explanations for a broad range of phenomena exhibited by people making choice under risk. I demonstrate that apparent cognitive biases can emerge from computational rational processing. Furthermore, I propose a unifying framework for modelling risky choice phenomena. Deep reinforcement learning has the potential to help discriminate between various explanations because it provides a means of computing computationally rational policies given both ecological and cognitive bounds.

Acknowledgements

I am eternally grateful to have had the opportunity to pursue my academic interests and passions through the great journey of this Ph.D. It would not have been possible without the help, support, and collaboration of those generous and bright people around me.

First and foremost, I am most grateful to my supervisor, Prof. Andrew Howes, who believes in me to explore and do independent research while consistently making insightful suggestions, leaving thoughtful comments, cheering me on, and lending me a helping hand along the way. I would have been completely lost without his guidance. His knowledge, laser focus on work, and way of thinking have made a lasting impression on me. Moreover, he is also like a friend to me and supports me in pursuing a colourful life outside of research. His encouragement, passion, and optimism guided me through the dark hours of life after the COVID-19 pandemic. I am more than lucky to have had him as my supervisor and could not have hoped for a better one.

I would like to thank Richard L. Lewis and Xiuli Chen for helping develop these ideas and influencing me on many occasions during my study. I would also like to thank Antti Oulasvirta, Jussi P. P. Jokinen, Aditya Acharya, and Logan Walls for conducting the conversations that directly influenced parts of my thesis projects.

My gratitude extends to Hyung Jin Chang and Jeremy Wyatt for co-supervising me and giving me advice on writing, presentation, and professional development. I would like to thank my excellent thesis group committee, Ata Kaban, Hyung Jin Chang, and Mohan Sridharan, who have given valuable comments and advice in every thesis group meeting.

I would also like to thank the examiners of my thesis, Howard Bowman and Bradley C. Love and the chairman of my PhD viva, Christopher Baber. Many thanks go to Howard for reading through my thesis so meticulously and providing the exact marked-up copy. Thanks to Brad for giving valuable feedback and suggestions for future work.

I thank the School of Computer Science for providing a scholarship for my Ph.D programme and the University of Birmingham for providing a perfect environment for research. I especially thank Sarah Brookes for her help with all the administrative work.

And finally, last but by no means least, I want to dedicate this thesis to my great parents, Xinglan Hu and Fajiang Chen, in appreciation of their nurturing, love, and unconditional support. In particular, I want to express my gratitude to Tiantian Zhu, my beloved girlfriend, who brought happiness and passion to my life. I appreciate her two years with the company in Birmingham and Beijing and her constant spiritual support.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Overview

Deciphering risky decision making is a central goal for the science of human psychology. The goal concerns understanding the processes by which people choose between two or more options, each of which has an uncertain outcome. In risky choice problems, options are often defined in terms of multiple attributes and minimally in terms of a gain and a probability. Making a choice is then difficult when none of the available options dominate on all attributes; rather, one option might have a relatively large gain, while the other has a relatively large probability. In the development of scientific theories of these processes, it has often proved productive to compare observed human behaviour to the expected gain that would be made by an agent that always made the optimal choice given the available information. Accordingly, over the last 50 years, people have been observed to exhibit a range of behaviours that appear to depart from this ideal agent. These include, for example, a tendency to risk aversion, meaning the tendency to prefer safer smaller gains over riskier large gains. Also, a tendency to let irrelevant contextual factors affect decision making.

Inevitably, given the richness of the phenomena, a great number of theories have been posited to describe and explain the complex pattern of phenomena. One of the earliest theories, for example, explained risk aversion as a consequence of attempts to maximize utility rather than maximize financial gain. Under this theory, very large gains suffer diminishing returns because of the diminishing utility of more money. Further advances, including subjective expected utility theory (Savage, 1972), built on these insights, showing how utility could be maximized under risk. Such utility maximizing theories are often known as rational theories (Colman, 2003). At the same time, other theories were developed in response to apparent inconsistencies in the choices made by people. Unlike Savage's approach, the "heuristics and biases" approach of Kahneman and Tversky (1979), for example, does not assume utility maximization but rather seeks a descriptive account of the outcomes of human decision making.

One of the properties of the descriptive theories – but not the rational theories – is that they build in irrationalities. In other words, they are based on the assumption that people sometimes make choices against their own self interest. While some researchers have promoted the idea that people are biased, others have pointed to an explanatory gap. This gap has recently started to be populated by a new generation of utility maximizing theories (Howes et al., 2009, 2016; Lieder and Griffiths, 2019; Tsetsos et al., 2016). A key feature of these theories is that they explain human decision making behaviour as a rational consequence of the limited mental information gathering processes that precede choice.

The extension of these modern rational theories is the focus of the current thesis. The approach taken starts from the assumption of rationality and then explores the cognitive and ecological bounds that limit risky choice behaviour. Here, bounds are limits imposed on behaviour in the sense used by Simon (1955). Noise, for example, is one bound that is thought to play a crucial role in limiting human decision making. Howes et al. (2016) for example, shows how contextual choice effects in risky decision making are rational consequences of noise rather than consequences of bias. The thesis builds on previous work such as this with an investigation of how to use the latest machine learning methods to explore the implications of such bounds. It does so by using machine learning to find approximately optimal policies given a bounded optimization problem that is a theory of the risky choice problem faced by humans.

The ambition, then, is to explain the phenomena of risky choice in terms of policies that are optimised to the bounds. Two types of models are presented in this thesis. The first type (Chapter 3) trains deep neural networks on samples of risky choice problems that are presented as rectangles in a pixel array. The networks are trained to choose the largest rectangles and subsequently are shown to generate some risky choice phenomena – even though they were not explicitly trained to do so! Importantly, unlike, for example, Decision Field Theory (Busemeyer and Townsend, 1993), there is nothing special about these networks. In fact, they make use of public domain image classification networks. The second type of model (Chapters 4, 5 and 6) uses deep reinforcement learning to acquire rational policies given partial observations of the risky choice problem. Here, risky choice phenomena derive from uncertainty (noise) in the model's ability to determine expected utility. A feature of this approach is that evidence is accumulated over time by virtue of the sequential decision making process and that rational decisions about evidence gathering precede rational choice. The simulations thereby provide evidence about the underlying information processing mechanisms used by humans. These models push towards the goal of explaining a broad range of phenomena in a single framework. These phenomena include context effects (Wedell, 1991) and the fourfold pattern of choice (Kahneman and Tversky, 1979): the models exhibit risk seeking over low-probability gains, risk aversion over high-probability gains, risk aversion over low-probability losses, risk seeking over high-probability losses.

The contributions of the thesis are as follows:

- A number of models that offer novel and rational explanations for a broad range of phenomena exhibited by people making choice under risk. *Apparent* cognitive biases emerge from computationally rational processing.

- A unifying framework for modelling risky choice phenomena. Deep RL has the potential to help discriminate between various explanations because it provides a means of computing computationally rational policies given both ecological and cognitive bounds.

- A demonstration that context effects can emerge from an optimal (rational) "classifier" – a neural network – that chooses the option with the highest expected value. The network takes a bit-array as input (unlike existing models which are symbolic). The array contains bars or rectangles which vary in size, and the model is trained to prefer larger bars/rectangles. The bars/rectangles can also be presented with different layouts, and some layouts make comparisons between options easier.

- Some novel, previously unexplored, predictions concerning the *sequential* information gathering that precedes risky choice tasks. The generated sequential decision policy is not pre-programmed but learns to choose what information to gather about which options, calculates option values, and makes comparisons between options as the unfolding task demands, only that information that helps it maximise utility.

Chapter 2

Background

How does cognitive information processing lead to risky choice behaviour in humans? Are there systematic biases or is the process rational given limited capacities? Is there a unifying framework for modelling various fundamental cognitive components of human decision making required to explain risky choice? This chapter examines recent work aimed at answering these and related questions.

One thing is immediately apparent to the reader of the risky choice literature: there is a disconnect between the models of risky choice based on expected utility theory (rational models) and models of cognitive processing (process models). These are two threads of risky choice modelling, each taking a different direction. In the first thread, theorists are concerned with models, which map the properties of options to predictions of decision-makers' choices. Following prospect theory (Kahneman and Tversky, 1979), dozens of descriptive models of human decision making have been proposed. However, amongst these there are few *sequential* models and the underlying information processing mechanism has remained a mystery. The normative models propose theories of utility (subjective expected utility, regret theory etc.), but are mute as to the cognitive processes that generated the choices. As pointed out by (Simon (1978), p. 14), "in the past, economics has largely ignored the processes that rational man uses in reaching his resource allocation decisions". In parallel, psychologists have developed the second thread: process models. These include: heuristic models and accumulation process models or sequential sampling models, including Decision

Field Theory (DFT), Leaky Competing Accumulators (LCA), Decision by Sampling (DbS), Parallel Constraint Satisfaction (PCS), and Diffusion Decision Model (DDM) (Busemeyer and Townsend, 1993; Fiedler and Glöckner, 2012; Forstmann et al., 2016; Glöckner and Betsch, 2008; Glöckner and Herbold, 2011; Ratcliff, 1978; Ratcliff and McKoon, 2008; Ratcliff and Smith, 2004; Stewart et al., 2006; Usher and McClelland, 2001).

Unfortunately, the two approaches tend to have been pursued separately with their own interests and emphasises, which may not best serve science. Simon (1978, 1979) was one of the first to suggest greater cooperation. He advocated "building a theory of procedural rationality to complement existing theories of substantive rationality" and recommended that "some elements of such a theory can be borrowed from the neighboring disciplines of operations research, artificial intelligence, and cognitive psychology", but remarked that "an enormous job remains to be done to extend this work and to apply it to specifically economic problems" (Simon (1978), pp. 14-15). But, 40 years later, there have been few developments along these lines. McClelland et al. (2010) also advocated "an integrated approach to cognition in which functional considerations are grounded in, and informed by, the performance characteristics of the underlying neural implementation". Loomes (2019) emphasises this point in the recent book "Taking process into account when modeling risky choice".

Many in both the rational and process tradition have argued that there is evidence of systematic deviations between people's choices and the predictions of expected utility theory. To reduce the deviation, many researchers have proposed modifications to the theory by introducing additional parameters (e.g., outcome sensitivity, loss aversion, and probability sensitivity) with core elements of expected utility (Pachur et al., 2018). There are at least 62 prominent models of risky choice so far (He et al., 2020). As a consequence, "the existence of so many decision models complicates our understanding of choice behavior, impeding scientific progress" (He et al., 2020).

One type of process account is the evidence-accumulation models (Busemeyer et al., 2019; Busemeyer and Townsend, 1993; Ratcliff and McKoon, 2008; Stewart et al., 2006; Usher and McClelland, 2001). Commonly, these consider decision making as a stochastic process, known as a random walk in the discrete cases or diffusion process in the continuous cases. Moreover, these process models assume that the evaluation of attributes is a stochastic sequence. However, the rational basis of these models is not clear. They provide a description of the process without justifying the utility of the decisions made. In contrast, Navarro-Martinez et al. (2018), for example, suggests Boundedly Rational Expected Utility Theory (BREUT), which highlights the influence of the process, specifically the boundedly rational deliberation process. Despite its EUT core, the model could predict the choice patterns that deviate from EUT in line with several phenomena.

In what follows, I review the two threads of risky choice modelling (rational and process). Then, the disconnections, theoretical fragmentation, redundancy, and limitations of these approaches are pointed out and a new framework based on computation rationality (Lewis et al., 2014) is introduced to tackle the open problems. To apply the theoretical framework to risky choices, various computational methods are introduced and used to generate the policy for a wide range of tasks. The chapter closes with a preview of how the subsequent parts will build on and apply these computational principles.

2.1 Theories of Human Rationality

What does it mean to view a phenomenon as rational? The definition of *rationality* is a fundamental question for the neighbouring disciplines of Artificial Intelligence, cognitive science, economics, and neuroscience (Chater et al., 2018; Colman, 2003; Felin et al., 2017; Sturm, 2012). Philosophers and economists have built normative views of human rationality. These include the rules of logic, probability theory, and expected utility theory. However, a great number of human phenomena *appear* to violate these rules and theories. These psychological phenomena are considered as cognitive biases, which are taken as evidence that human cognition is irrational. The empirical evidence suggesting that people violate the expected utility theory, motivated behavioural economics and psychology to modify the normative theory of risky choice. The most notable contribution is prospect theory (Kahneman and Tversky, 1979) and its extension (Tversky and Kahneman, 1992).

2.1.1 Classical Rationality

The core hypothesis of rationality is that humans make decisions conforming to the maximisation of expected utility (Colman, 2003; Friedman and Savage, 1948; Savage, 1972; Von Neumann and Morgenstern, 1953). From the philosophers, such as Aristotle's theories of ethics and Bentham's theories of utilitarianism, to economic theories such as the marginal utility theory of value, have come the basic principles of human rationality: that people reason according to the normative rules of logic and mathematics.

Aristotle considers the human aim to become a good person, and he is credited with laying the foundation of formal logic. Utilitarianism's view is that the best action is the one that results in maximum utility, which is defined as that which produces the greatest well-being of the greatest number of people. Bentham provides a method to calculate the utility by summing all the pleasure resulting from an action and subtracting the pains involved in the action. The theory of utility, established by William Stanley Jevons, assumes that people make decisions to maximise utility. Decision theory is built on the mathematical foundation and connected with quantities from then on (Fishburn, 1970). These could be considered as practical rationality, which focuses on action. The classical rationality theory is formed based on these principles. For decision making under certainty, logic is the basic and normative theory of deductive reasoning. For example, people prefer option A to option B and option B to option C. People should then prefer option A to option C (transitivity). The optimal decision-maker would always choose the option with the highest utility. However, human reasoning usually involves risk and uncertainty in most real-world scenarios. For risky choices, we should use statistical inference based on probability theory when all involved information of alternatives, probabilities, and consequences are known. Therefore, the option with the maximum expected utility would be chosen. For decisions under uncertainty, people should choose the actions that maximising the expected utility when partial information of alternatives, probabilities, and consequences are known. The classical rational theory hypothesises that people reason according to the rules of logic, mathematics, and the theory of utility.

2.1.2 The Debate about Human Rationality

However, a series of experiments appear to show that human judgements violate the axioms of logic (Tversky et al., 1986; Wason, 1968) and statistical theory (Tversky and Kahneman, 1974). People's decisions appear to deviate from the predictions made by expected utility theory. Specifically, people's choices change between two options when adding a third alternative, known as preference reversal (Wedell, 1991). Also, Kahneman and Tversky (1979) show that people tend to be averse to risk when there would be sure gain and seek risk when there must be a loss. They also theorised that the human utility function is convex in gains and non-convex in losses. This is known as prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). People tend to use heuristics to make judgements or decisions with imperfect information. These violations of rationality are sometimes described as cognitive biases. This theory has attracted a great deal of attention among researchers studying human decision-making and risky choices.

However, risky choice has not only been studied in economics. In order to have a better understanding of the risk sensitivity of decision making, the study of risky choice should be integral across human cognition (Kacelnik and Bateson, 1996), behavioural economic theory (Bernoulli, 1954), and neural mechanisms (Platt and Huettel, 2008). Researchers in neuroscience have hypothesised that human reinforcement learning is a risk-sensitive process that learns about outcomes from trial-and-error experience (Niv et al., 2012). Neuroscientists consider that these phenomena stem from the brain's inherent capacity limitations (Shenhav et al., 2017).

One fact that finds support across various disciplines is that it is hard to acquire accurate information from the task environment, since perception and subjective representation are noisy (Findling and Wyart, 2021; Howes et al., 2016; Kahneman et al., 2021; Khaw et al., 2017; Steiner and Stewart, 2016; Stocker and Simoncelli, 2006a,b). Another fact is that it is difficult to identify the best action since computational cost needs to be taken into account for action selection (Bossaerts and Murawski, 2017; Russell, 1997).

2.1.3 Bounded Rationality

Economists have often used classical rationality as the model of economic entities to evaluate and forecast the market. However, for real-world problems, people should make decisions with limited computational resources and limited time in a particular environment. To bridge the gap between the theory and practical problems, Simon proposed the concept of bounded rationality, which revises the assumption of classical rationality. This ideal models human decision making in practical problems rather than in perfect situations. For Simon, people are "partly" rational, but human rationality is limited by the cognitive mechanisms of the mind and the structure of the environment. In bounded rationality, human decision-making is a rational process of finding a "satisficing" decision strategy given the limited computation resources and the available information.

Bounded rationality (Simon, 1955, 1956) provides the following basis for understating human cognition: people rationally behave with limited cognitive capacity as well as limited information and time. Simon's notion of bounded rationality has inspired various psychological theories, for instance, rational analysis, ecological rationality, Bayesian rationality, and computational rationality. These theories argue that human behaviour is rational, but these models of human cognition focus on different viewpoints: they vary from the structure of the mind to the environmental structure. Whereas the psychological theories influenced by bounded rationality are often descriptive and qualitative, many researchers in AI work on a formal and mathematically precise theoretical framework of bounded *optimality* (Russell and Subramanian, 1994). The theory of bounded optimality provides rigour and a general framework and reduces the gap between theory and practice in understanding rationality and intelligence (Russell, 1997). As an application of bounded optimality, recent work proposes the term computational rationality (Lewis et al., 2014) as a framework to advance our understanding of human cognition.

Descriptive Theories Inspired by Bounded Rationality

Inspired by Simon's work of bounded rationality, various assumptions and viewpoints have been explored for a better understanding (rational explanation) of apparently irrational cognitive biases.

Rational analysis emphasises the goals and problems that the cognitive system faces. Predicted behaviour is derived as an optimal solution for the problem under environmental constraints. According to the methodology of rational analyses (Anderson, 1991; Anderson et al., 1990), there are six steps to deriving a theory of the cognitive processes of human behaviour:

- 1.Specify precisely the goals of the cognitive system.
- 2.Develop a formal model of the environment to which the system is adapted.
- 3.Make minimal assumptions about computational limitations.
- 4.Derive the optimal behavioural function given 1-3 above.
- 5.Examine the empirical literature to see whether the predictions of the behavioural function are confirmed.
- 6.Repeat, iteratively refining the theory.

Rational analysis has been applied to memory and reasoning (Chater and Oaksford, 1999). The core objective of rational analysis is to specify the problem that the cognitive system is attempting to solve. But it does not provide the underlying computational processing directly. While Anderson thought of rational analysis as being inspired by bounded rationality, Simon was sceptical because of the extent to which bounded rationality minimises the role of cognitive bounds in theorising.

Following rational analysis, Bayesian rationality (Jones and Love, 2011; Oaksford and Chater, 2007, 2009) takes a probabilistic approach to human reasoning and rationality, which is defined by the ability to reason under uncertainty rather than by normative rules of logic. According to this theory, human reasoning is viewed as solving probabilistic inference problems rather than logical inference problems in real life. There is some evidence that people are poor at numerical calculation and the internal representation of the value of options is not stable (Noguchi and Stewart, 2018; Sanborn and Chater, 2016; Stewart et al., 2006; Vlaev et al., 2011). Bayesian modelling assumes that people make inferences according to the mathematics of probability theory. It also notes that the direct probabilistic calculating method is not acceptable since it is relatively computationally expensive (Perfors et al., 2011).

Also inspired by bounded rationality, there is much highly influential work on cognitive heuristics (Gigerenzer, 2004; Gigerenzer et al., 1999). Here it is hypothesised that people actually make judgements and decisions more accurately by *ignoring* part of the information. Reasoning proceeds without adhering to the rules of logic, calculating probabilities and maximising expected utilities but rather by using frugal, good-enough, heuristics. Using ecological rationality (Gigerenzer and Gaissmaier, 2011; Goldstein and Gigerenzer, 2002) showed that heuristics are rational strategies that are adapted to the structure of the environment and, perhaps, that they co-evolve with the fundamental cognitive mechanisms. Models of ecological rationality are mainly concerned with the environmental task structure.

Computational Rationality

In the field of AI, many researchers focus on the topic of AI and bounded rationality. Much of the work focuses on designing agents with limited resources in a practical way. First, the term bounded optimality is used to refer to "the optimisation of computational utility given a set of assumptions about expected problems and constraints in reasoning resources" (Horvitz, 1987). Then work by (Russell and Subramanian, 1994; Russell and Wefald, 1991) made the notion more precise and built the theoretical framework of bounded optimality. In this view, the program of a bounded-optimal agent (Russell and Subramanian, 1994) is the best solution to the constrained optimisation problem presented by its machine architecture and the specific task environment. It also provides a methodology to construct a provably bounded optimal agent:

- Specify the properties of the environment in which actions will be taken and the utility function on the behaviours.

- Specify a class of machines on which programs are to be run.
- Propose a construction method covering the processes and techniques used in the building process.
- Prove that the construction method succeeds in building bounded optimal agents.

There are two forms of bounded optimality used to explain human cognition (computational rationality (Lewis et al., 2014) and resource rationality (Lieder and Griffiths, 2019)). The former is more mechanistic using an internal causal structure, and the latter starts by focusing on the decision problem. As an application of bounded optimality, computational rationality (Lewis et al., 2014) is a framework to advance our understanding of human cognition. It not only explains why people make decisions but also how people make decisions. Moreover, it shows that human decision making is an adaptive process to the structure of the environment and the cognitive mechanisms (Howes et al., 2014).

Computationally rational agents are not unbounded optimal agents, but they do the best that they can with the computational resources that they have available. Lewis et al. (2014) proposed computational rationality as a general framework for understanding cognition that computes behavioural predictions from theories of resource constraints. For example, Howes et al. (2016) calculated the implications of noisy observations on risky choice tasks and showed that uncertainty in estimates of expected value could lead to apparent biases in a behaviour known as preference reversals.

2.2 Risky Choice Tasks and Human Behaviour

In this Section, I review two sets of behavioural phenomena that have been influential on theories of risky choice. The first set concerns choice context, and the second set concerns the fourfold pattern of risk attitudes.

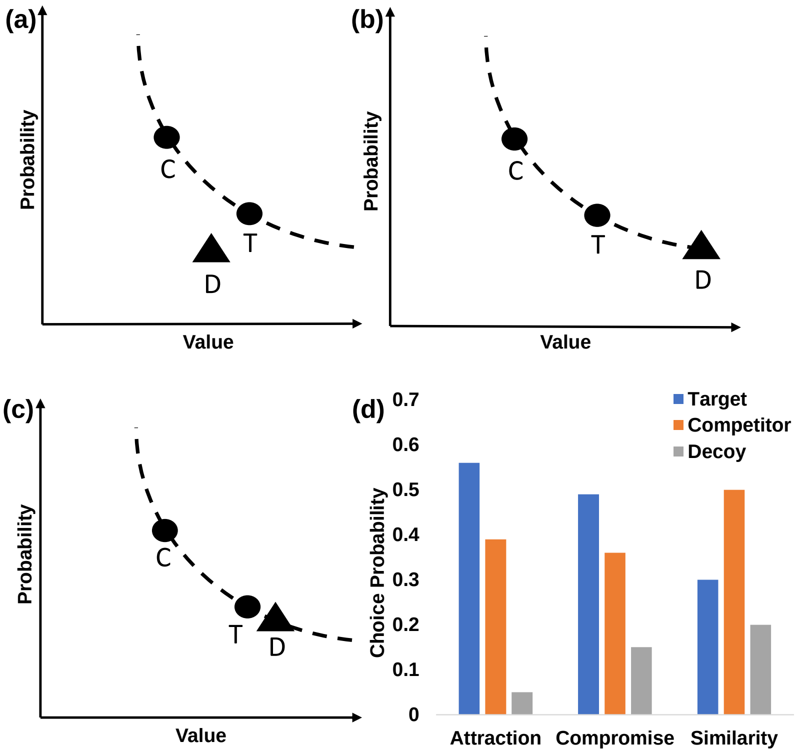


Fig. 2.1 (a)(b)(c) An illustration of the options in three types of contextual choice tasks – called the attraction (a), compromise (b), and similarity (c) tasks. The Target T and Competitor C are two options that have equal expected values (the dotted line). Option D is a decoy designed for comparison with the Target T. In the attraction task (a), T dominates D. In the compromise task (b), T is a compromise between D and C. In the similarity task (c), D has a similar expected value to T. (d) The proportion of choices of each of the three options (Target, Competitor, and Decoy) in each of the three contextual choice tasks (Attraction, Compromise, and Similarity). The Target is preferred in the Attraction and Compromise tasks, and the Competitor is preferred in the Similarity task. Data are reproduced from (Trueblood, 2012).

2.2.1 The Effect of Choice Context on Humans

Some of the human behaviours that have influenced this thesis are those exhibited in decision-making tasks in which people appear biased by seemingly irrelevant context. Three of the most well known contextual decision task are the attraction, compromise, and similarity tasks. These are illustrated in Figure 2.1a, b, c. For the attraction type task, there are two best options (the Target and the Competitor) with similar expected value. Each option is best on one dimension but not the other. One of these two options (the Target) dominates a third option, called the decoy, on both dimensions. It is difficult to choose between the two best options, since each option dominates the other on one of the attributes. Experiments studying these three tasks have been reported by many authors.

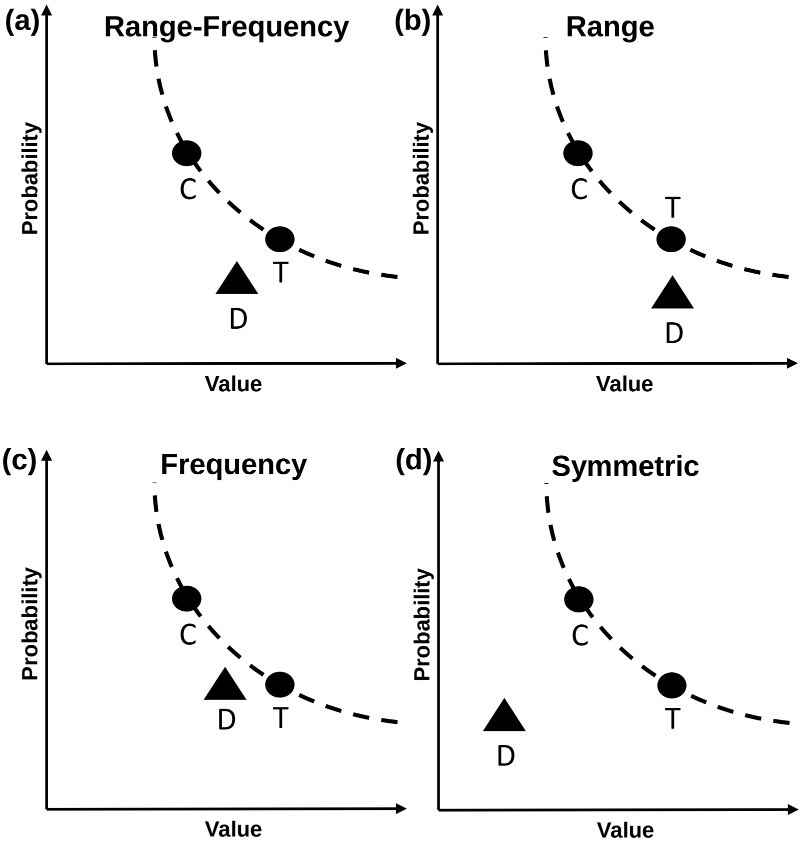


Fig. 2.2 The option positions are plotted in a two-dimensional space defined by the values of two attributes (probability and value). The Target T and Competitor C are two options that have equal expected values (the dotted line). Option D is a decoy designed for comparison with the Target T. As reported by Wedell (1991), the decoy options are range-frequency (a), range (b), frequency (c), and symmetric (d).

In the attraction-effect experiment, four different types of the decoy relative to the target were tested: range, frequency, range-frequency, and symmetric (Wedell, 1991). There positions for the decoy are shown in Figure 2.2. A range decoy is an option that is slightly weaker than the target alternative on the weakest attribute of the target alternative. On the target alternative's strongest attribute, the frequency decoy is a slightly inferior option to the target alternative. The range-frequency decoy refers to an option that is always dominated by the target option on both feature dimensions. The range decoy, frequency decoy, and range-frequency decoy are in the asymmetric conditions where the decoy is positioned significantly closer to the target option than the competitor option. A context effect was expected to be observed in the asymmetric conditions. However, simply introducing a decoy is insufficient to cause context effect. Furthermore, no effect is observed in the symmetric condition where the decoy is positioned far from the target and competitor options.

Consider the results of an experiment in which participants were asked to make decisions about criminal suspects (Trueblood, 2012). Participants were presented with a sequence of tasks each consisting of three suspects and were asked to decide which suspect was most likely to have committed a crime. There were two types of evidence, of varying strength, about each of the three suspects, such that the suspects had likelihoods of criminality in patterns identical to the three patterns presented in Figure 2.1d.

In the attraction condition of the experiment, there were two equally likely criminal suspects and a decoy suspect who was less likely than the other two (Figure 2.1a) but dominated by only one of the other choices. The experimental results showed that the Target suspect who dominates the decoy was chosen more frequently than the Competitor suspect. In the compromise condition of the experiment (Figure 2.1b), the findings showed that the suspect who is in-between the Competitor and Decoy is chosen more frequently than the Competitor. In the similarity condition (Figure 2.1c), the results showed that the suspect who is very similar to the decoy is chosen less frequently than the Competitor.

These effects have contributed to shaping a number of cognitive theories of human decision making (Busemeyer et al., 2019; Howes et al., 2016; Noguchi and Stewart, 2018; Roe et al., 2001; Ronayne and Brown, 2017; Spektor et al., 2019; Usher and McClelland, 2001; Wollschlaeger and Diederich, 2020). In some, though not all of these theories, human behaviour has been assumed to be biased because the irrelevant context (the decoy option) has consequences for the choice between the other two options (Tversky and Simonson (1993), p. 1188). The most commonly used operationalization of irrationality among decision researchers has been based on violations of value maximization. Preferring a dominated option or expressing different preferences depending on the framing of options demonstrates irrational decisions.

The significance of any irrationality, if that is what they are, cannot be understated given the potential for catastrophic real world consequences. However, the conclusion that choice under uncertainty provides evidence of irrationality may be incorrect (Howes et al., 2016; Tsetsos et al., 2016). Substantive analysis of the value of comparing options has shown that they are in fact informative and are required, under conditions of uncertainty, for reward maximization (Howes et al., 2014, 2016). The substantive structure of these analyses has informed the design of the agent that we present below. The key idea that is borrowed from human behaviour is the use of option comparison to inform decision making under uncertainty. The comparison was extensively explored by Stewart (Stewart et al., 2006; Vlaev et al., 2011), who has documented its use in a range of human decision-making tasks. For example, there is eye tracking evidence (Noguchi and Stewart, 2014) that people tend to make more eye movements that switch between options than eye movements that gather all of the evidence about a single option; evidence which is consistent with the use of comparisons.

2.2.2 Kahneman and Tversky's Fourfold Pattern of Risk Attitudes

In risky choice problems, options are often defined in terms of multiple attributes and minimally in terms of a gain and a probability. In common settings, a choice between two gambling options is used to demonstrate decision-making under uncertainty. The subjective estimation of a risky alternative is based on the product of the subjective evaluation of the outcomes and risks involved. Expected utility is one kind of such a subjective estimation. As reviewed in the previous section, expected utility theory has been the normal (or rational) theory of decision-making under risk. However, several classes of choice problems appear to exhibit violations of expected utility theory. For example, the framing of choice options (in terms of gain or loss, high or low probability) yields systematically different decisions in people (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). In particular, people behave according to the fourfold pattern of risk attitudes: “risk aversion for gains and risk seeking for losses of high probability; risk seeking for gains and risk aversion for losses of low probability” (Tversky and Kahneman (1992), p. 297). For example, people tend to choose the certain gain of £1000, when given a choice between the two options of a certain gain of £1000 or a 50% chance of a £2000 gain or a 50% likelihood of no gain. This phenomenon is so-called risk aversion in that people opt for gains of high probability. While, people tend to choose the risky option, when given a choice between the two alternatives of a certain loss of £1000 or a 50% chance of a £2000 loss or a 50% likelihood of no loss. This occurs since people become risk seeking for losses of high probability.

This fourfold pattern was demonstrated empirically by Kahneman and Tversky (1979) in a series of choice problems that showed that humans were subject to a set of choice biases. These included the certainty effect, isolation effect, the value function, and weighting function. Many other biases have been noted in the literature, but these, along with the contextual preference reversals (above), are the subject of this thesis.

2.3 Models of Human Choice

As I have said, behavioural studies have influenced a number of models of human decision making. Choice reversals, by themselves, have given rise to a number of models (Busemeyer et al., 2019; Frazier and Angela, 2008; Noguchi and Stewart, 2018; Roe et al., 2001; Ronayne and Brown, 2017; Trueblood et al., 2014; Usher and McClelland, 2001). Many of these models have focused on neurally plausible sequential processing, capturing the fact that decision making usually requires accumulation of evidence and integration of information across time (Gold and Shadlen, 2007; Tsetsos et al., 2016). Other models have focused on the way that people solve this problem by sampling comparisons between option attributes and thereby imposing a rank order on options (Noguchi and Stewart, 2018). However, none to my knowledge, have shown that preference reversals and fourfold pattern are an emergent consequence of an optimal solution to a Partially Obsevable Markov Decision Process (POMDP).

POMDPs provide a mathematical framework for sequential decision processes (Kaelbling et al., 1998). POMDPs have previously been used for modelling and explaining various aspects of human decision making (Daw et al., 2006; Dayan and Daw, 2008; Rao, 2010). An early contribution was Daw et al. (2006)'s model of the dopamine system which incorporated semi-Markov dynamics and partial observability. Rao (2010) proposed a model of neural information processing based on POMDPs and tested this model on perceptual tasks such as the random dot motion task. Further work in perceptual decision making, has used the POMDP framing to explore model confidence (Khalvati and Rao, 2015) and understand the role of priors (Huang et al., 2012b). POMDPs have also been used to model social decision making (Khalvati et al., 2016) and Theory of Mind (Baker et al., 2011, 2017; Rabinowitz et al., 2018). More recently, meta-level Markov decision processes (meta-MDP), a closely related framework, have been used for modelling higher level decision making (Griffiths et al., 2019). The Meta-MDP model is similar to the belief-MDP version of the POMDP, but replacing physical actions with cognitive operations. Meta-MDPs have been used to model strategy selection and heuristics in decision making (Callaway et al., 2022a; Kruegera et al., 2022; Lieder et al., 2017), attention allocation in simple choice (Callaway et al., 2021) and strategy in human planning (Callaway et al., 2022b).

2.4 Cognitive Process of Human Choice

The Observation of Comparison and Calculation

The literature empirically supports the assumption that people make comparison and calculation observations in choice tasks (Krajbich et al., 2010; Vlaev et al., 2011). Vlaev reviews these well-established cognitive and neural principles on the comparison and value evaluation scheme of choice. By an analysis of eye-tracking data, recent studies (Cataldo and Cohen, 2019; Noguchi and Stewart, 2014) show that the order in which people gather information supports comparison. The decision by sampling model (Noguchi and Stewart, 2018; Ronayne and Brown, 2017), which makes use of comparison, has substantial evidence. Also, the choice model (Howes et al., 2016) shows that comparison predicts performance in decision tasks.

Sequential Decision Making Process

Why are sequential processing models needed for human choice? This is a fundamental assumption of the model and is normatively motivated to understanding how context affects choice. Models of decision making based on sequential processing and evidence accumulation have been the dominant theory in cognitive science and decision neuroscience (Busemeyer et al., 2019; Forstmann et al., 2016; Ratcliff et al., 2016). A number of computational cognitive process models (Busemeyer et al., 2019; Noguchi and Stewart, 2018; Ronayne and Brown, 2017; Trueblood et al., 2014) hypothesize algorithms that model the sequential nature of human choice. These works focus on the effects that choice context and relational judgement has on human preferences. They explain context effects with algorithms that concern what evidence to gather and how that evidence is accumulated. For example, (Noguchi and Stewart, 2018) focuses on the way that people sample comparisons between option attributes and thereby impose a rank order on options.

Selective Attention and Accumulation to Threshold

Another fundamental principle of models of decision making is accumulation to threshold (Bogacz et al., 2006; Busemeyer et al., 2019; Krajbich et al., 2010; Krajbich and Rangel, 2011; Ratcliff et al., 2016; Song et al., 2019). Assumptions about how attention is allocated among the attributes are necessary for attribute processing models (Bowman and Wyble, 2007; Busemeyer et al., 2019; Wyble et al., 2009, 2011). The core assumption is that choices are made by gathering evidence for each choice option from the environment, continuing until a threshold (pre-specified value) is reached, at which point decision maker makes a choice in favour of the first one to reach the boundary or threshold. These models describe the dynamics of sequential decision making. The thresholds or boundaries, which represent the amount of noisy evidence that is required to accumulate before a decision is made, play a central role in decision theory. Recently, there has been growing interest in building a model of human decision making within the framework of computational rational analysis (Griffiths et al., 2015; Lewis et al., 2014). It offers explanations for why people use particular information gathering and decision strategies given the limitations on the resources available to them. They do not specify the mechanism to gather information and the threshold to reach a decision. Rather, they assume that the agent selects what item to attend to in an (approximately) optimal way by maximizing cumulative reward. They also assume that the decision is reached optimally, that the potential benefit of gathering information should be greater than the cost of gathering information.

Noise in the Process

They assume that human cognition is subject to noise (Findling et al., 2019; Findling and Wyart, 2021; Kahneman et al., 2021). Firstly, it is biologically plausible, building upon two widely accepted neural substrates: “that decisions are realized in a hierarchy of cortical layers and that processing at each layer is corrupted by independent neuronal noise” (Tsetsos et al., 2016). It is the noisy and distributed nature of the neural information process, which leads to irreducible observation noise. Secondly, it is consistent with the influential sequential sampling models (Busemeyer et al., 2019; Noguchi and Stewart, 2018). The mechanisms describe the stochasticity in the choice process. In addition, from a behavioral perspective, we know that people make errors, and noise is everywhere (Kahneman et al., 2016).

2.5 Machine Learning Models of Optimal Choice

Recently machine learning has achieved great progress, especially deep learning systems have attained human- and superhuman-level performance in a number of domains and also have the ability to mimic human-like representation. Human sequential decision-making could be framed as Partially Observable Markov Decision Processes (POMDP) (Dayan and Daw, 2008), which can be solved using a reinforcement learning method. Recent progress in machine learning suggests it to be a promising approach to modeling humans. It assumes that human decision-making is optimal or noisily optimal, whereas humans appear to deviate from optimisation as described by apparent human biases (Kahneman, 2016). Therefore, the combination of reinforcement learning and deep learning could be a promising and powerful method to model sequential decision making in a single general framework.

2.5.1 Markov Decision Process

There is a great deal of work on the tasks which could be considered to be Markov Decision Process (MDP) problems (Dayan and Daw, 2008) in the field of both psychology and neuroscience. The framework of MDP is a classical formalization of sequential decision making.

A finite MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where

- $\mathcal{S}$ is a finite set of states.
- $\mathcal{A}$ is a set of actions.
- $\mathcal{P}(s^{'}, r|s, a) := \Pr\{s_{t+1} = s^{'}, r_{t+1} = r|s_t = s, a_t = a\}$ is a probability of the next state $s^{'}$ and the next reward r occurring at time t, given the state s and the action a, where the Pr is the complete probability distribution for all $s^{'}$, r, s_t, and a_t and

$$\sum_{s^{'} \in \mathcal{S}} \sum_{r \in \mathcal{R}} \mathcal{P}(s^{'}, r|s, a) = 1, \text{for all } s \in \mathcal{S}, a \in \mathcal{A}.$$

- $\mathcal{R}(s^{'}|s, a) := \mathrm{E}\{r_{t+1}|s_t = s, a_t = a, s_{t+1} = s^{'}\}$ is the expected value of the next reward.
- $\gamma \in [0, 1]$ is a constant called the discount factor.

At each time step t, the agent receives the representation of the environment's state, $s_t \in \mathcal{S}$, and on that basis chooses an action, $a_t \in \mathcal{A}(s_t)$, where $\mathcal{A}(s_t)$ is the set of actions available in state s_t. Then, the agent receives a numerical reward, $r_{t+1} \in \mathcal{R}$, and finds itself in a new state, s_{t+1} (Sutton and Barto, 2018).

A (deterministic) policy $\pi : S \rightarrow A$ maps each state to an action that should be taken by the agent when in that state. The state value function of policy π, denoted by $V^{\pi}(s)$, maps each state s to the expected discounted cumulative reward that the agent could get starting from s and following policy π. The action value function of π, denoted by $Q^{\pi}(s, a)$, maps each state-action pair (s, a) to the expected discounted cumulative reward starting from s taking action a, then following π thereafter. Every roll-out of a policy accumulates rewards from the environment, resulting in the return. The goal of RL is to find an optimal policy, π^* that achieves the maximum expected return from all states:

$$\pi^* = \underset{\pi}{argmax}\, \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid \pi]$$

where $\mathbb{E}_\pi$ denotes the expected value of a random variable given that the agent follows policy π and I make the standard assumption that future rewards are discounted by a factor of γ per time-step (Mnih et al., 2015; Sutton and Barto, 2018).

A key concept underlying RL is the Markov property that only the current state affects the next state, or, in other words, the future is conditionally independent of the past given the present state. Although this assumption is held by the majority of RL algorithms, it is somewhat unrealistic, as it requires the states to be fully observable. A generalization of MDPs are partially observable MDPs (POMDPs), in which the agent receives an observation where the distribution of the observation is dependent on the current state and the previous action (Kaelbling et al., 1998).

2.5.2 Reinforcement Learning

Sequential decision making involves taking a sequence of actions to achieve the goal, maximising the expected cumulative rewards (Sun and Giles, 2001). Reinforcement Learning (RL) (Sutton and Barto, 2018) is such a general goal-driven learning framework that it can learn to accomplish the goal efficiently without hand-crafted rules and explicitly programming. RL could be a promising potential approach to model human sequential decision-making problems. Michael suggests that: "Reinforcement learning is also providing a valuable conceptual framework for work in psychology, cognitive science, behavioural economics, and neuroscience that seeks to explain the process of decision making in the natural world" (Littman, 2015).

RL is learning to take actions to maximise the cumulative reward signal by interacting with the environment. It is a computational approach to understanding and automating goal-directed learning and decision making (Sutton and Barto, 2018), using the framework of the Markov Decision Process (MDP), which is a classical formalisation of sequential decision making. The latest model of preference reversals (Howes et al., 2016) is not a sequential model, and there are still challenges underlying the information processing mechanisms of human choices. Recent work (Oulasvirta et al., 2018), explores the feasibility of modelling sequential decision problems as a Partially Observable Markov Decision Process (POMDP), which provides a rigorous mathematical framework for decision-making modelling. The RL framework is a good application for POMDPs and provides an efficient model of human learning processing. Research (Acuna and Schrater, 2009; Chen et al., 2015) have shown that optimal behavioural strategies could be learned in decision-making tasks using RL, which has made substantial progress in policy selection.

There are three classes of RL methods: value-based, policy-based and hybrid actor-critic models. These RL approaches learn a policy, which determines the agent's behaviour, in different ways: selecting actions based on value functions, action selection based on policy search, and hybrid methods employing both value functions and policy search. We currently model the rational choices using three reinforcement learning methods: Q-learning, Monte-Carlo-Policy-Gradient, and Actor-Critic, to verify the feasibility of this approach. The value function or policy function is updated by rewards in each task according to the methods described in the following sections.

The Value-based Model

Value-based methods learn the value of actions and choose actions based on estimating the expected value by acting greedily. The Q-learning (Watkins, 1989) is one of the most popular value-based RL algorithms and the most important breakthrough in the development of an off-policy Temporal-Difference (TD) learning algorithm (Sutton and Barto, 2018). One-step Q-learning is used in this report and defined by

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \left[r_{t+1} + \gamma \cdot \underset{a}{max}\, Q(s_{t+1},a) - Q(s_t,a_t) \right]$$

More formally, a deep convolutional neural network is used to approximate the optimal action-value function

$$Q^*(s,a) = \underset{\pi}{max}\, \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \mid s_t = s, a_t = s, \pi \right]$$

which is the maximum sum of rewards r_t discounted by γ at each time-step t, achievable by a behaviour policy $\pi = \mathcal{P}(a|s)$, after making an observation s and taking an action a (Mnih et al., 2015).

The action-value function Q directly approximates the optimal action-value function Q^* which is independent of the following policy. Thus it is known as off-policy control. The function Q is initialised to zero before training. In each episode, the agent runs with an $\varepsilon - greedy$ policy. In order to enable early convergence, the ε descends from 0.9 to 0.05 uniformly in the first 500 episodes. Thus, the model could converge within 1000 episodes instead of 5000 episodes.

The Policy-based Model

Unlike value-based methods in which policy is based on a value function, policy-based methods directly learn a parameterized policy which could choose actions. The policy gradient method changes the policy parameters in a way that improves the performance. The notation θ is used for the policy's parameter vector. The policy gradient method learns the policy parameter based on the gradient of some scalar performance measure $J(\theta)$ with respect to the policy parameter. These methods seek to maximize performance. The policy gradient theorem for the episodic case establishes that

$$\triangledown J(\theta) \propto \sum_s d^{\pi_\theta} \sum_a q_\pi(s,a) \triangledown_\theta \pi(a \mid s, \theta)$$

where d^{π_θ} is the stationary distribution of the Markov chain for on-policy π_θ. The REINFORCE (Williams, 1992) algorithm, also known as the Monte Carlo Policy Gradient, uses the full return G_t from time t, which accumulates all future rewards until the end of the task, as an unbiased sample of the value function $q_\pi(s,a)$. Whereas it is of high variance leading to slow learning because of Monte Carlo sampling. In order to reduce the variance of policy gradient, the advantage function $A^{\pi_\theta}(s)$ is used with a baseline function $B(s)$. The update rule of the REINFORCE algorithm with baseline is

$$\theta \leftarrow \theta + \alpha^{\theta}\gamma^{t}A^{\pi_{\theta}}(S)\nabla_{\theta}ln\pi(A_t \mid S_t, \theta)$$

where α^{θ} is parameter which denotes the step size for values. The state-value function $\hat{v}(S_t, \mathbf{w})$ could be a natural choice for the baseline, where w is a vector of state-value weights. Thus the advantage function is

$$A_t{}^{\pi_{\theta}}(S_t) \leftarrow G_t - \hat{v}(S_t, \mathbf{w})$$

where the state-value function $\hat{v}(S_t, \mathbf{w})$ is updated by the rule

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\gamma^{t}A_t{}^{\pi_{\theta}}(S_t)\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w})$$

where $\alpha^{\mathbf{w}}$ is parameter which denotes the step size for the policy.

The Actor-critic Model

A hybrid method has grown in popularity, known as actor-critic, which learns approximations to both value and policy functions. The policy function (‘actor’) learns the parameters from the feedback of the value function (‘critic’). Actor-critic methods combine policy search with learned value function, thus it could reduce variance and accelerate learning. Instead of using the complete return of REINFORCE, one-step actor-critic algorithms update with the one-step return as follows:

$$\theta \leftarrow \theta + \alpha^{\theta}\gamma^{t}A^{\pi_{\theta}}(S)\nabla_{\theta}ln\pi(A \mid S, \theta)$$

where uses a learned state-value function $\hat{v}(S_t, \mathbf{w})$ as the baseline

$$A^{\pi_{\theta}}(S) \leftarrow R + \gamma\hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$$

where S is the current state and S' is the next state. The state-value function $\hat{v}(S_t, \mathbf{w})$ is updated by the same rule described above.

In the following, we will distinguish between actor-critic and REINFORCE-with-baseline methods since they learn both the state-value and the policy function. The actor-critic methods use a learned state-value function as a critic and a baseline, while the REINFORCE-with-baseline methods use the state-value function only as a baseline. Unlike REINFORCE-with-baseline, which is unbiased and tends to learn slowly, the actor-critic methods significantly reduce the high variance of the policy gradient and also introduce the bias of the estimated value function. Thus it would learn faster and avoid converging on a local minimum.

2.5.3 Deep Reinforcement Learning

The theory of RL provides a method to improve an agent's ability to make decisions through interacting with the environment and evaluating feedback. The environment in practical problems is difficult to represent and generalise for the agent confronted with complex real-world scenarios. The RL method may easily lose its viability for solving these large problems. The generality of POMDPs would lead to mass computation and a large feature space for acquiring optimal strategies. Deep neural networks (Bengio, 2009; Hinton and Salakhutdinov, 2006; Krizhevsky et al., 2012), which have shown powerful ability in representation learning and function approximation, provide a new approach to overcome the problems described above. Deep learning (LeCun et al., 2015) approaches have made remarkable progress on the preference inference problem but do not directly address policy selection. Systems combining deep learning and RL, such as deep Q-network (DQN) (Mnih et al., 2015) and asynchronous advantage actor-critic (A3C) (Mnih et al., 2016), can successfully learn control policies in a range of different environments and achieve a higher level understanding of the environment. The recent model (Acharya et al., 2017) optimises strategies for visual search using DQN, which is a solution to a POMDP.

Deep Q-Network

DQN is a value-based DRL algorithm, using a deep convolutional neural network to approximate the optimal action-value function $Q(S,A)$. The authors (Mnih et al., 2015) use experience replay (Lin, 1992) and the target network to address unstable and convergent problems caused by using nonlinear function approximator. The experience replay mechanism could remove the correlations in the observation sequence by sampling the batches of observed experience. Updating action-value function $Q(S,A)$ periodically could reduce the correlations with the target which is defined as

$$Y_t^{DQN} = R_{t+1} + \gamma \cdot \underset{a}{max}\, Q(S_{t+1}, a; \theta_t^-)$$

where θ_t^- are the network parameters used to compute the target at iteration i. The DQN agent could achieve scores that are comparable to a professional human game tester across 49 Atari video games. The initial DQN algorithm has heavy data demands, which require around 38 days of video game experience. A series of techniques have been developed to reduce the data requirements.

The Double Q-learning (Hasselt, 2010) algorithm could reduce the overestimate bias due to using the maximum action value as an approximation of the maximum expected returns. Inspired by the idea of Double Q-learning that separates the action selection and action evaluation, Double DQN updates its target as

$$Y_t^{DoubleDQN} = R_{t+1} + \gamma \cdot Q(S_{t+1}, \underset{a}{argmax}\, Q(S_{t+1}, a; \theta_t); \theta_t^-)$$

where θ_t are the parameters of the Q-network at iteration i. The Double DQN architecture evaluates the policy according to the online network, while estimating its value based on the target network. This minimal adjustment could result in significantly better performance in the Atari 2600 domain.

Another way to adjust the DQN architecture is to separate the representation of state values and action advantages and innovate a neural network architecture named dueling architecture (Wang et al., 2016b). The Q-function is constructing as

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s,a';\theta,\alpha) \right)$$

where a is the current action and a' is the next action. The agent choose an action from a discrete set $a_t \in \mathcal{A} = \{1,...,|\mathcal{A}|\}$. The dueling network decomposes the Q-network into the stream $V(s;\theta,\beta)$ estimating the value function and the stream $A(s,a;\theta,\alpha)$ estimating the advantage function. In doing so, the agent could learn the state-value function efficiently.

In order to improve data efficiency, the framework of prioritized experience replay (Schaul et al., 2015) in DQN replays more frequent important transitions from which one could learn more effectively rather than sample uniformly from the replay memory at random.

Asynchronous Advantage Actor-critic

Actor-critic methods have gained popularity as a practical way to combine the advantages of policy search techniques with learnt value functions (Arulkumaran et al., 2017), that can learn from full returns and/or TD errors. Both policy gradient approaches and value function methods can benefit from improvements.

Deterministic policy gradients (DPGs) (Silver et al., 2014), which extend the common policy gradient theorems for stochastic policies (Williams, 1992) to deterministic policies, are one recent progress in the field of actor-critic algorithms. While stochastic policy gradients integrate over both state and action spaces, DPGs only integrate over the state space, requiring fewer samples in problems with large action spaces. This is one of the main advantages of DPGs. In addition, the Distributed Distributional Deep Deterministic Policy Gradient method (D4PG) (Barth-Maron et al., 2018), which incorporates several minor enhancements to the DDPG algorithm, achieves cutting-edge performance across a variety of control applications.

Utilizing parallel computation is an alternative strategy for accelerating learning. Computation can be efficiently distributed over both processing cores in a single central processing unit (CPU) and across CPUs in a cluster of machines by maintaining a canonical set of parameters that are read by and updated in an asynchronous fashion by numerous copies of a single network. One of the most well-known DRL methods in recent years is the asynchronous advantage actor-critic (A3C) algorithm (Mnih et al., 2016), which was developed for both single and distributed machine settings. Asynchronously updated policy and value function networks that were parallel-trained over many processing threads are the foundation of A3C, which integrates advantage updates with the actor-critic formulation. Multiple agents operating in separate, independent contexts not only stabilize parameter improvements but also provide an additional benefit by enabling further exploration. The distributed agent IMPALA (Importance Weighted Actor-Learner Architecture) (Espeholt et al., 2018) is a recent advancement that aims to solve a large number of tasks using a single reinforcement learning agent with a single set of parameters. IMPALA is the first Deep-RL agent that has been tested in such large-scale multi-task environments with success, and it would offer a straightforward yet scalable and reliable foundation for building better Deep-RL agents.

2.6 Outlook

Cognitive modelling has a central role in the computational foundations of the machine and human intelligence, which converges the insight between computer science, cognitive science and neuroscience. Building computational models of human cognition are useful for two main reasons. Firstly, it is a meaningful way to explore and understand the nature of cognitive progresses (McClelland, 2009), which is essential and critical to predicting human behaviours. Computational models could be used to explore and investigate the implications of our understanding of human behaviours in the same way as a psychological experiment on human participants in the laboratory. Depending on fundamental new insights, the new frameworks for thinking and modelling would emerge with increasing computing power (McClelland, 2009). Secondly, it enables tractable evaluating and testing of cognitive mechanisms and representations since computational models are much easier and cheaper to interpret and manipulate compared to natural stimuli with limited laboratory settings (Peterson et al., 2018). The deep neural network could be used to engage in complex cognitive tasks and indicate brain information processing (Kriegeskorte, 2015). The complex psychological phenomena could be recreated in the computer. What is more, some unobservable mental processes (McClelland, 2009), and mental representations (Peterson et al., 2018) taken in people's minds could be simulated by computer programs, which enable a novel path to investigate human cognition.

Meanwhile, computational models contribute to building more intelligent machines in the Artificial Intelligence (AI) field by re-implementing the transfer of the insights gained from cognitive science and neuroscience (Hassabis et al., 2017; Van Gerven, 2017). Humans have long been a source of inspiration for how to build intelligent machines (Hassabis et al., 2017; Lake et al., 2017). There is now a series of successful examples of where knowledge about the brain and mind has been used to develop new types of Machine Learning (ML), including artificial neural networks (McCulloch and Pitts, 1943), convolutional neural networks inspired in part by the hierarchical organization of vision (LeCun et al., 2015), and Reinforcement Learning (RL) which was inspired by decision making and learning under uncertainty in humans and other animals (Littman, 2015). A more recent example is provided by the promise of the utility of uncertainty, which has demonstrated that incorporating human-like uncertainty about object classifications can help obtain more robust and better performing machine classification (Peterson et al., 2019). Many recent advances have come from modelling uncertainty in ML. For example, capturing uncertainty can improve model performance in regression and classification tasks (Kendall and Gal, 2017; Kendall et al., 2018), estimating uncertainty can improve deep learning algorithms (Gal and Ghahramani, 2016; Maddox et al., 2019; Osawa et al., 2019), and representing the uncertainty of an agent's policy can aid more efficient exploration in RL (Fortunato et al., 2018; Janz et al., 2019; O'Donoghue et al., 2018). Another example of the influence of the human sciences on ML is how selective attention in human perception and neural information processing, has motivated rapid progress in object recognition (Ba et al., 2015), visual object tracking (Choi et al., 2018, 2016, 2017), human action recognition (Lee et al., 2015), image caption generation (Xu et al., 2015), and machine translation (Bahdanau et al., 2015; Vaswani et al., 2017). In sum, progress on multiple fronts suggests that human cognition offers a productive source of inspiration for improving ML.

Recent efforts toward a computational understanding of the human mind have been invigorated by advances in Artificial Intelligence (AI) (Cichy and Kaiser, 2019; Gershman et al., 2015; Lewis et al., 2014; Lieder and Griffiths, 2019). As machine learning has progressed, reinforcement and deep learning algorithms have generated systems that attained human- and superhuman-level performance in a number of domains, and it is believed by many researchers that modern AI not only has the capacity to equal human performance but also to help inform deeper understandings of human cognition (Bommasani et al., 2021; Hassabis et al., 2017; Lake et al., 2017; LeCun et al., 2015). In other words, building computational models of human cognition, informed by modern machine learning, offers a potential way to advance our understanding of cognitive processes (Fontanesi et al., 2019; Lieder et al., 2012; Lieder and Griffiths, 2017; McClelland, 2009; Milli et al., 2017).

In this thesis, I try to model human behaviours through machine learning techniques, which shows to be a novel and promising approach to both human decision-making modelling and AI application. As data about complex user behaviour has proliferated in recent years, it provides the intelligent computer with the opportunity to understand humans and explain actions from large amounts of data. The proposed research will construct a cognitive model and estimate its parameter values to pursue an understanding of human cognition as computationally rationality. The aim is to discover and improve human decision making in an integrated modelling framework.

Chapter 3

A Deep Learning Model of Contextual Choice Reversals

3.1 Introduction

Machine learning based methods have been used to solve a diverse set of complex scientific problems across a broad range of disciplines. Over recent years, there has been a rapidly growing interdisciplinary field that has been driven by the contributions from Artificial Intelligence (Hassabis et al., 2017), neuroscience (Lowet et al., 2020), cognitive science (Garcez et al., 2022), social science, chemistry, and more. Among them, recent efforts toward a computational understanding of the human mind have been invigorated by advances in Artificial Intelligence (AI) (Cichy and Kaiser, 2019; Gershman et al., 2015; Lewis et al., 2014; Lieder and Griffiths, 2019; Luo et al., 2021). It is believed by many researchers that modern AI not only has the capacity to equal human performance but also to help inform deeper understandings of human cognition (Hassabis et al., 2017; Lake et al., 2017). However, the extensive literature on suboptimal phenomena in perceptual decision-making tasks (Rahnev and Denison, 2018) seem to pose a severe challenge to this contention. In order to explore this tension further, in this Chapter, I will use Artificial Neural Networks as a computational model for human perceptual choice tasks.

The contributions of this Chapter are: (1) A demonstration that context effects (including attraction, compromise, and similarity) can emerge from an optimal (rational) classifier – a neural network – that chooses the option with the highest expected value. The network is implemented as a CNN. It takes a bit-array as input (unlike existing models of choice, which are symbolic). The array represents bars or rectangles which vary in size, and the model is trained to prefer larger bars/rectangles. The bars/rectangles can also be presented with different layouts, some layouts make comparisons between options easier. (2) A model which predicts context effects for some presentation layouts, but not others. In particular, context effects emerge – as for people – when the layout encourages attribute comparisons. The results provide evidence that the comparison process is important for contextual choice effects in artificial neural networks. (3) The approach shows that machine learning has the potential to accelerate the exploration and verification of psychological hypothesises.

Recent studies have found that the format in which choices are presented can affect contextual choice decisions (Cataldo and Cohen, 2019; Spektor et al., 2018). In these studies, choices are presented as rectangles or bars, and participants are asked to choose the largest rectangle/bar on the display. While these studies are not strictly speaking risky choice tasks, they do have an isomorphic structure. For risky choices, the two attributes are the outcome and the probability. But for the choice between rectangles, there is height and width. In both cases, there is uncertainty about the attribute values and also about the value of the outcome – the utility of getting the choice right.

The findings from these studies pose a serious challenge to computational cognitive models of contextual choice. Leading models do not make different predictions for different representational formats because they take values or symbols representing numerical values as input, which directly represent gains and probabilities. However, these symbolic inputs do not represent the arrangement and distance of the options in the images. To solve this research problem, this Chapter explores a new framework for modelling perceptual choice, which borrows from recent progress in computer vision. In this framework, deep neural network models are trained to make optimal decisions given training data that is sampled from distributions that correspond to the types of tasks that are presented to humans in experiments. The models reported in this Chapter take these sampled images as input and are then tested on the bar/rectangle arrangements used in human experiments. The analysis demonstrates when they capture human data and when they do not.

3.2 A Neural Network Model of Trueblood et al. (2013) Rectangle Choice Task Results

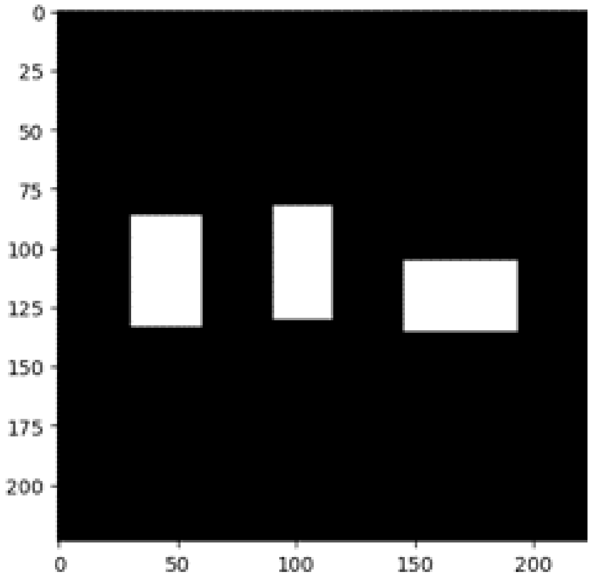


Fig. 3.1 An example of a range-decoy trial from the attraction-effect experiment. The axes give coordinates on the 224 × 224 pixel display. There are three white rectangles on the black background. The middle rectangle represents the decoy option, and the target option is on the left. The left rectangle and the middle rectangle have the same width, but the height of the left rectangle is greater than the middle rectangle. The right rectangle represents a competitor option, which has the same area as the left rectangle.

In the first study reported by Trueblood et al. (2013), participants were presented with three rectangles on each trial. They were asked to select the rectangle that had the largest area by pressing a key corresponding (from left to right) to that rectangle. The rectangles varied in both height and width, which can be thought of as two attributes representing the choices. To solve the task, participants might calculate the area of each rectangle by multiplying width and height, or perhaps by comparing the widths or heights of different rectangles. A rectangle that has a wider width and a larger height also has a larger area. A typical experimental trial is shown in Figure 3.1 which is reproduced from Trueblood et al. (2013). Each participant completed 720 trials, which were divided into three different types of attraction effect tasks. The results show that there was a main effect of the decoy on choice in all three of the conditions. Participants were more likely to select a rectangle if it dominated an irrelevant choice.

In the second study reported by Trueblood et al. (2013) participants were asked to perform a similar task to the first study, but this time the rectangle sizes were chosen to represent similarity-type contextual choice problems. Finally, in the third study, participants were asked to choose between rectangles in which one rectangle was a compromise between the other two.

3.2.1 Methods and the Experimental Setup

In this Section I describe my unified experimental setup, which is used throughout this section. I use the TensorFlow library for transfer learning using GPUs (NVIDIA GeForce RTX 3090). More details of my setup are provided below.

Source and Datasets. For training, I generated a large-scale image dataset containing 100,000 (100K) images and 3 distinct categories. Each image was a three-channel image with a resolution of 224 × 224 and had a single label. The categories (classes) include the labels: the left rectangle has the largest area, the middle rectangle has the largest area, and the right rectangle has the largest area. I sampled two separate sets of 50k examples from the full training set and used them as the training and testing sets, respectively.

The width and length of each rectangle were specified in pixels. For each rectangle, the value of width and length were sampled from a bivariate normal distribution ($mean = [50, 80], covariance\ matrix = [[100, 0], [0, 100]]$) following the settings in Trueblood et al. (2013). So the mean of the width is 50, and 80 for the length. In each task, three solid white rectangles were presented on the black background from left to right, as shown in Figure 3.1. The rectangles were randomly oriented vertically or horizontally, with the constraint that they do not all orient in the same direction. In other words, there are two situations:

two rectangles orient vertically while another orients horizontally, or two rectangles orient horizontally while another orients vertically.

Model and Training Details. A common approach to applying deep neural networks to a new domain is starting with a network that has been pre-trained on a very large data set. I used the models from keras.applications[1] developed by Google. All pre-trained networks were trained on ImageNet, one of the large-scale image data sets. I train all of my models using SGD with momentum 0.9. I use batch size 64 for all the models. I use a cosine learning rate schedule for 20k steps. I linearly warm up the learning rate for 2000 steps and sweep over two learning rates 0.03, 0.001.

To test the performance of the deep neural networks, I used various pre-trained networks as feature extractors and added one full-connected layer as a classifier on top. I trained the model in two steps:

- STEP1: For feature extraction, I trained the model for 20 epochs. The weights of the pre-trained network were frozen and not updated during training. The pre-trained network functions as a feature extractor. The newly-added classification layer was trained to choose the option with the highest area using what the pre-trained network had learned to represent the visual world's general features.

- STEP2: For fine-tuning, I trained for 40 epochs at this step. The weights of the last 30 layers of the pre-trained network were unfrozen. The newly-added classifier layers and the last 30 layers were trained jointly to fine-tune the higher-order feature representations, which are relevant to rectangle-sized perceptual tasks.

In summary, the model was trained to choose the one with the largest area out of three alternatives.

Metrics and Performance. For training and evaluating, I report top-1 classification accuracy as our main metric. The accuracy is the rate of the tasks in which the model predicted the right label–the rectangle with the largest area. Hyper-parameters were selected by the results from the validation split, and final numbers were reported from the test split.

[1] https://keras.io/api/applications/

After a total of 60 epochs, the model converged and demonstrated stable performance in randomly sampled tasks. For example, the MobileNetV2 model achieves 94.2% top-1 accuracy on a randomly sampled task set, as shown in Figure 3.2. It shows that the trained model performs well in optimizing correct decisions.

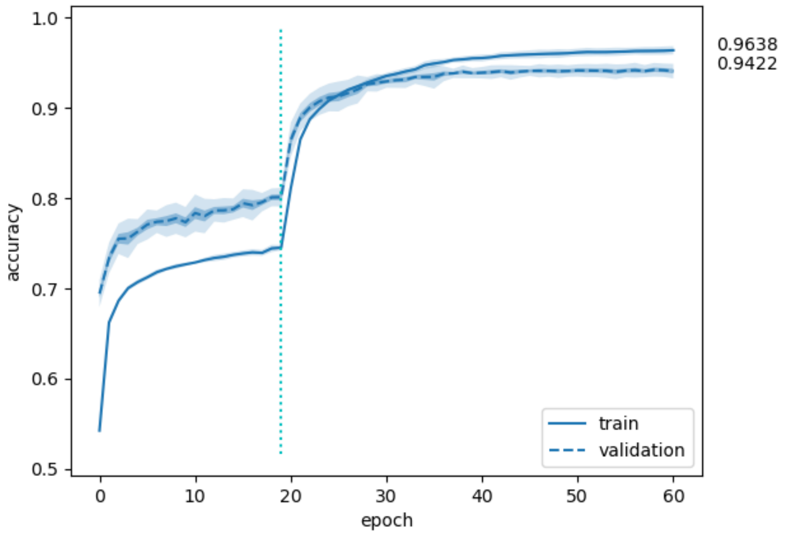


Fig. 3.2 Learning curves for the MobileNetV2 model on a randomly sampled task set. Both training and validation curves are shown. There are 20 feature-extracting epochs and 40 fine-tuned epochs. The results are the average classification accuracy on the different 5 runs.

I evaluated the trained model on contextual choice tasks used in human psychological experiments and compared its performance to the context effects observed by the participants. I generated an image dataset for contextual choice tasks, which contains 10K images. The values of the width and length of the rectangles were designed following the experiment used in Trueblood et al. (2013). The three rectangles (representing target, competitor, and decoy options) were placed randomly from left to right. All the details on contextual choice tasks can be found in Appendix A.

Table 3.1 A survey of the pre-trained models on context effect tasks on rectangle settings. The Classic(+), Null(0) and Reverse(-) range(r), frequency(f), range-frequency(rf), compromise(c) and similarity(s1 and s2) effect are tested for each model, when trained as feature-extractor (e) and fine-tuning(f) models. The overall numbers of models capturing the Classic(+), Null(0) and Reverse(-) effects for each type of effect appear at the bottom of the table. Each result is the average for 3 runs. The performance of the feature extractor is shown on a green background.

Pre-trained model	**r(e)**	**r(f)**	**f(e)**	**f(f)**	**rf(e)**	**rf(f)**	**c(e)**	**c(f)**	**s1(e)**	**s1(f)**	**s2(e)**	**s2(f)**
M0(MobileNetV2)	+	+	+	+	+	+	+	+	+	+	-	+
M1(Xception)	+	+	+	+	+	+	+	0	0	0	+	-
M2(EffNetB0)	+	+	+	+	+	+	+	+	-	+	-	0
M3(EffNetB7)	+	+	+	+	+	+	+	-	-	+	+	0
M4(ResNet50)	0	+	-	-	+	+	+	0	-	0	-	-
M5(DenseNet201)	+	+	+	+	+	+	0	-	0	0	-	-
M6(InceptionResNetV2)	+	+	+	+	+	+	+	+	-	+	-	+
M7(ResNet101V2)	+	+	+	+	+	+	-	0	+	+	+	0
M8(InceptionV3)	-	+	-	+	-	+	+	+	+	+	+	0
M9(ResNet50V2)	+	+	-	-	0	+	+	0	+	+	+	+
M10(ResNet152)	+	-	+	-	+	-	-	+	-	-	+	-
M11(ResNet152V2)	+	+	+	+	+	+	+	-	+	+	0	-
M12(ResNet101)	0	0	-	-	0	0	0	0	-	+	-	-
M13(DenseNet121)	+	+	-	-	+	+	0	-	-	+	-	-
M14(DenseNet169)	0	+	0	+	0	+	-	-	-	+	-	-
M15(EffNetB1)	+	+	+	+	+	+	+	0	-	+	-	-
M16(EffNetB2)	+	+	0	-	+	+	+	+	0	+	0	-
M17(EffNetB3)	+	-	+	+	+	+	+	-	-	+	-	+
M18(EffNetB4)	+	+	+	0	+	+	0	+	0	+	-	-
M19(EffNetB5)	+	+	+	-	+	+	+	+	-	-	0	0
M20(EffNetB6)	+	+	+	+	+	+	+	0	-	+	0	-
M21(NASNetMobile)	+	+	+	+	+	+	-	-	+	+	+	+
M22(MobileNetV3Small)	+	+	-	-	0	0	+	+	-	-	-	-
M23(MobileNetV3Large)	+	-	0	-	0	-	+	+	+	+	-	-
Classic(+)	20	20	15	14	19	20	16	10	7	18	7	5
Null(0)	3	1	3	1	5	2	4	7	4	3	4	5
Reverse(-)	1	3	6	9	1	2	4	7	13	3	13	14

3.2.2 Results

The results show that the majority of models (around 83% of 24 models) capture the Classic(+) attraction and compromise effect both at feature extraction and fine-tuning steps, as shown in Table 3.1. Also, these models capture the Reverse(-) similarity effect (s1 and s2) at the feature extraction step. This is consistent with the human data (Trueblood et al., 2013) when the stimuli consist of bars as the representation format of the values (i.e., rectangles with a common edge).

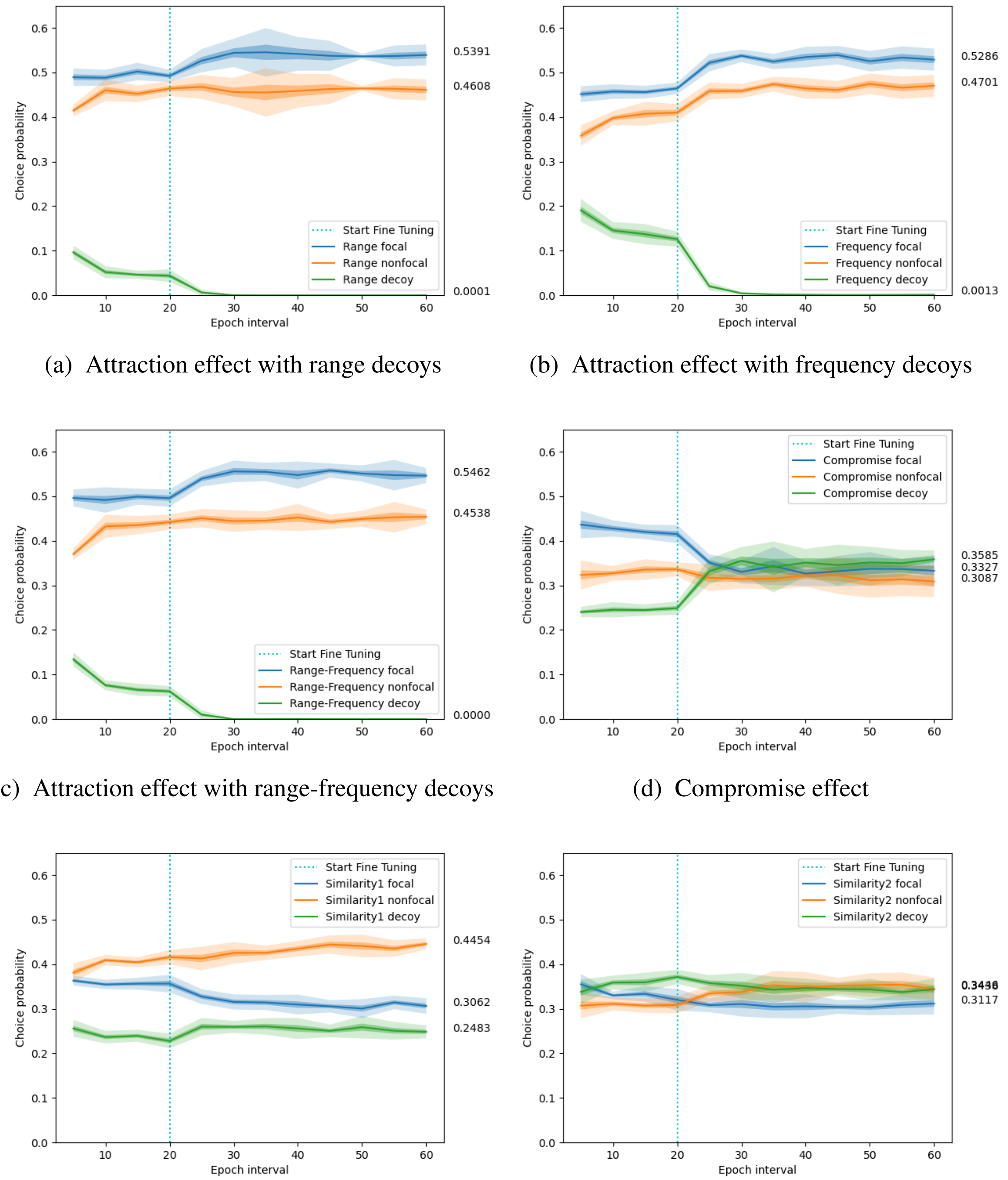
These initial results show that CNN models can model risky choices. Importantly, the models were not fitted to the data, rather they were trained to optimize correct decisions on the randomly sampled task set. The range of values used for training was significantly greater than the one used for evaluating the size of the context effect. In addition, the models were trained to prefer larger rectangles. Then I evaluated the models on context choice tasks. The results show that the models capture the context effects robustly.

I traced the evolution of the context effect during the training process, as shown in Figure 3.3. The results show that the trained networks generate contextual choice robustly, and the effects are stable after initial training. In Figure 3.3 a, b, c, the proportion of the decoy option (presented as a green line) decreases as the training epoch increases. As we know, the target and competitor options have the same area, and the attraction decoy option is inferior to the target option. So the attraction decoy option has the smallest area among the three alternatives. These findings demonstrate that the models learn to optimize correct decisions. After fine-tuning training for 10 epochs, the proportion of the decoy option diminishes and the proportion of the target option increases. Therefore, the size of the attraction effect increases as the model performs better in choosing the largest area rectangles.

I tested the effect of task distribution on the attraction effect (See Fig 3.4). Overall, all of the effects are predicted across three different training distributions (bivariate normal distribution, univariate normal distribution and uniform distribution) and different levels (from 1 to 8 levels) of variance in the distribution of the random sampling values.

It is also the case that the size of the attraction effect (range, frequency, and range-frequency) increases as the level of variance increases. In contrast, the size of the compromise and similarity (1 and 2) effect decreases as the level of variance decreases.

For compromise, the effect size is strong when the level of variance is low. One possible explanation is that the model learns a biased classifier because there is a low chance of learning from the extreme values in the training data set. The values are either extremely large or small in the compromise setting since the gambles at either side of the compromise option are at the extremes of the distribution. The compromise effect diminishes when the model learns from the values with a high level of variance.



(a) Attraction effect with range decoys

(b) Attraction effect with frequency decoys

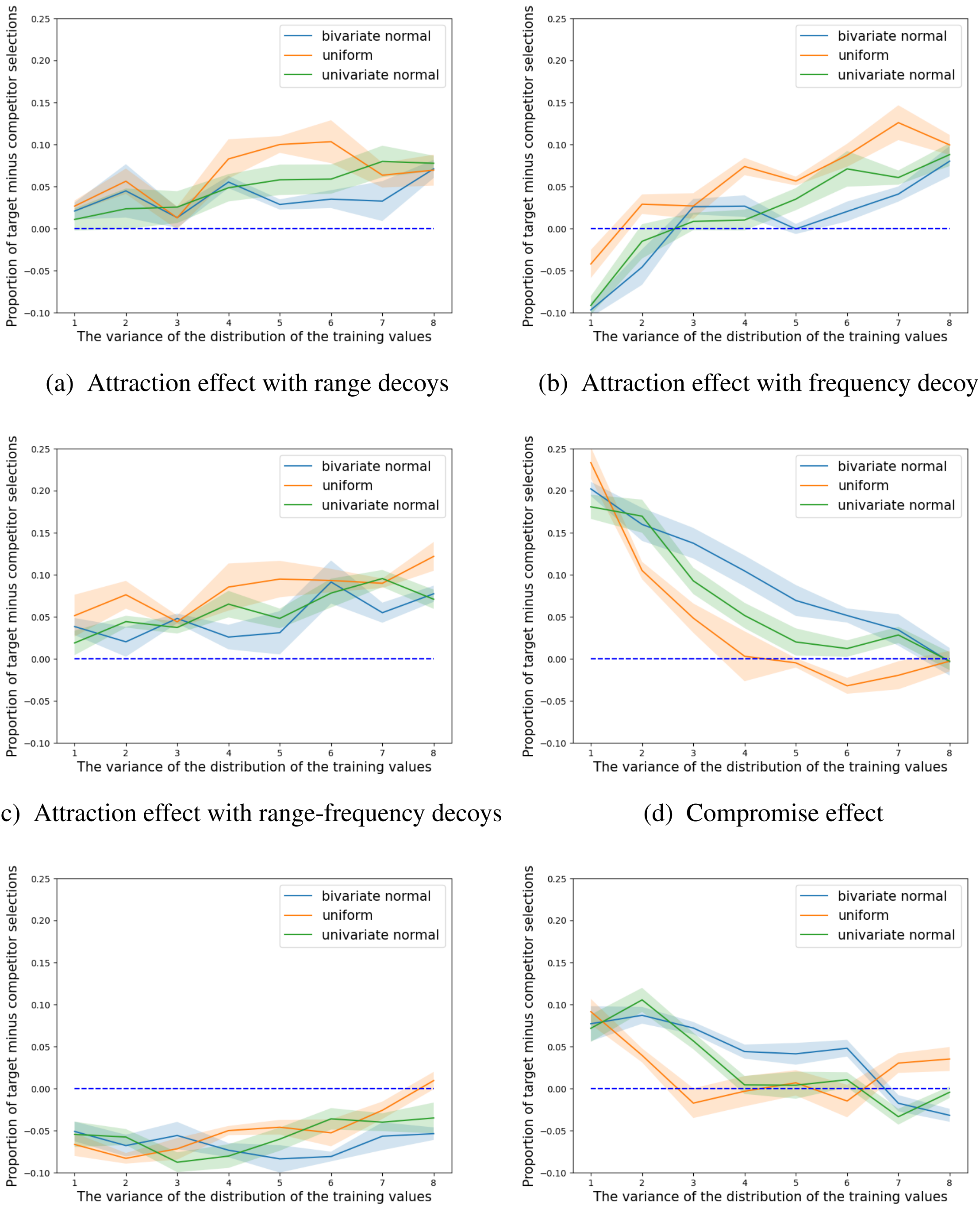
(c) Attraction effect with range-frequency decoys

(d) Compromise effect

(e) Similarity effect on choice set 1

(f) Similarity effect on choice set 2

Fig. 3.3 The choice proportion of the MobileNetV2 model over the training process. Each panel is for a different contextual choice effect. Trueblood et al. (2013) called the target option as “focal” option, and the competitor option as “nonfocal” option. I also use these items in the plot to make it easy to compare. The results are the average proportion of the different 5 runs. Confidence bands indicate a 95% confidence interval.



(a) Attraction effect with range decoys

(b) Attraction effect with frequency decoys

(c) Attraction effect with range-frequency decoys

(d) Compromise effect

(e) Similarity effect on choice set 1

(f) Similarity effect on choice set 2

Fig. 3.4 The proportion of target choice minus competitor choice against the variance of the distribution of the random sampling values. The pre-trained MobileNetV2 models are used as feature extractors. There are 3 types of distribution. The mean of the width was 50 pixels, and the mean of the length was 80 pixels. For bivariate normal distribution (shown as blue), the 1 ∼ 8 level variance are [5, 10, 20, 30, 50, 70, 100, 200], with no correlation between variance in two values. For univariate normal distribution (shown as green), the 1 ∼ 8 level variance are [2, 4, 6, 8, 10, 12, 14, 16]. For uniform distribution (shown as orange), the 8 levels are [5, 10, 15, 20, 25, 30, 40, 49]. For example, the sample range is from (30=50-20) to (70=50+20) when the level is 20 and the mean of the width is set to 50. Each result is the average for 5 runs. Confidence bands indicate (95%) confidence interval.

3.3 A Neural Network Model of Spektor et al. (2018) Representation Format Exploration Results

In the experiment reported by Spektor et al. (2018), participants were requested to do the same tasks as in Trueblood et al. (2013), i.e., to select the largest rectangle among three alternatives. All the experimental settings in Trueblood et al. (2013) were replicated, but the options were represented differently. In order to explore the effect of representation, Spektor introduced a new experimental condition. In this new condition (the triangular condition), the three rectangles were arranged as shown in Figure 3.5. This arrangement contrasts to Trueblood et al. (2013) where they were arranged along a straight line (the straight-line condition). The hypothesis was that participants would find it harder to make comparisons in the triangular condition, and context effects would therefore diminish. Spektor recruited 301 participants for 4 experiments. The results showed the attraction effect in the horizontal arrangement but the repulsion effect in the triangular arrangement. Also, the attraction effect is present when the target-decoy attribute distance is short, but the repulsion effect is present when the distance is large. In other words, distance moderates the context effects.

3.3.1 Methods and the Experimental Setup

I extended the model reported in the previous section of the thesis to the task reported by Spektor et al. (2018). As before, the model uses a pre-trained network as a feature extractor. One fully-connected network was added on top of the pre-trained feature extractor. This layer functions as a classifier that chooses the option with the largest area based on the features generated by the pre-trained network.

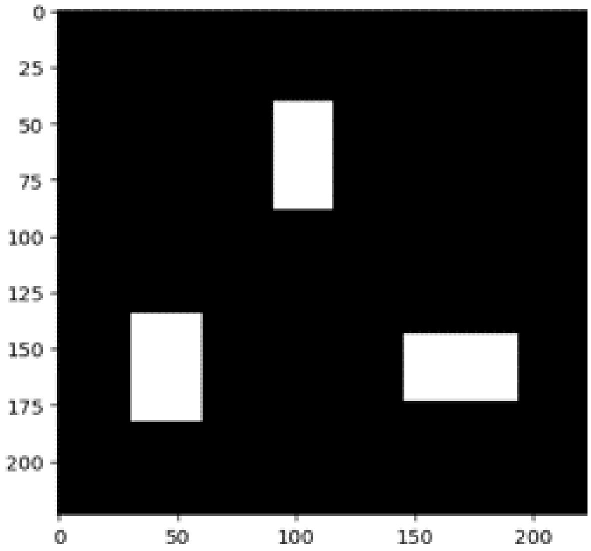
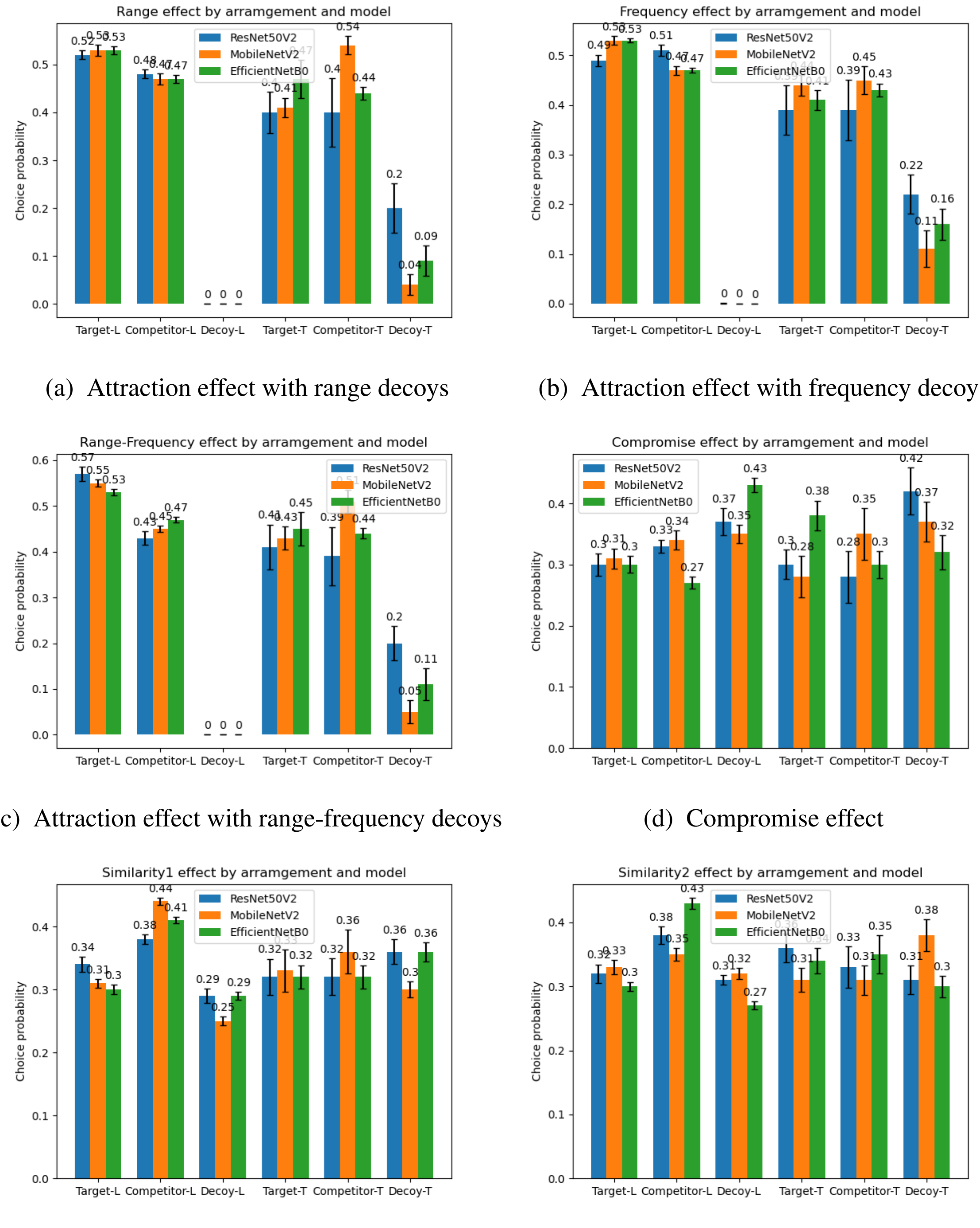


Fig. 3.5 Example of an experimental trial with triangular arrangement.

The model was trained with 100,000 sampled images with three rectangles. A sampled image consisted of three rectangles. For each rectangle, the width and length were sampled from a bivariate-normal distribution. The means were 50 and 80 for width and length, respectively. These means are the same means used for generating the test materials in Trueblood et al. (2013). The variance in each dimension was 20. These distributions give a range of materials within which the test materials used by Trueblood et al. (2013) and Spektor et al. (2018) fall. In addition, separate models were trained using two different "offsets". The offset was the amount of jitter that was used to perturb straight-line arrangements and generate a distribution that included some triangular arrangements. Two levels of offset were used (offset = 10 and offset = 30). After 50 epochs of training, the model was tested on images of which the width and height were those used in the human studies reported by Spektor et al. (2018).



(a) Attraction effect with range decoys

(b) Attraction effect with frequency decoys

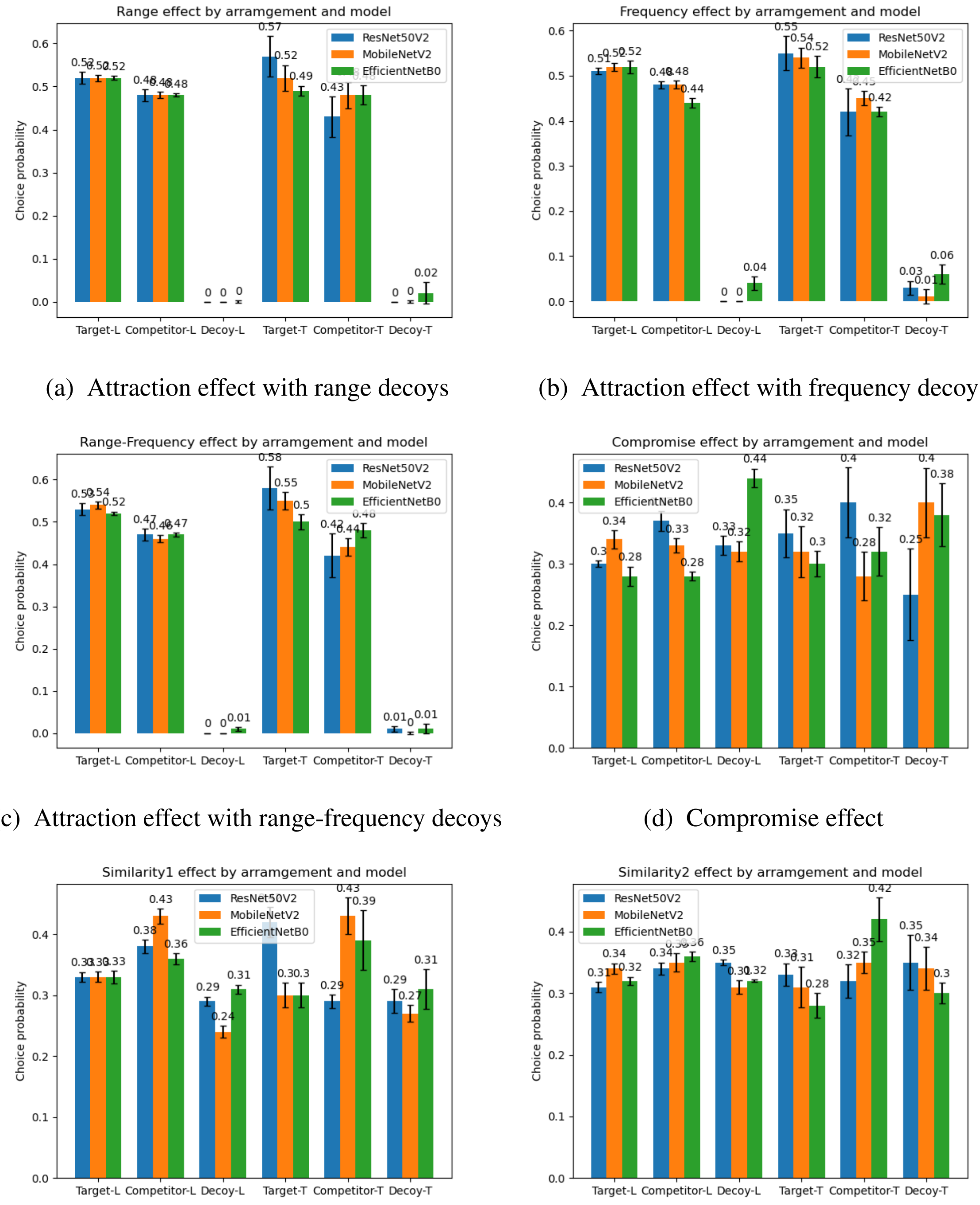
(c) Attraction effect with range-frequency decoys

(d) Compromise effect

(e) Similarity effect on choice set 1

(f) Similarity effect on choice set 2

Fig. 3.6 Six panels each represent the choice probability for target, competitor, and decoy in the straight line (L) and triangular (T) arrangements of rectangles. The training offset is set to 10. Each result is the average of 5 runs. Error bars indicate (95%) confidence interval.



(a) Attraction effect with range decoys

(b) Attraction effect with frequency decoys

(c) Attraction effect with range-frequency decoys

(d) Compromise effect

(e) Similarity effect on choice set 1

(f) Similarity effect on choice set 2

Fig. 3.7 Six panels each represent the choice probability for target, competitor, and decoy in the straight line (L) and triangular (T) arrangements of rectangles. The training offset is set to 30. Each result is the average of 5 runs. Error bars indicate (95%) confidence interval.

3.3.2 Results

The Effect of Straight Line and Traingular Arrangement on Contextual Choice Effects

After training, with offset = 10 used to generate sample tasks, the model exhibited the attraction effect for a straight-line arrangement of rectangles but not for the triangular arrangement. This effect is illustrated in the Figure 3.6 a, b, c. In each of these panels, the choice probability is plotted for the target, competitor, and decoy in both straight-line trials (L) and triangular arrangement trials (T). The model's predictions are consistent with the human experimental data reported in Spektor et al. (2018). The stimulus arrangement moderates the effect of contextual choice. The model also predicts a similarity effect but not a compromise effect.

A new model was trained with an offset of 30. The results for this model are shown in Figure 3.7. Here, the model generates an attraction effect for the straight-line arrangement and the triangular arrangement.

While the prediction for the straight line condition is consistent with Spektor et al. (2018)'s human data, the prediction for the triangle condition is not. One reason for the difference between the predictions for offset = 10 and offset = 30 may be that, with offset = 30, the training set includes more rectangles arranged in detectable triangles and that the model therefore gets more experience with the triangle arrangements. [2] With more relevant experience in offset = 30, the model is able to learn to use comparisons even for materials that are not arranged in a straight line, and the attraction effect is therefore observed.

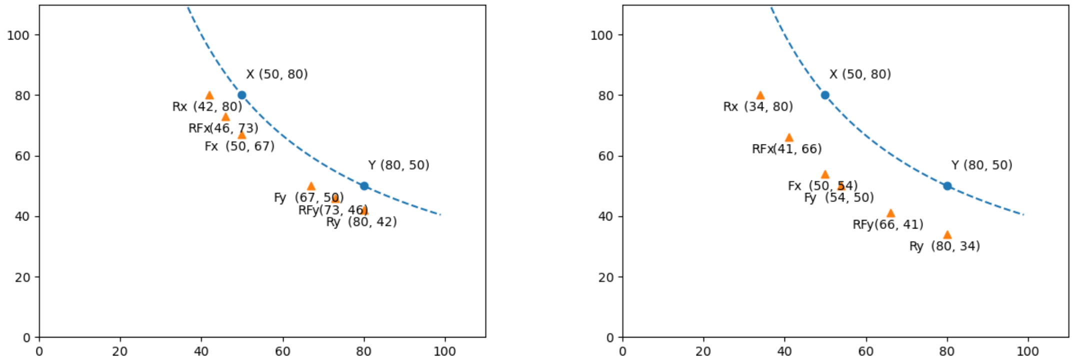
The Effect of Target-decoy Distance on Contextual Choice Effects

The target-decoy distance means the difference between the target and decoy in size. For example, when the distance is 16, the difference in size between the target and decoy is 16% of the area of the rectangle, as shown in Figure 3.8. In Trueblood et al. (2013), the distance is around 16% for the range and range-frequency decoys, and 10% for the frequency

[2] I say "detectable" here because the same proportion of triangle arrangements will be generated irrespective of the offset, but with the smaller offset the triangles will be flatter and more dissimilar to the test materials for the triangular tasks.

decoy. Of the 24 models surveyed in Table 3.1, 3 successfully captured the attraction effect in range, frequency, and range-frequency decoy types. These were: InceptionResNetV2, MobileNetV2 and EfficientNetB0. After training on the task, these models could exhibit the attraction effect in all 3 decoy types both before and after fine-tuning. Therefore, I extended these models to the experiment design reported by Spektor et al. (2018).

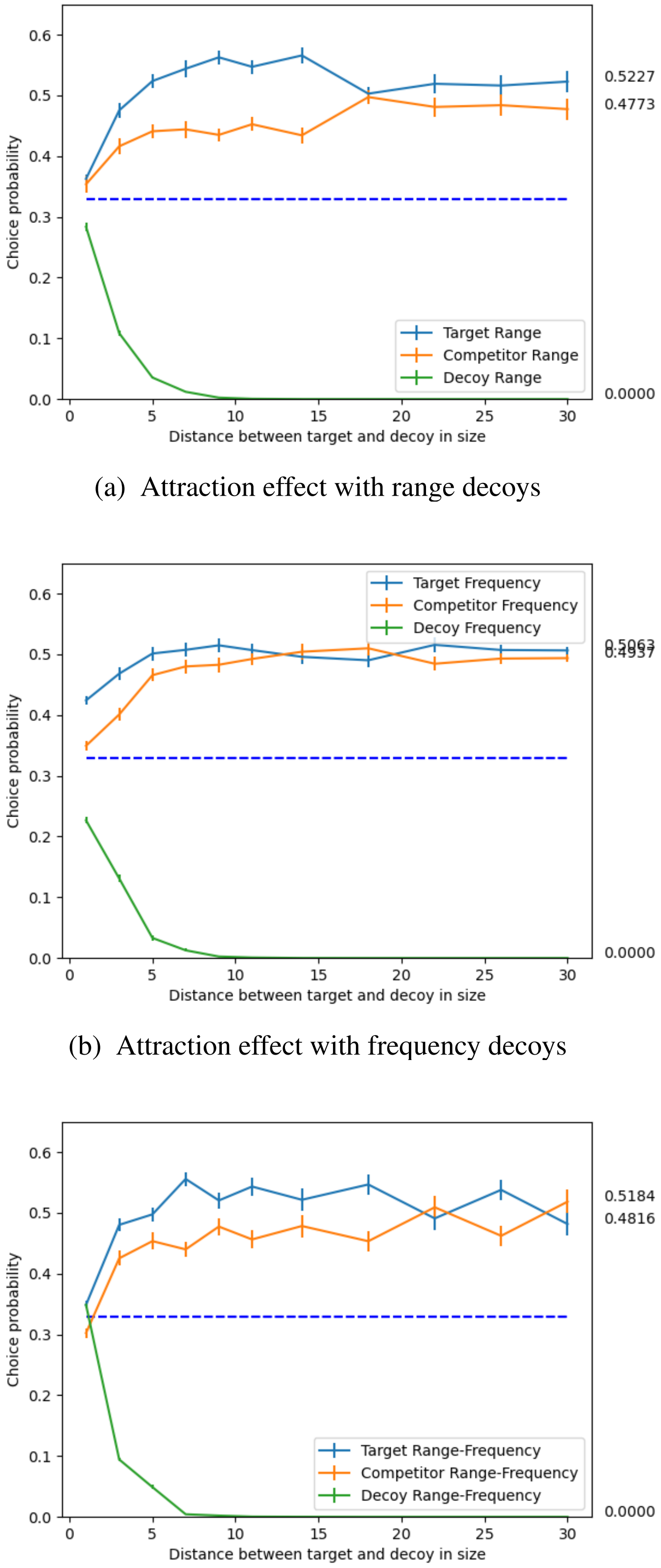
Figure 3.9, 3.10, 3.11 depict the distance effects. The size of the attraction effect increases as the distance increases. Then the size of the effect achieves maximal value at a certain distance. The proportions of the target and competitor increase as the distance increases from 2 to 10, while the proportion of the decoy decreases. This prediction is consistent with human experiments (Spektor et al., 2018). Therefore, it is plausible that people choose the decoy less and the target and competitor more since the task is easier as the distance is larger.



(a) Distance = 16

(b) Distance = 32

Fig. 3.8 Examples of the locations of options for attraction effects with the different target-decoy distances.

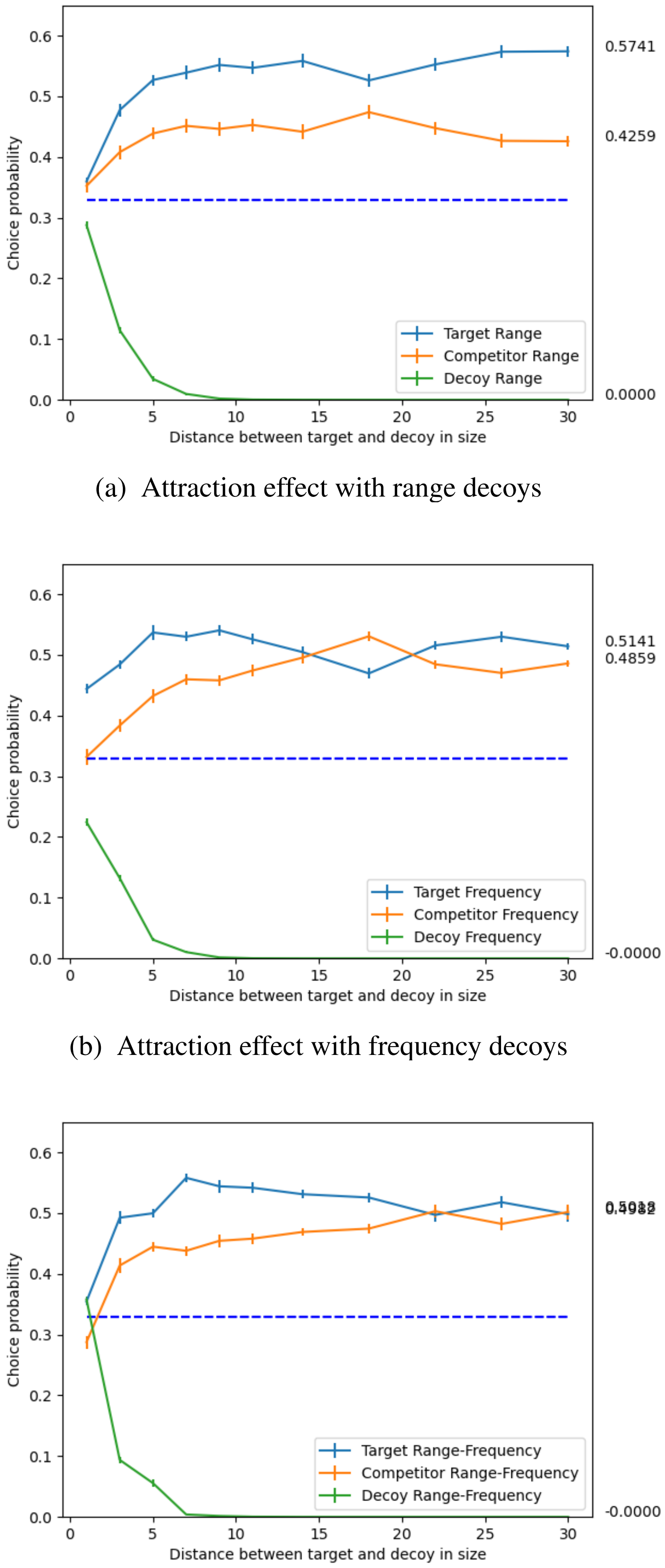


(a) Attraction effect with range decoys

(b) Attraction effect with frequency decoys

(c) Attraction effect with range-frequency decoys

Fig. 3.9 Choice probability against distance for the InceptionResNetV2 pre-trained model. The blue line represents the target choice probability. The orange line is for the competitor, and the green line is for the decoy. For each model, the data is the average of 10 runs for each level of distance. The error bars indicate (95%) confidence intervals.

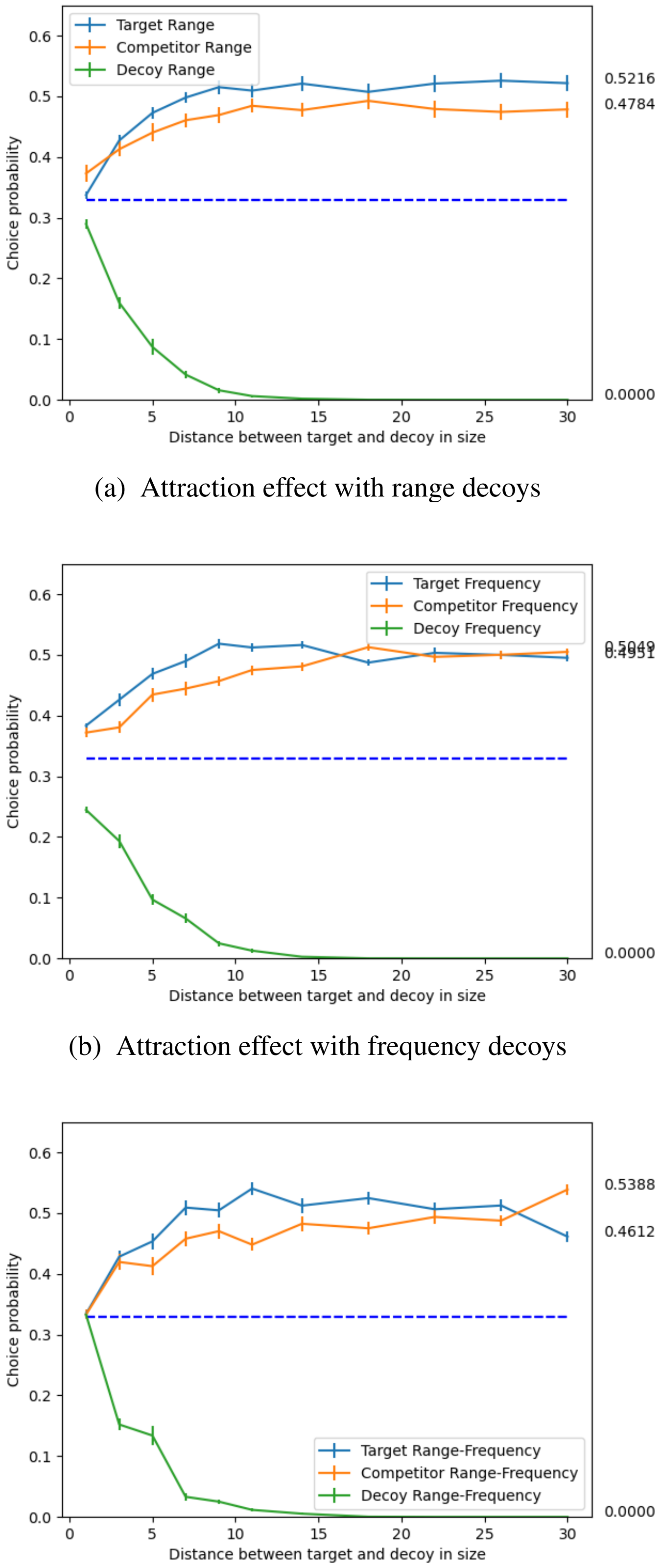


(a) Attraction effect with range decoys

(b) Attraction effect with frequency decoys

(c) Attraction effect with range-frequency decoys

Fig. 3.10 Choice probability against distance for the MobileNetV2 pre-trained model. The blue line represents the target choice probability. The orange line is for the competitor, and the green line is for the decoy. The performance of each model is averaged over 10 runs at each distance. The error bars indicate (95%) confidence intervals.



(a) Attraction effect with range decoys

(b) Attraction effect with frequency decoys

(c) Attraction effect with range-frequency decoys

Fig. 3.11 Choice probability against distance for the EfficientNetB0 pre-trained model. The blue line represents the choice probability of the target. The orange line is for the competitor, and the green line is for the decoy. Each model's prediction is the average of 10 runs on each distance. The error bars indicate (95%) confidence intervals.

3.4 A Neural Network Model of Cataldo and Cohen (2019) Bar Choice Task Results

In the previous two Sections, the choice options were represented as rectangles. Each rectangle indicates one alternative. The height and width of the rectangle represent the two attributes of a choice option. In this experimental setting, both the neural network models and human participants exhibit contextual preference reversals when the representation format encourages comparison (i.e., one side of the rectangles is aligned). To further test the influence of choice representation, Cataldo and Cohen (2018, 2019) introduced new stimuli settings in which the choice options are represented as pairs of bars. In the new case, each couple of bars represents one alternative. For each couple, the horizontal lengths of two bars correspond to the values of two attributes of one option. See Figure 3.12.

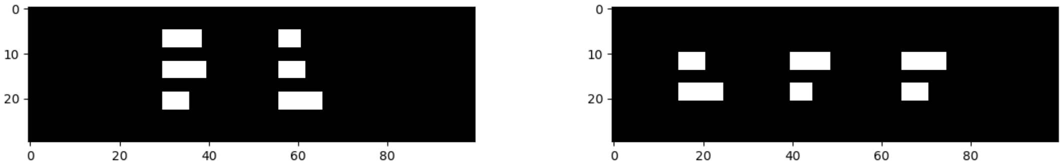


Fig. 3.12 An example of the choice tasks used in Cataldo and Cohen (2019); displayed by-dimension (left) and by-alternative (right).

The choice set is the same as before in that each task consisted of three options varying in dimension values (size and location). Here, a group of two bars represents the two values of a single choice. The participants in Cataldo and Cohen (2018) were required to press a key (1, 2, or 3) corresponding to the desired alternative. To do the task, the participants might compare the dimension values among different options or calculate the utility of an alternative. There are two types of presentation formats, which are by-dimension and by-alternative. For by-dimension, each alternative is represented by a group of bars on a horizontal. In this condition, the arrangement of bars encouraged participants to make comparisons within a dimension by looking up and down the vertical arrangement of each of the two dimensions. Whereas for by-alternative, each alternative was represented by a group of bars on a vertical axis. In this condition, it was harder for participants to compare the values of attributes among options.

3.4.1 Methods and the Experimental Setup

All the models were trained to choose the group of bars with the largest sum of the horizontal length among the three alternatives. I also extended the model reported in the previous sections of this Chapter to the task reported by Cataldo and Cohen (2019). Besides the pre-trained network, I also explored the effect of the neural network's architecture. Finally, I used a specific designed neural network to fit the human performance in Cataldo and Cohen (2019). More details of my unified experimental setup are provided below.

For training, I generated two large-scale image datasets, each containing 100K images. One choice image set was displayed in the by-alternative presentation format, and another was displayed in the by-dimension format. In both cases, the presentation strongly encouraged comparisons within columns rather than within rows (Cataldo and Cohen, 2019). For each couple of bars, the values of the horizontal length were sampled from a discrete uniform distribution ($interval = [1, 100]$). In addition, the vertical height of the bar was set to constant (20) pixels across presentation format conditions.

For evaluation, I generated two image datasets for contextual choice tasks, each containing 10K images. The values of the horizontal length of the bars were designed following the experiment used in Cataldo and Cohen (2019). The three groups (representing target, competitor, and decoy options) were placed randomly. The two options (X and Y) had the equal sum of the values of the horizontal length for each couple. The decoy option was either A_X or A_Y for the attraction effect, C_X or C_Y for compromise effect, and S_X or S_Y for the similarity effect. The subscripts index the target option. The details on the value used in contextual choice tasks can be found in Appendix A.

3.4.2 The Survey on the Various Model on the Bars Setting

After a total of 120 epochs, all the models converged and demonstrated stable performance in randomly sampled tasks. For example, both the ResNet50 and ResNet152 models achieve 99.3% top-1 accuracy on a randomly sampled task set, as shown in Figure 3.13. It shows that the trained models perform well in optimizing correct decisions.

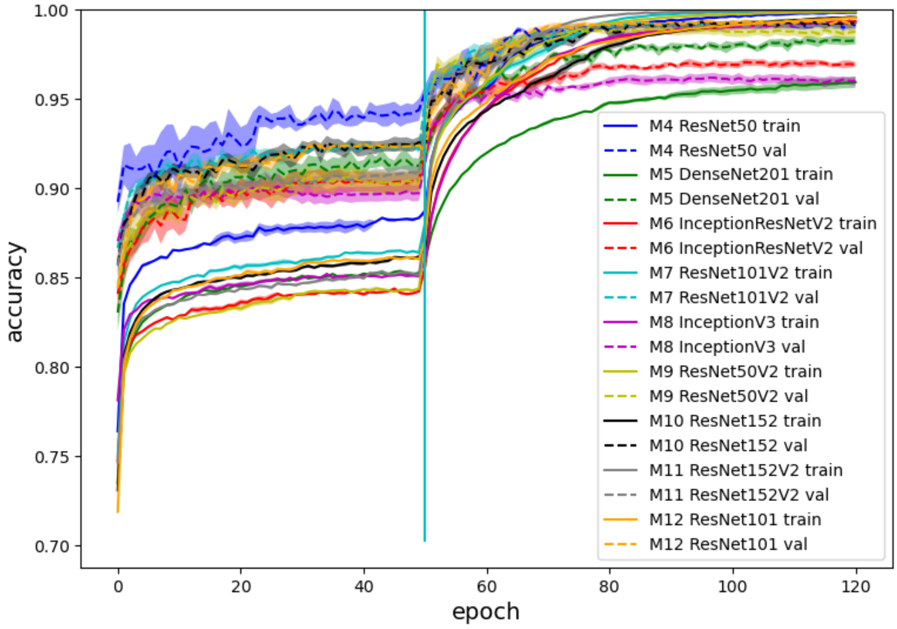


Fig. 3.13 The learning curves for pre-trained models. Both training and validation curves are shown. There are 50 feature-extracting epochs and 70 fine-tuned epochs. The results are the average classification accuracy on the different 5 runs. Confidence bands indicate (95%) confidence interval.

I evaluated the trained models on contextual choice tasks used in human stimulus and compared their performance to the context effects observed by the participants. I only evaluated the models on the choice tasks displayed in the by-dimension representation format because participants in the by-alternative condition show a weak or null attraction and compromise effect (Cataldo and Cohen, 2019). Nevertheless, the predictions of my models show that over half of the models capture the Classic(+) attraction and Reverse(-) similarity effect at feature extraction steps, as shown in Table 3.2. This finding is consistent with the human data (Cataldo and Cohen, 2019).

However, there were no significant effects on other conditions. The results indicate that the neural network's architecture plays an important role in capturing the context effects. In the rest of this Section, I will explore the effect of the neural network's architecture on modelling contextual choice tasks.

Table 3.2 A survey of the pre-trained models on context effect tasks on bar settings. The Classic(+), Null(0) and Reverse(-) attraction(a), compromise(c), and similarity(s) effect are tested for each model, when trained as feature-extractor (e) and fine-tuning(f) models. The overall numbers of models capturing the Classic(+), Null(0), and Reverse(-) effects for each type of effect appear at the bottom of the table. Each result is the average of 3 runs. The performance of the feature extractor is shown on a green background.

Pre-trained model	**a(e)**	**a(f)**	**c(e)**	**c(f)**	**s(e)**	**s(f)**
M0(MobileNetV2)	+	+	-	-	-	-
M1(Xception)	+	0	+	+	-	+
M2(EffNetB0)	+	0	-	-	+	+
M3(EffNetB7)	-	+	0	0	+	+
M4(ResNet50)	+	+	0	0	+	+
M5(DenseNet201)	-	-	-	-	+	0
M6(InceptionResNetV2)	0	-	0	+	-	+
M7(ResNet101V2)	+	0	-	-	-	+
M8(InceptionV3)	+	+	+	+	-	-
M9(ResNet50V2)	-	+	0	0	-	-
M10(ResNet152)	+	0	+	0	0	0
Classic(+)	7	5	3	3	4	6
Null(0)	1	4	4	4	1	2
Reverse(-)	3	2	4	4	6	3

3.4.3 Explore the Architectures of the Neural Network

In the last section, I test the feasibility of neural network models in capturing the context effect. The results show that the performance varies in different pre-trained models with generic features of the visual world. The deep neural network might be easy to fit the data since the training images are within a narrow range. To test the effect of the neural network's architecture, I focus on kernel size k and block size b.

I use a ResNet stem block ($k \times k$ convolution + batch normalization + ReLU + max-pooling) followed by a variable number of bottleneck blocks (He et al., 2016; Simonyan and Zisserman, 2014). The kernel size k ranges from 3 to 14, which refers to a very small receptive field (3×3) to a large one (14×14). The convolution stride is fixed to 1 pixel. The convolution layer follows the batch normalization layer. All hidden layers are equipped with the rectification (ReLU (Krizhevsky et al., 2012)) non-linearity. Max-pooling is performed over a 2×2 pixel window, with stride 2.

A stack of ResNet stem blocks (which has a different depth in different architectures) is followed by one Fully-Connected (FC) layer. The block size b is the number of ResNet stem blocks that refer to the convolutional network depth. The FC layer has 3 channels (one for each class) and performs 3–way classification. The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks. All the parameters for the networks are those of He et al. (2016).

To avoid the effect of other factors, 3 representation parameters are set to randomly sampled values. Examples of these situations are shown in Figure 3.14.

First, I evaluated the trained models on the randomly sampled choice tasks, as shown in Figure 3.15. The results show that the model with a small kernel size and shallow depth (small block size) cannot achieve high accuracy (0.98). Therefore, I excluded the results using these parameters in the rest of this Section.

Next, I used these models to predict which context effects were present. The predicted size of each context effect are presented in Figure 3.16, 3.17, 3.18. The results demonstrate that the attraction effect was displayed robustly throughout the models with various architectures (Figure 3.17). This finding is well supported by the evidence that most effect sizes are significantly larger than 0 (above the red line in Figure 3.17). However, the models predicted a reversed compromise effect when the decoy options targeted option Y (Figure 3.16). And the performance was not stable when the decoy options targeted option X (Figure 3.16). It is interesting to find that the models capture a reversed similarity effect when option X2 was the target (Figure 3.18 c), whereas a classic similarity effect occurs when the decoy options targeted option Y1 (Figure 3.18 b).

I note that the results presented in this Section are consistent with the predictions made by pre-trained models, which can capture the attraction effect robustly but fail to capture the stable compromise and similarity effects.

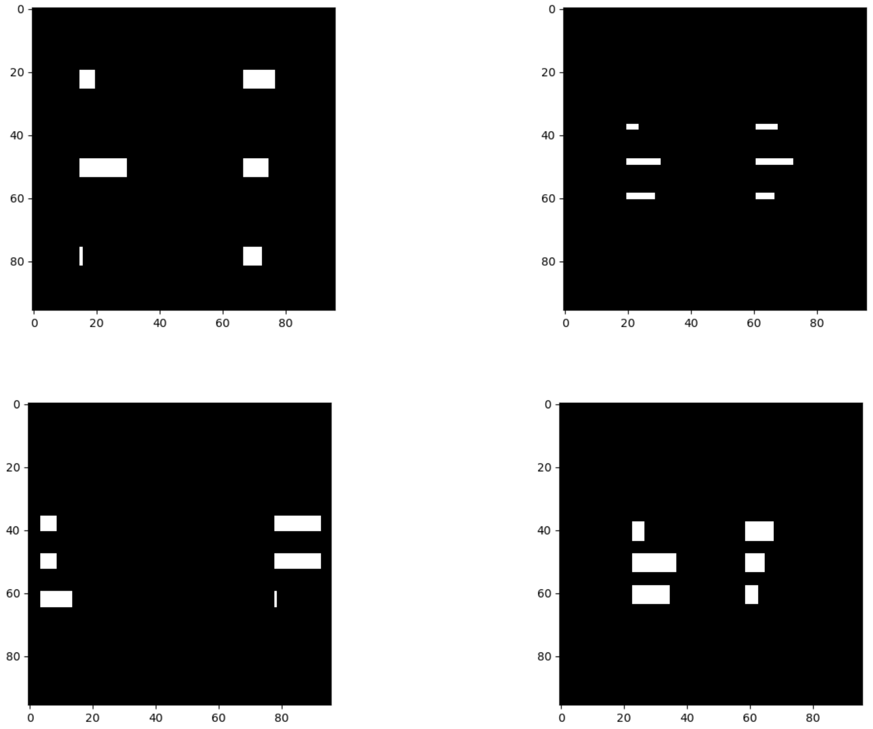


Fig. 3.14 Examples of training images. There are 3 random factors: the gap between dimensions, the gap between options and the height of the bars.

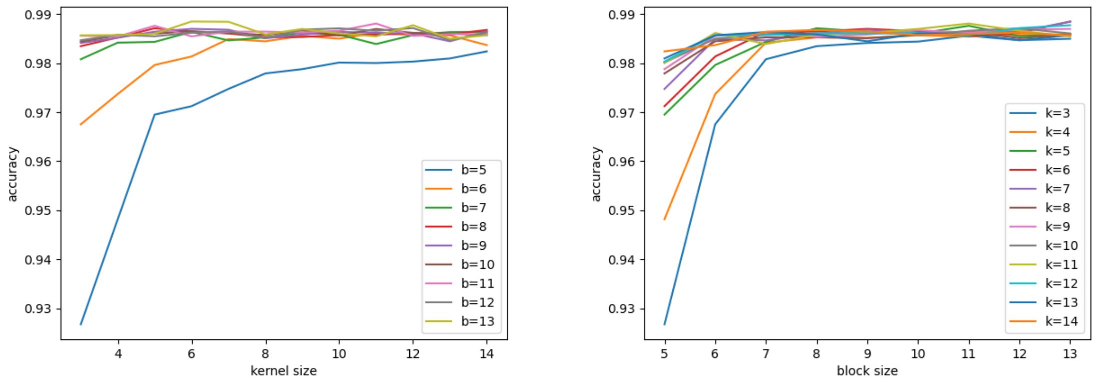
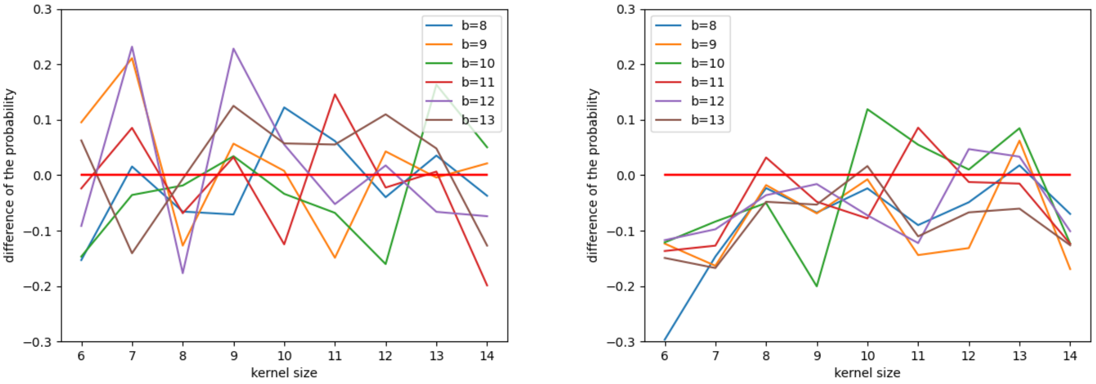


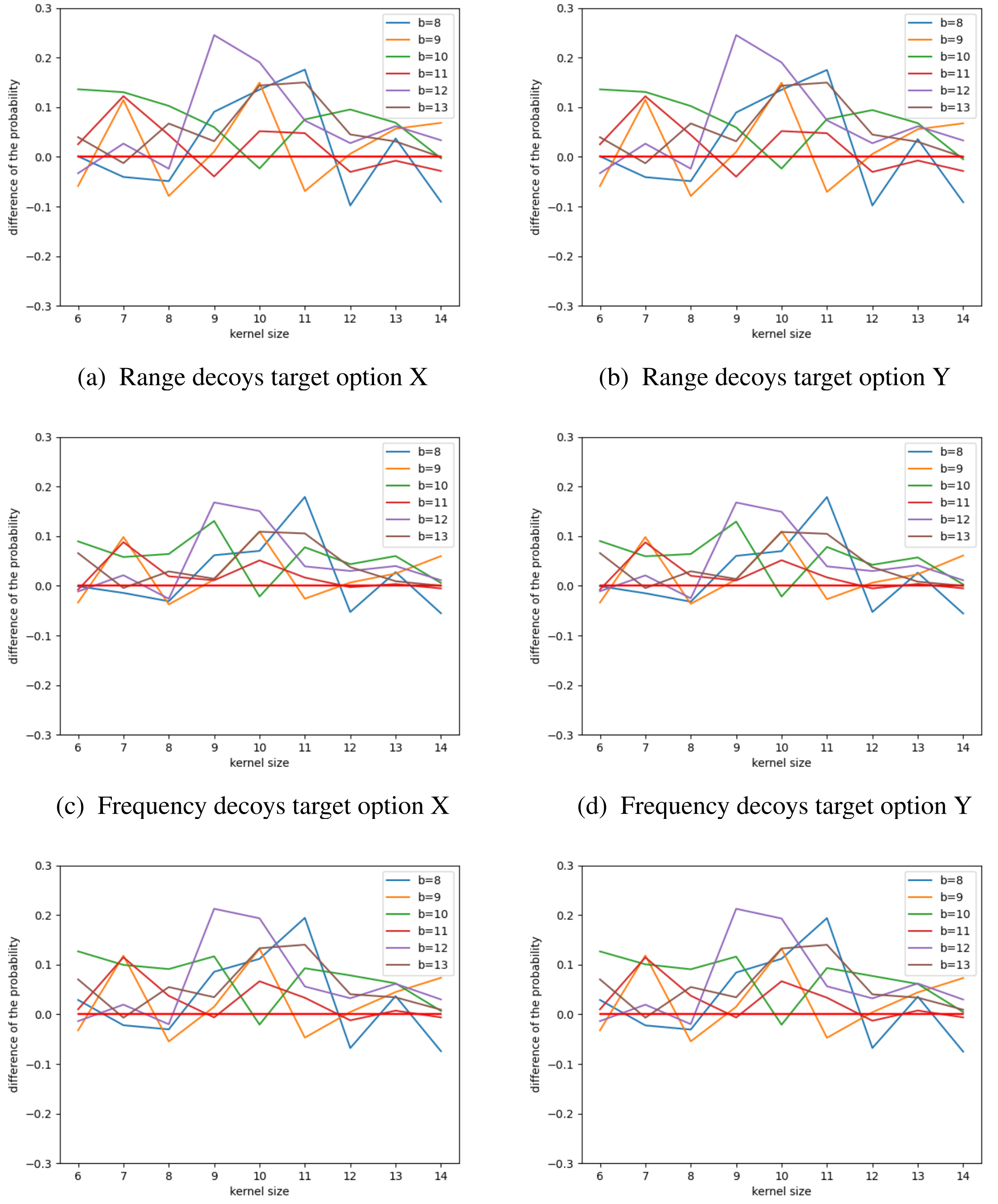
Fig. 3.15 The accuracy of the models against the kernel size and block size. The results are the average classification accuracy on the different 5 runs.



(a) Compromise decoys target option X

(b) Compromise decoys target option Y

Fig. 3.16 Sizes of compromise effect across different models. For each architecture, I train 5 models on randomly sampled data subsets. Then I evaluate the models on contextual choice tasks. A dot represents the average of these five runs. Furthermore, the line connects the dots. Each dot in a line denotes the difference in the choice proportion between the target and the competitor. In other words, it represents the size of the effect. It indicates a null compromise effect when the dot is plotted around the red line. Otherwise, the model displays a classic compromise effect when the dot is positioned significantly above the red line; a reversed compromise effect when the dot is below the red line.



(a) Range decoys target option X

(b) Range decoys target option Y

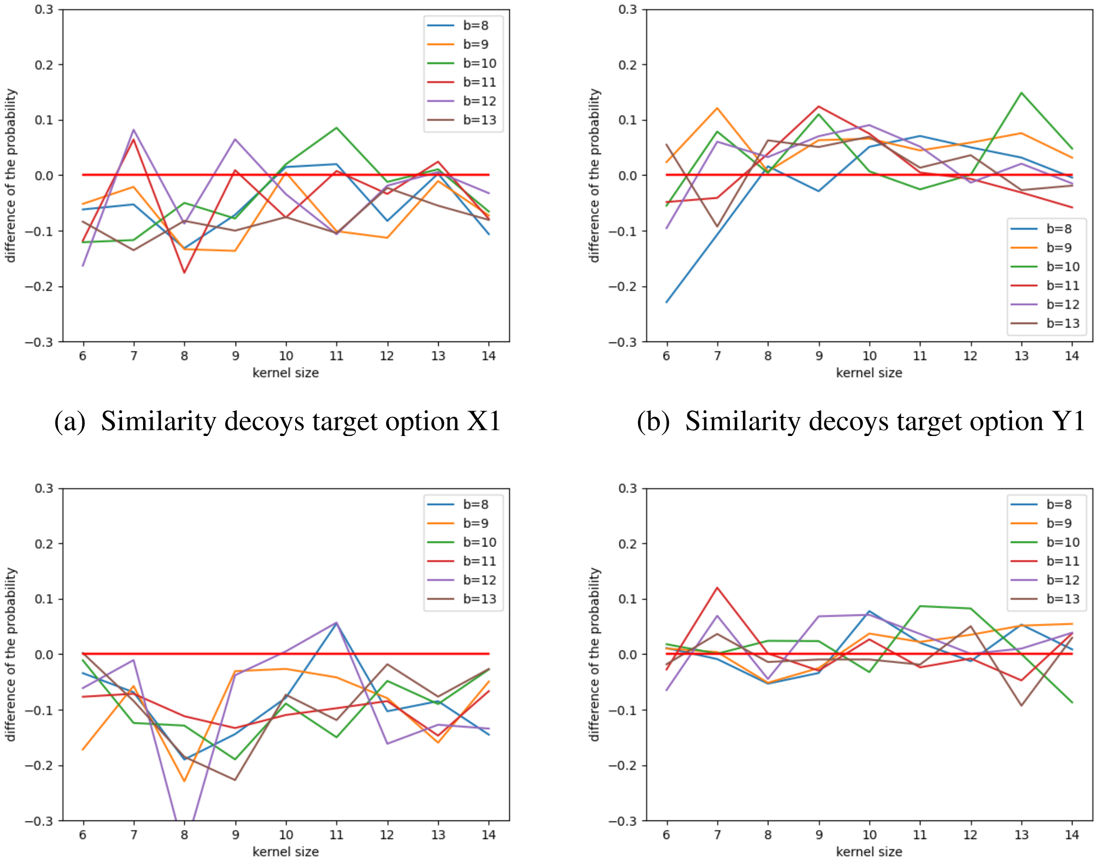
(c) Frequency decoys target option X

(d) Frequency decoys target option Y

(e) Range-frequency decoys target option X

(f) Range-frequency decoys target option Y

Fig. 3.17 Sizes of attraction effect across different models. Each dot is the average for 5 runs. It indicates a null attraction effect when the dot is plotted around the red line. Otherwise, the model displays a classic attraction effect when the dot is positioned significantly above the red line; a reversed attraction effect when the dot is below the red line.



(a) Similarity decoys target option X1

(b) Similarity decoys target option Y1

(c) Similarity decoys target option X2

(d) Similarity decoys target option Y2

Fig. 3.18 Sizes of similarity effect across different models. Each dot is the average for 5 runs. It indicates a null similarity effect when the dot is plotted around the red line. Otherwise, the model displays a classic similarity effect when the dot is positioned significantly above the red line; a reversed similarity effect when the dot is below the red line.

3.4.4 Predictions with a Specific Designed Neural Network

In this Section, I used the Keras Tuner[3] to perform hypertuning for fitting human performance reported in Cataldo and Cohen (2019). The Bayesian optimization algorithm is used to search for the model hyperparameters (i.e., kernel size and block size) and the algorithm hyperparameters (i.e., the learning rate and dropout rate). The learning curves of the resulting model are shown in Figure 3.19.

[3] https://github.com/keras-team/keras-tuner

The attraction, compromise, and similarity results are plotted in Figure 3.20. The top panel shows the results for the by-dimension condition. This is where it is easier for human participants to compare values and also where human participants exhibit attraction, compromise, and similarity effects. As can be seen in the plot, the model clearly predicts these findings. As previously stated, the qualitative effects are not a consequence of data fitting but rather of being trained to generate utility maximizing responses.

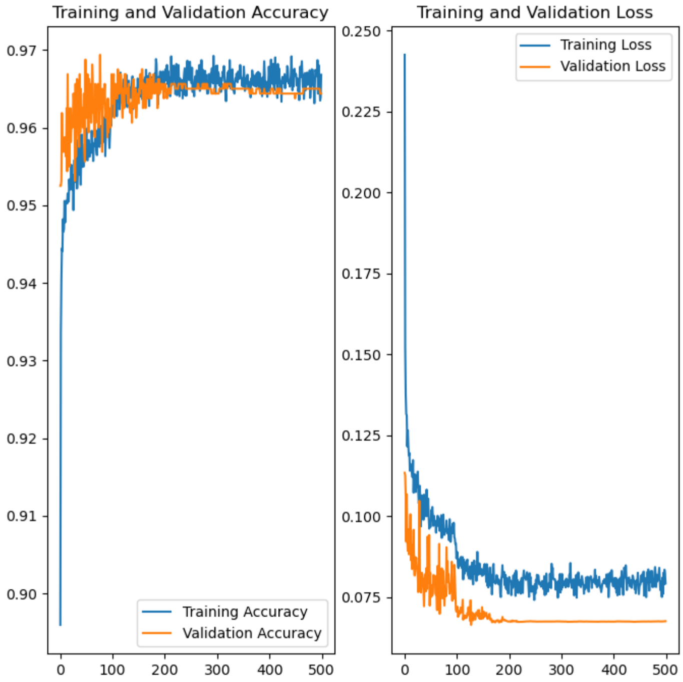


Fig. 3.19 The learning curves and the loss of the specific designed neural network in the training process.

In the bottom panel of Figure 3.20 the results of the model's performance on the by-alternative plots are presented. Here, any effect of the condition is extremely small (much smaller than humans) and/or in the wrong direction. The difference in effect between the top and bottom panels mirrors the effect of conditions on humans.

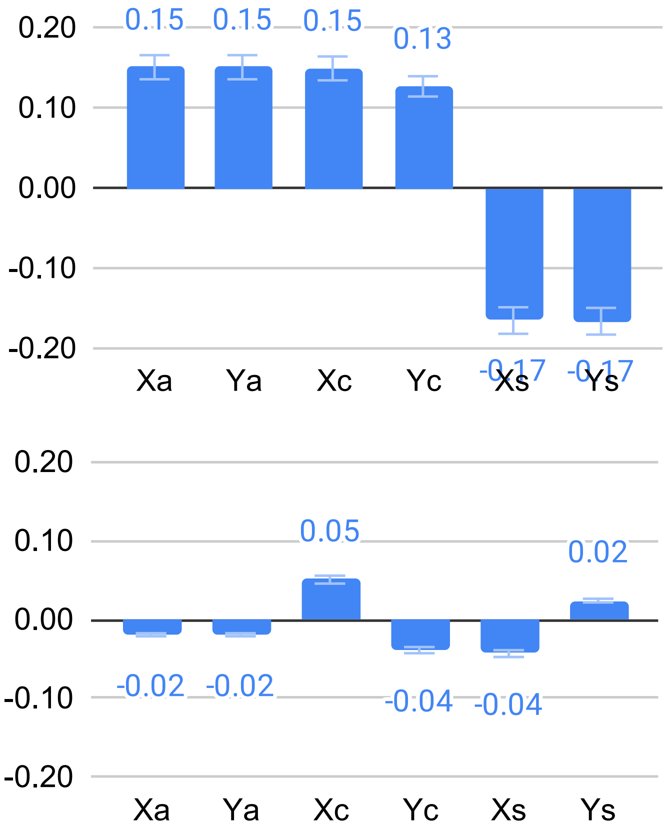


Fig. 3.20 The relative proportion of choices against conditions (a - attraction, c - compromise, s - similarity). The specifically designed model makes these predictions. The example of the choice set is displayed by-dimension (top) and by-alternative (bottom). The performance of the model was averaged over 10 runs. The error bars indicate (95%) confidence intervals. The results fit the human data reported in Fig.2 by Cataldo and Cohen (2019).

3.5 Discussion

The results presented in this Chapter show that CNN models produce the context effect when they are trained to choose the best option. Moreover, the presentation format matters. When the information is presented in a way that for humans would facilitate comparison, the CNNs generate contextual choice effects, but when comparison is not facilitated then contextual choice effects do not emerge.

I applied a number of CNN models to perceptual choice problems. To the best of my knowledge, these studies are the first to apply convolution neural networks to model human perceptual decision making. The majority of the models predicted all three major contextual choice effects (attraction, compromise, and similarity) both when options were represented as rectangles and when they were represented as bars.

From the perspective of computation rationality, flexible comparison strategies adapt to the representation format of the choice set. Nevertheless, it corresponds to the core assumption that human decision making could be generated by an optimal cognitive process adapted to the environment and cognitive constraints.

The results also show that context effects emerge for bars but not for randomly position rectangles. This may be because option representations support comparisons, although I have not analysed how the convolution network performs comparison. This behaviour is supported by a number of human experiments (Cataldo and Cohen, 2018, 2019).

An important point to note is that the models reported in this Chapter are trained to optimise an objective function which involves choosing between the available options to maximize the number of best choices (accuracy). The trained networks generated contextual choice robustly and the effects are stable after initial training (See Figure 3.3). This is evidence against the idea that contextual choice effects are irrational. Instead, the effects emerge in a network that is trained to optimise its objective.

Furthermore, our results show that the model could achieve higher accuracy through flexible information processing. The higher accuracy in choosing the options with highest expected value could be achieved through optimal underlying information processing which is adapted to the image representations of the scenario/environment. The logic is that: in order to be more accurate, the model learns to use optimal information processing with specific representations.

It is worth noting that the modeling approach taken in this thesis does not require big data. The method used to explain human behaviour has not involved training a model on large amounts of human data and then deriving explanations from the structure of the model. Rather, the models in this Chapter (and others) are trained on the problem faced by people and the behaviour of the model is then compared to the behaviour of people.

However, there are still difficulties interpreting the models reported in this Chapter. These stem from the well known difficulty of interpreting the weights in a neural network. While the CNNs above generate behaviour that corresponds to human behaviour, it is hard to interpret why. What do the CNNs tell us about the structure of information processing in the human mind? The underlying information processing mechanisms are central to understanding human cognition but artificial neural networks are mute to the cognitive process. Therefore, in the next Chapter I develop a set of models of these underlying processes and fit them to human data to explicate the psychological processes underlying choice.

One future avenue worth exploring might be the use of Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005; Felzenszwalb et al., 2009) as input. Then make the comparison of the performance between HOG and CNN. We assume that the CNN model could do better on the accuracy. But the HOG model could also model the perceptual choice tasks. Another way forward would be to test a more neurally plausible (human brain-like) network (Kubilius et al., 2019; Schrimpf et al., 2020a,b).

Chapter 4

A Deep Reinforcement Learning Model of Preference Reversals

Howes et al. (2016) reported an analysis showing that choice reversals are rational under reasonable assumptions concerning human cognition. According to this account, observed phenomena are a consequence of optimal adaptation to the bounds on cognitive limitations and the environment. However, although their model predicted the outcome and choice patterns accurately, Howes et al. (2009) does not describe the choice process. In this chapter I report a computationally rational sequential decision making model that not only predicts the choice outcomes but also predicts the process of decision-making; outcomes are a result of a decision making process that involves an integrated, sequential process of gathering information and choosing.

4.1 Introduction

As I have said in previous Chapters, recent research has begun to show that people may exhibit rationality more often than supposed (Braunlich and Love, 2022; Chen, 2015; Chen et al., 2017; Frazier and Angela, 2008; Howes et al., 2016; Juechems et al., 2021; Lewis et al., 2014; Lieder and Griffiths, 2019; Todd and Gigerenzer, 2012; Tsetsos et al., 2016). In response, I present a normative decision-making model for contextual choice tasks based

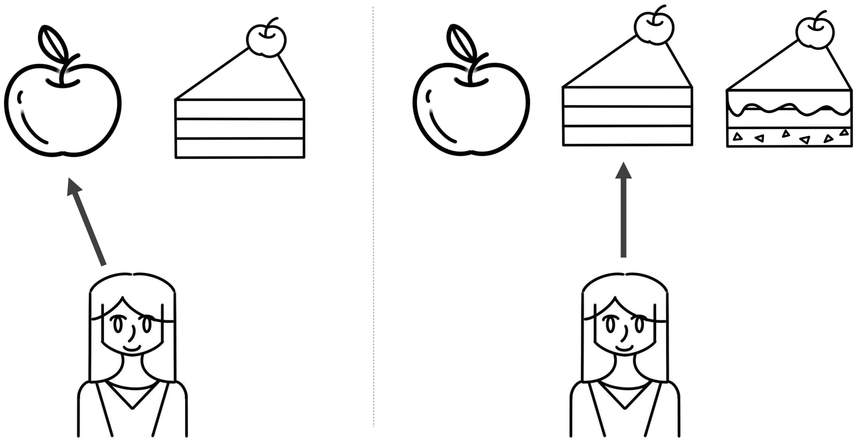


Fig. 4.1 An example of a contextual choice effect. If a person chooses an apple over a cake on the grounds of health, but then chooses the same cake when the choice is between an apple, the cake and another cake with extra sugar, then the clearly inferior (on health grounds) "cake with extra sugar" has influenced the choice between two superior alternatives.

on POMDPs, which provides a unifying framework for modelling various fundamental cognitive components of human decision making required to explain contextual preference reversal. The model is inspired by the demonstration (Howes et al., 2016) that apparent irrationalities of choice can emerge from rational processing. The approach is an application of computational rationality (Griffiths et al., 2015; Howes et al., 2009; Lewis et al., 2014) to the problem of human decision making. It extends Howes et al. (2016) by modeling contextual choice tasks as sequential decision problems and formulating them as POMDPs. Previous work by (Dayan and Daw, 2008; Frazier and Angela, 2008; Howes et al., 2018; Rao, 2010) and others has established the value of POMDPs and related formalisms for modeling humans.

In this Chapter, a Reinforcement Learning (RL) agent, designed to solve a POMDP, acquires a sequential decision policy that chooses what information to gather about which options, calculates option values, and makes comparisons between options as the unfolding task demands. The agent is trained and tested on sampled choices between three gambles, each expressed as a probability and a value. It learns the relative value of (1) noisy calculation of option values (e.g., by multiplication of a probability by a value), (2) noisy comparisons (e.g., comparing two probabilities to see which option is riskier), and (3) acting (making a choice). The agent is not pre-programmed to gather all information but learns to gather only that information that helps it maximize utility. I contrast the policies acquired by this agent to other simpler agents and show that the human-inspired agent performs better (achieves higher cumulative reward) than an agent that makes independent assessments of each option value, replicating the results of Howes et al. (2016) but in the POMDP setting.

The contribution of this Chapter is:

- A computationally rational model of contextual choice formulated as a POMDP. The model shows that preference reversals are a consequence of rational policies that prefer higher value policies. To avoid confusion, it is also important to say that the model is not a model of human learning processes. It is a model of the emergent sequential decision policy.
- Novel predictions concerning optimal sequential information gathering in contextual choice tasks. In particular, the model shows how the ratio of option comparisons and expected value calculations is influenced by the level of uncertainty in the observation function.
- An extension to the analysis of Howes et al. (2016) that accounts for the impact of sequential information gathering costs on contextual choice.
- Advancing a general understanding of how rationality, uncertainty, and apparent biases are connected. These connections are critical to the future of AI systems that work with people.

4.2 Contextual Choice as a POMDP

Unlike existing models of the contextual choice task, this Chapter presents a normative decision-making model based on POMDPs, which provides a unifying framework for modelling various cognitive components of human decision making including noisy evidence accumulation, reward maximization, costs and rewards of actions, uncertainty evaluation, etc.

I view contextual choice tasks as sequential decision making problems and formulate them as POMDPs that include, in the action space, comparison actions to assess choice option values. Given this formulation, I use a deep reinforcement learning model to discover an approximately optimal choice policy and demonstrate its capacity to simultaneously maximize reward and model humans. A crucial property of the model is that gathering information is costly, so that more information costs more but also increases the probability of a better, more rewarding, choice.

I start with a standard definition of a POMDP as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\mathcal{O}$ is the observation space. At each time step t the agent is in the latent state $s_t \in \mathcal{S}$, which is not directly observable to the agent. When the agent executes an action $a_t \in \mathcal{A}$, the state of the process changes stochastically according to the transition distribution, $s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t, a_t)$. Then, to gather information about the state, the agent makes a partial observation $o_{t+1} \in \mathcal{O}$ according to the distribution $o_{t+1} \sim \mathcal{Z}(o_{t+1}|s_{t+1}, a_t)$. The agent receives a reward $r_{t+1} \in \mathcal{R}$ according to the distribution $r_{t+1} \sim \mathcal{R}(o_{t+1}|s_{t+1}, a_t)$ after performing an action a_t to take the agent to a particular state s_{t+1}. The agent must rely on its observations to inform action selection since the hidden states are not directly observable. In each time step t, the agent acts according to its policy $\pi(a_t|h_t)$ which returns the probability of executing action a_t, and where $h_t = (o_0, a_0, o_1, a_1, \cdots o_{t-1}, a_{t-1})$ are the histories of observation-actions pairs. The goal of the agent is to learn an optimal policy π^* that maximizes the expected cumulative rewards, $\pi^* = \underset{\pi}{argmax}\, \mathbb{E}\left[\sum_{t=1}^{t=T} \gamma^{t-1} r_t\right]$, where $0 < \gamma < 1$ is the discount factor.

Each choice task had 3 options (X, Y, Z) which were represented with two attributes: a randomly sampled probability p and a randomly sampled value v. I assumed that probabilities p were sampled from a $\beta-$distribution and values v were sampled from a $t-$distribution. These distributions represented the ecological distributions experienced by participants in the human behaviour experiments reported by Wedell (1991). I view contextual choice tasks as sequential decision making problems and formulate them as POMDPs as follows.

The state space $\mathcal{S}$ for each task was generated from a sampled choice task. More formally, a state was $\{(p_X, v_X), (p_Y, v_Y), (p_Z, v_Z)\}$, where probabilities p were sampled from a $\beta-$distribution and values v were sampled from a $t-$distribution. The agent selected actions from a set $\mathcal{A}$ which included 6 comparison actions (e.g., compute the comparison relation for p_X and p_Y), 3 calculation actions (e.g., compute the expected value for X given p_X and v_X) and the 3 choice actions (choose X, choose Y, choose Z). The reward for comparison and calculation actions was negative c. The reward for a choice action was 10 if the agent chose the option with maximum expected value, otherwise, it was -10.

There was therefore a trade-off between the cost of information gathering and choice accuracy. More information cost more but was more likely to lead to a better response and therefore a higher reward. As a consequence of the selected action, the subsequent observation $o_{t+1} \in \mathcal{O}$ was of computing the most recent comparison or calculation with noise. Following Howes et al. (2016), each observation of a comparison had 4 possible outcomes, which indicated that the relation was unknown, greater, equal and less. These are:

$$f(m_i, m_j) = \begin{cases} none, & unknown \\ >, & m_i > m_j + \tau_m \\ \equiv, & |m_i - m_j| \leq \tau_m \\ <, & m_i < m_j - \tau_m \end{cases} \tag{4.1}$$

The function f represents this pairwise order relation between the two values or two probabilities of two gambles. The magnitude $m \in \{v, p\}$ is the magnitude of value or probability. The relation is defined as equal if their magnitudes m are within their corresponding tolerance τ_m. The probability of comparison error $P(error_f)$ was the probability that the relations were sampled uniformly random from the comparison set $O = \{>, \equiv, <\}$.

$$O = \{f(p_A, p_B), f(p_A, p_D), f(p_B, p_D), f(v_A, v_B), f(v_A, v_D), f(v_B, v_D)\}$$

The observation of a calculation was computed using:

$$E_i = p_i^{\alpha} \times v_i + \varepsilon \qquad \varepsilon \sim N(0, \sigma_{calc}^2) \tag{4.2}$$

where the probability p was weighted by an exponential parameter α. The purpose of using parameter α was to model **subjective probability** following Savage (1972), which is used extensively in econometrics because it is mathematically well behaved.

The evidence state is the history of the partial and noisy observation of the latent state. The history of observation set $\mathcal{O}_h$ is the noisy encoding of the partial orderings of probabilities and values:

$$\mathcal{O}_h = \{f(p_X, p_Y), f(p_X, p_Z), f(p_Y, p_Z), f(v_X, v_Y), f(v_X, v_Z), f(v_Y, v_Z), E_X, E_Y, E_Z\} \tag{4.3}$$

The observation o_t is the same as the latent state s_t, when there is no noisy observation: the tolerance $\tau_m = 0$, the exponential parameter $\alpha = 1$, the probability of comparison error $P(error_f) = 0$ and the calculation noise $\sigma_{calc} = 0$.

It is intractable to compute a policy to solve the defined POMDP, but it is possible to approximate the optimum through learning (Cushman and Morris, 2015; Igl et al., 2018; Wang et al., 2018). I solve the POMDP by casting it as a Markov Decision Process (MDP) whose state space is the history of observation o_h. I used a deep reinforcement learning method, called ACER, to find an approximately optimal policy for the POMDP (Wang et al., 2016a). For all reported experiments, I built the environments within OpenAI Gym (Brockman et al., 2016) and used the OpenAI Baselines [1] implementation of the deep RL algorithms.

4.3 Wedell Results

In order to test the model, I designed three different agents: The **integrated agent** could use both calculation and comparison selectively. States represent the results of calculation

[1] https://github.com/openai/baselines

and comparison actions. The model can learn which observations are useful and not every observation needs to be made. There is no explicit integration of comparison and calculation. Instead, the results of comparison and calculation accumulate in the history and choice action values are conditional on these histories. The **comparison-only agent** was the same as the integrated agent but could only use comparison actions, and states only represented the comparison information. The **calculation-only agent** was the same as the integrated agent but could only use calculation actions, and states only represented the calculation information. The important difference between the three models was the availability of the different kinds of observation. All three agents learnt approximately optimal policies from experience given the bounds imposed by these difference observation capacities.

In what follows, I first show that the new reinforcement learning model replicates previous findings (Howes et al., 2016) and then show that it makes new predictions derived from the sequential nature of the model.

First, I investigate the economic value of using contextual information in an environment where uncertainty is introduced through the noise in the partial observation of the state. The analysis shows the benefit of using comparisons in noisy and costly environments. Second, I show that the human-like agent can generate three critical types of contextual effect and fits well with human behavioural data. Third, I report the impact of noise on the agent's choice behaviour. Fourth, I look into the information gathering process of the learnt policy. The analysis shows that the uncertainty of information influences the decision process and leads to selectivity between comparison and calculation, which can help make decision making more efficient.

4.3.1 Is it Beneficial to Compare Options?

In order to answer this question, I first fitted the distributions of the environment to those used in a prominent human experiment (Wedell, 1991). The probabilities p are sampled from a $\beta-$distribution $(a = 1, b = 1)$ and the values v are sampled from a $t-$distribution $(location = 19.60, scale = 5, degree\ of\ freedom = 100)$. For all the experiments below, I used the same distributions. Reported results are averaged over 10 runs, each with a different seed, after training on 3 million samples.

We built the environments within OpenAI Gym (Brockman et al., 2016) and used the OpenAI Baselines [2] implementation of the deep RL algorithms: ACER. Hyperparameters were as follows:

- Policy Network: (64, tanh, 64, tanh, Linear) + Standard Deviation variable; Value Network (64, elu, 64, elu, linear)
- Number of timesteps = 10 million
- Batch size = 1024
- Learning rate for RMSProp = 7e-4
- Schedule of learning rate: linear annealed
- Actor-learner threads = 32
- Other hyperparameters are default as in OpenAI Baselines implementation

All agents were tested with different levels of observation noise and the resulting performance is shown in Figure 4.2. The maximum expected value that could be achieved by any agent was 16.29 (horizontal upper bound in Figure 4.2), which was calculated by averaging the maximum expected value of 3 options across 1 million choice sets sampled from the above distributions.

In Figure 4.2 it can be seen that the integrated agent, using both calculation and comparison observations, can approximate the optimal policy when actions could be conducted without noise. Also, calculation-based and comparison-based agents are able to perform close to optimum when there is no noise. However, the noise has a negative effect on the performance of all types of agent. The average obtained value of choices decreases as noise increases.

[2] https://github.com/openai/baselines

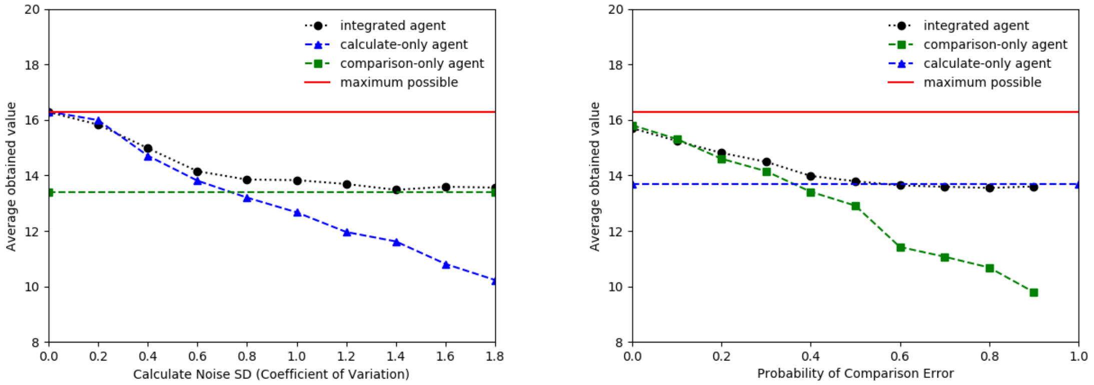


Fig. 4.2 The mean expected value obtained by agents with different levels of noise: the coefficient of variation for the calculation noise (left panel) and the probability of comparison error (right panel). In the left panel, the comparison noise is fixed at $P_{error} = 0.3$. In the right panel, the calculation noise is fixed at $\delta_{calc} = 30$, corresponding to a coefficient of variation is 0.3. Results for 3 types of agent are presented in each panel: the comparison-only agent (green-doted line), calculation-only agent (blue-doted line) and integrated agent (black-doted line). This Figure replicates Figure 3 in Howes et al. (2016).

Figure 4.2 also shows that the integrated agent combines the strengths of both noisy comparison and noisy calculation to make better decisions than the other agents in all noise conditions. The average expected value of the choices made by the integrated agent is greater than the other agents. In other words, the human-like integrated agent performs better in accumulating reward than the agent that makes independent assessments of each option value. The results suggest that when there is observation uncertainty, both humans and artificial agents will gain higher reward if they compare options, rather than merely evaluate each option independently.

These results also demonstrate that the comparison-only agents learned an optimal policy in the choice task by trying to maximise the expected value. Furthermore, the human-inspired integrated model learns to make use of all available and valuable information to obtain higher expected rewards in the situation of uncertainty. Finally, it demonstrates that the resulting policy is a computationally rational process.

4.3.2 Does the Integrated Agent Predict Human Performance?

To determine whether the integrated agent (the agent that uses both comparison and calculation) predicts human performance, I measured its behaviour on the the attraction, compromise and similarity tasks. The human behaviour on these tasks is shown in Figure 2.1d. I used one fixed setting of the agent policy and parameter values. Following this common rule, I use the proposed model to produce all the three context effects and one specific account of the attraction effects observed in human psychological experiments.

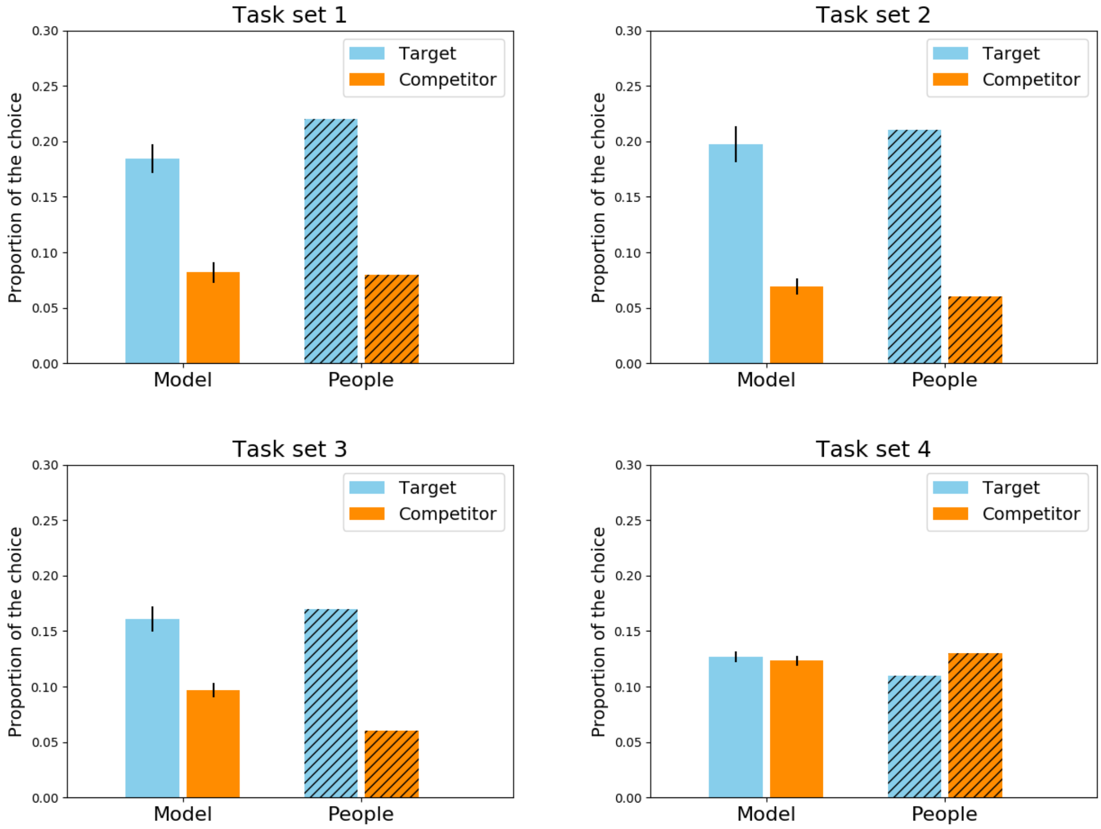


Fig. 4.3 The integrated agent exhibits the attraction effect. A sample of agents was tested on each of four variants of the attraction effect task (in which the decoy is in slightly different positions). People and agent exhibit more target choices than Competitor choices in task sets 1, 2, and 3. As expected, neither the integrated agent, nor people, exhibit the effect in task set 4 where the decoy was not dominated by only one of the options and was therefore in a neutral position. Task 4 thereby acts as a control The human data is from Wedell (1991). The results are averaged over 10 runs with different seeds. The error bars indicate confidence interval (95%) of the predictions made by the agent. This Figure replicates Figure 8 in Howes et al. (2016).

Agents were trained on tasks which were randomly sampled from a $\beta-$distribution ($a = 1, b = 1$) for the probability p and $t-$distribution ($location = 19.60, scale = 8.08, df = 100$) for the value v. After 3 million training samples, the agent converged and demonstrated stable performance.

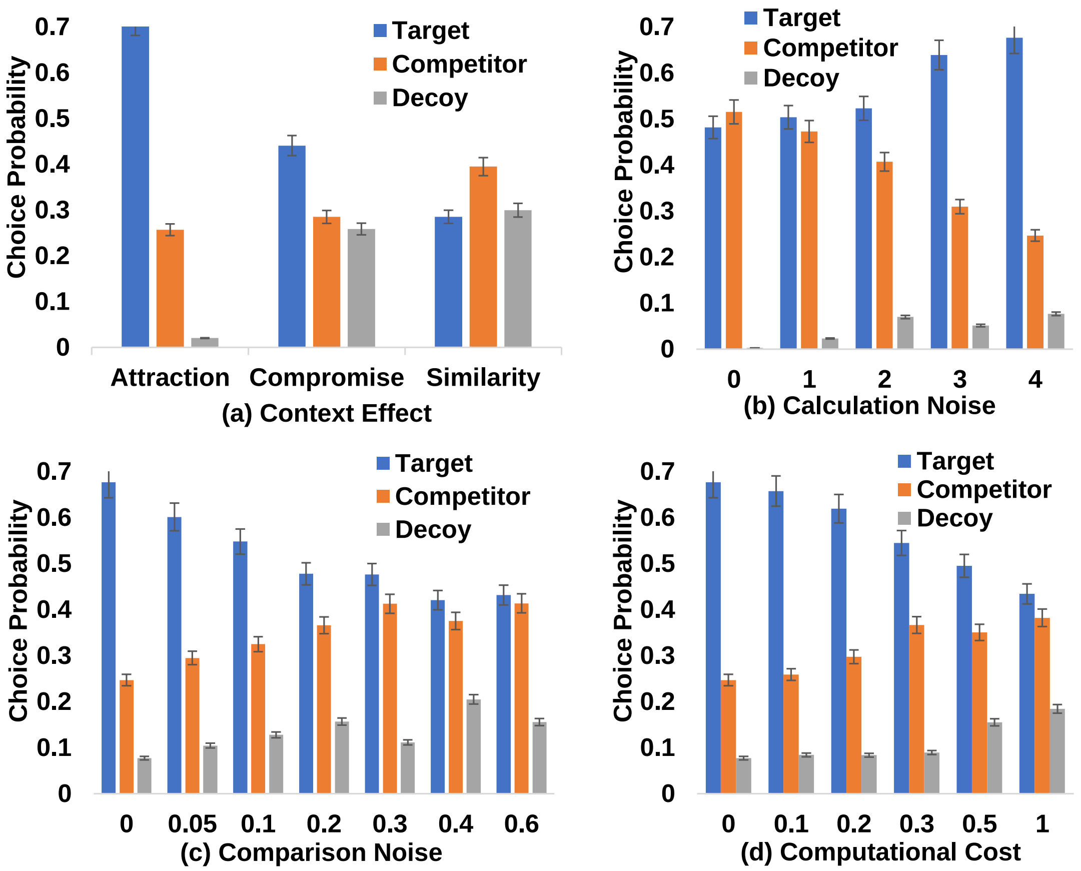
The agent was repeatedly trained with adjusted values of the comparison noise, calculation noise, probability weighting parameters, cost of comparisons and calculation cost until the qualitative effects fitted the human performance (Trueblood (2012); Figure 2.1d). The fitted parameter values were: calculation noise $\sigma_{calc} = 4$, comparison error $P(error_f) = 0.1$, probability weighting parameters $\alpha = 1$, the perceived cost of comparison $C_{comparison} = -0.01$ and the calculation cost $C_{calc} = -0.1$. I do not claim to have achieved the best possible fit, nor a better fit than other models. The point of the fit was to show that the qualitative effects exhibited by humans was within the space of behaviours generated by the agent.

The results are averaged over 10 runs with different seeds and shown in Figure 4.4a. It shows that the agent generates the three context effects using one learnt policy and one fixed set of parameter values. Comparison of Figure 4.4a to Figure 2.1d) shows that all of the qualitative effects are predicted.

To further test the agent I fitted it to variations of the attraction effect in human performance (Wedell, 1991). The fitted values were: calculation noise $\sigma_{calc} = 0.50$, comparison error $P(error_f) = 0.1$, probability weighting parameters $\alpha = 1.5$, the perceive cost of comparison features $C_{comparison} = -0.01$ and the calculation cost $C_{calc} = -0.1$.

In total, 80 models were trained and tested. I tested the accuracy of each model using 10000 random sampled choice tasks. The results of 11 models were removed since their average accuracy was a very low 0.80, and were outliers from the others which had an average accuracy of 0.94. The 80 remaining models were tested on 50000 × 8 × 10 choice sets from the experiment.

The results in Figure 4.3 show that for both agents and people, the Target is selected more often than the Competitor in three of the task sets (1, 2, and 3). In contrast, and as expected, the Target and Competitor are selected equally often in the 4th task set by both agents and people. The decoy was positioned in a neutral position in task set 4 and does not therefore have an effect on the target choice rate.



(a) Context Effect

(b) Calculation Noise

(c) Comparison Noise

(d) Computational Cost

Fig. 4.4 (a) The behaviour of the integrated agent for 3 types of context effect: attraction, compromise and similarity. (b)(c)(d) The effect of noise and computational cost on the contextual choice effect exhibited by the integrated agent. (b) increased calculation noise increases the effect size, (c) Increased comparison noise reduces the effect size, and (d) increased computational cost reduces the effect size. The results are averaged over 10 runs with different seeds, and the error bars indicate (95%) confidence interval.

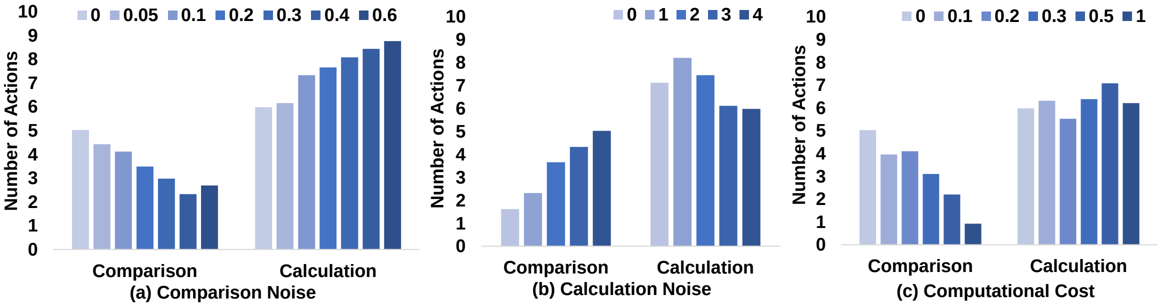
As I have analysed in the last section, the solution to the POMDP is approximately optimal. Moreover, it demonstrates that the context effect can emerge from approximately computationally rational processes. In the next section, I will investigate how the learning process is adaptive to computational cost and cognitive limits.

4.3.3 Does the Uncertainty of Information Influence the Decision Process?

The nature of the context effect is difficult to comprehend since existing cognitive process models are based on a variety of assumptions. They also offer a variety of explanations for the phenomenon. I analyse the processing of the context effect by testing the impact of potential influential factors. In this section, I'll experiment with various factors that affect the size of preference reversals in the choice task, such as computational cost and perceptual noise.

I tested the consequences of noise on choice. The results in Figure 4.4b, c, d show that: (1) The size of the attraction effect decreases as computational cost increases, (2) the attraction effect is weaker when the agent's accuracy of comparison is diminished with noise, (3) The effect is stronger when calculation noise is higher. While there is no human data that directly tests the effect of noise, a number of studies report that the rate of the context effect diminishes as time pressure increases (Pettibone, 2012; Trueblood et al., 2014). As shown in Figure 4.4b, c, d, the effects of time pressure on humans is consistent with the effect of increased noise in the model.

4.3.4 What are the Effects of Noise and Computation Cost?



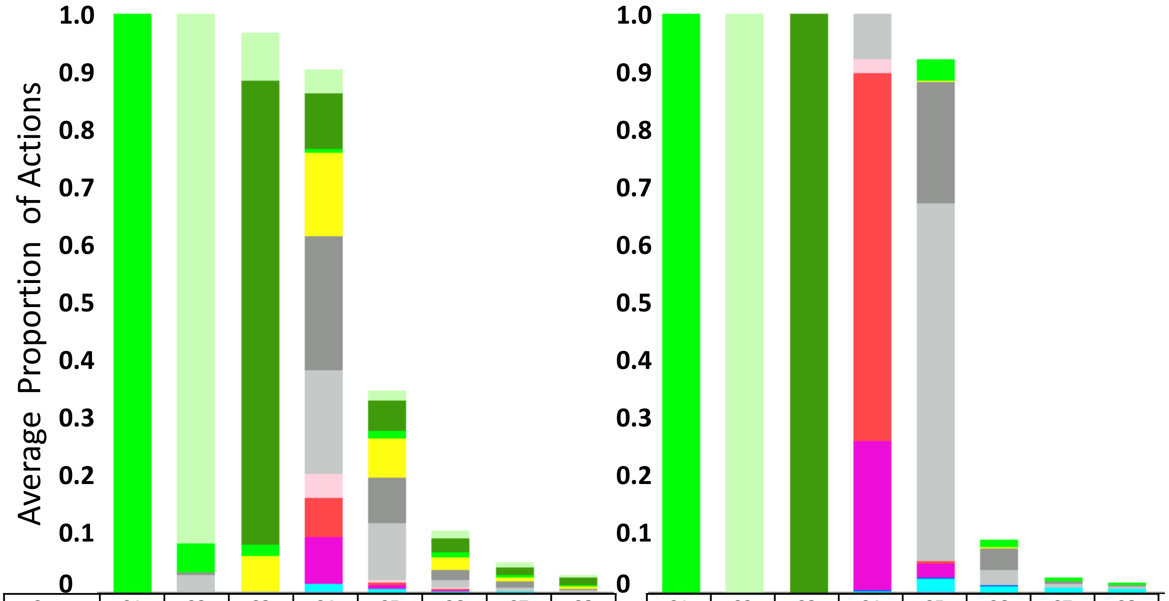
(a) Comparison Noise

(b) Calculation Noise

(c) Computational Cost

Fig. 4.5 Effect of noise and computational cost on number of comparison and calculation actions taken. The analysis shows that the model chooses observations selectively depending on their utility.

The effect of noise on the number of comparisons and calculation actions taken is shown in Figure 4.5. Increases in comparison noise leads to a selective reduction in the use of comparison and a selective increase in the use of calculation. Conversely, increases in calculation noise leads to a selective decrease in the use of calculation and an increase in the use of comparison. Increase in the cost of information gathering actions (comparison and expected value) reduces contextual effects on choice (Figure 4.5c) as less information is gathered.



Step	S1	S2	S3	S4	S5	S6	S7	S8	S1	S2	S3	S4	S5	S6	S7	S8
Visit	100	100	96	90	35	10	5	3	100	100	100	100	92	9	2	1
CD	0	0.91	0.08	0.04	0.02	0.01	0.01	0	0	1	0	0	0	0	0	0
CB	0	0	0.80	0.10	0.05	0.02	0.01	0.01	0	0	1	0	0	0	0	0
CA	1	0.05	0.02	0.01	0.01	0.01	0	0	1	0	0	0	0.05	0.01	0.01	0
D	0	0	0.06	0.14	0.07	0.02	0.01	0.01	0	0	0	0	0	0	0	0
B	0	0	0	0.23	0.08	0.02	0.01	0	0	0	0	0	0.21	0.04	0	0
A	0	0.04	0	0.18	0.10	0.02	0.01	0.01	0	0	0	0.08	0.62	0.03	0.01	0
VB?VD	0	0	0	0.04	0	0	0	0	0	0	0	0.03	0	0	0	0
VA?VD	0	0	0	0.07	0	0	0	0	0	0	0	0.63	0	0	0	0
VA?VB	0	0	0	0.08	0.01	0	0	0	0	0	0	0.26	0.02	0	0	0
PB?PD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PA?PD	0	0	0	0.01	0.01	0	0	0	0	0	0	0	0.02	0.01	0	0.01
PA?PB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 4.6 The left panel shows the average proportion of each action type taken by the model on each step when given randomly sampled tasks. The right panel shows the average proportion of each action type when given attraction effect preference reversal tasks. Actions that calculate the expected value of A, B or D are in green; actions that compare probabilities are shown in blue; actions that compare values are shown in red; actions that choose A, B or D are shown in white, grey and yellow respectively.

4.4 Sequential Effects

4.4.1 How Does Context Affect Decision Sequence?

A novel contribution of the model is that, by virtue of the sequential decision process, it predicts how action sequences should vary with task type. Figure 4.6 contrasts the model's action sequences on random tasks (left panel) and its action sequences on preference reversal tasks (right panel). Comparing the left and right panels, I can see that the model tends to use calculations of expected value in the first three steps regardless of task type. Despite this initial similarity, the fourth action is quite different for the two task types. Here, on average, for random tasks the model tends to pick one of the options. In contrast, for preference reversal tasks, the model tends to compare values and subsequently, shows a marked preference for the high probability option (option A). This preference is not visible in the random task action sequences. This, approximately bounded optimal, prediction conflicts with authors who have argued that people prefer comparisons to calculation of expected value (Noguchi and Stewart, 2018; Ronayne and Brown, 2017; Stewart et al., 2006; Vlaev et al., 2011).

4.4.2 Decision Tree Analysis of Decision Sequence

To explore the order in which ordinal information is gathered, the comparison-only agent is used to predict the information gathering process of ordinal features when the parameters are set as $[n = 0.0, df = 100, scale = 4, location = 20]$. In order to make it simple and clear, I use the comparison-only agent to explore the sequence predictions.

The results indicate that the agent mainly compares the probability attributes between options in the first few steps, then the value attributes and finally makes a choice. The results are consistent with the observed human behaviour (Noguchi and Stewart, 2014) that pair alternatives are compared on a single attribute dimension in each choice using eye-movement data. The agent almost chooses targets in both situations, although the information gathering process is the same. It is well supported that the model could adapt to different environments.

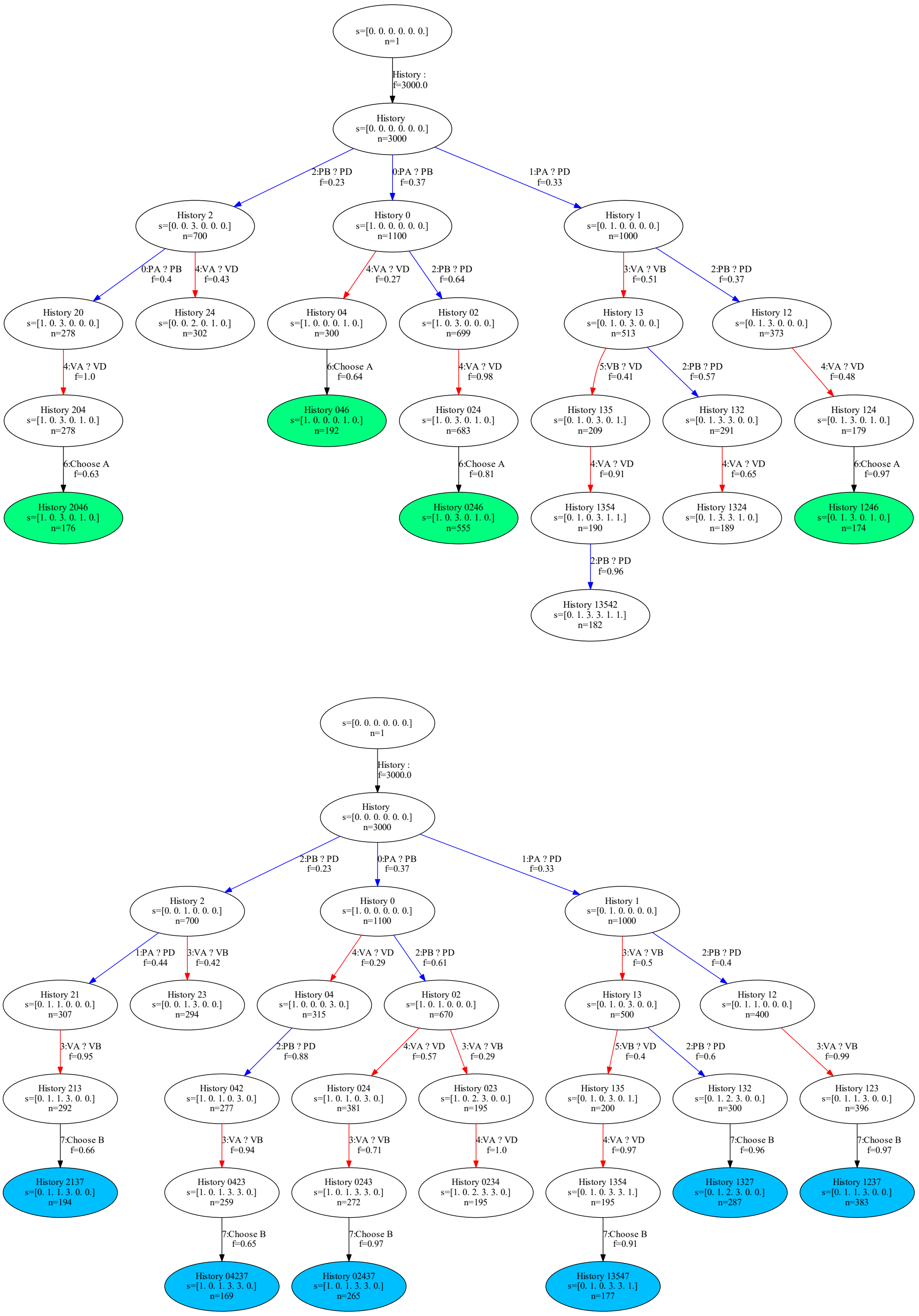

s=[0. 0. 0. 0. 0. 0.] n=1
History : f=3000.0
History s=[0. 0. 0. 0. 0. 0.] n=3000
2:PB ? PD f=0.23
0:PA ? PB f=0.37
1:PA ? PD f=0.33
History 2 s=[0. 0. 3. 0. 0. 0.] n=700
History 0 s=[1. 0. 0. 0. 0. 0.] n=1100
History 1 s=[0. 1. 0. 0. 0. 0.] n=1000
0:PA ? PB f=0.4
4:VA ? VD f=0.43
4:VA ? VD f=0.27
2:PB ? PD f=0.64
3:VA ? VB f=0.51
2:PB ? PD f=0.37
History 20 s=[1. 0. 3. 0. 0. 0.] n=278
History 24 s=[0. 0. 2. 0. 1. 0.] n=302
History 04 s=[1. 0. 0. 0. 1. 0.] n=300
History 02 s=[1. 0. 3. 0. 0. 0.] n=699
History 13 s=[0. 1. 0. 3. 0. 0.] n=513
History 12 s=[0. 1. 3. 0. 0. 0.] n=373
4:VA ? VD f=1.0
6:Choose A f=0.64
4:VA ? VD f=0.98
5:VB ? VD f=0.41
2:PB ? PD f=0.57
4:VA ? VD f=0.48
History 204 s=[1. 0. 3. 0. 1. 0.] n=278
History 046 s=[1. 0. 0. 0. 1. 0.] n=192
History 024 s=[1. 0. 3. 0. 1. 0.] n=683
History 135 s=[0. 1. 0. 3. 0. 1.] n=209
History 132 s=[0. 1. 3. 3. 0. 0.] n=291
History 124 s=[0. 1. 3. 0. 1. 0.] n=179
6:Choose A f=0.63
6:Choose A f=0.81
4:VA ? VD f=0.91
4:VA ? VD f=0.65
6:Choose A f=0.97
History 2046 s=[1. 0. 3. 0. 1. 0.] n=176
History 0246 s=[1. 0. 3. 0. 1. 0.] n=555
History 1354 s=[0. 1. 0. 3. 1. 1.] n=190
History 1324 s=[0. 1. 3. 3. 1. 0.] n=189
History 1246 s=[0. 1. 3. 0. 1. 0.] n=174
2:PB ? PD f=0.96
History 13542 s=[0. 1. 3. 3. 1. 1.] n=182
s=[0. 0. 0. 0. 0. 0.] n=1
History : f=3000.0
History s=[0. 0. 0. 0. 0. 0.] n=3000
2:PB ? PD f=0.23
0:PA ? PB f=0.37
1:PA ? PD f=0.33
History 2 s=[0. 0. 1. 0. 0. 0.] n=700
History 0 s=[1. 0. 0. 0. 0. 0.] n=1100
History 1 s=[0. 1. 0. 0. 0. 0.] n=1000
1:PA ? PD f=0.44
3:VA ? VB f=0.42
4:VA ? VD f=0.29
2:PB ? PD f=0.61
3:VA ? VB f=0.5
2:PB ? PD f=0.4
History 21 s=[0. 1. 1. 0. 0. 0.] n=307
History 23 s=[0. 0. 1. 3. 0. 0.] n=294
History 04 s=[1. 0. 0. 0. 3. 0.] n=315
History 02 s=[1. 0. 1. 0. 0. 0.] n=670
History 13 s=[0. 1. 0. 3. 0. 0.] n=500
History 12 s=[0. 1. 1. 0. 0. 0.] n=400
3:VA ? VB f=0.95
2:PB ? PD f=0.88
4:VA ? VD f=0.57
3:VA ? VB f=0.29
5:VB ? VD f=0.4
2:PB ? PD f=0.6
3:VA ? VB f=0.99
History 213 s=[0. 1. 1. 3. 0. 0.] n=292
History 042 s=[1. 0. 1. 0. 3. 0.] n=277
History 024 s=[1. 0. 1. 0. 3. 0.] n=381
History 023 s=[1. 0. 2. 3. 0. 0.] n=195
History 135 s=[0. 1. 0. 3. 0. 1.] n=200
History 132 s=[0. 1. 2. 3. 0. 0.] n=300
History 123 s=[0. 1. 1. 3. 0. 0.] n=396
7:Choose B f=0.66
3:VA ? VB f=0.94
3:VA ? VB f=0.71
4:VA ? VD f=1.0
4:VA ? VD f=0.97
7:Choose B f=0.96
7:Choose B f=0.97
History 2137 s=[0. 1. 1. 3. 0. 0.] n=194
History 0423 s=[1. 0. 1. 3. 3. 0.] n=259
History 0243 s=[1. 0. 1. 3. 3. 0.] n=272
History 0234 s=[1. 0. 2. 3. 3. 0.] n=195
History 1354 s=[0. 1. 0. 3. 3. 1.] n=195
History 1327 s=[0. 1. 2. 3. 0. 0.] n=287
History 1237 s=[0. 1. 1. 3. 0. 0.] n=383
7:Choose B f=0.65
7:Choose B f=0.97
7:Choose B f=0.91
History 04237 s=[1. 0. 1. 3. 3. 0.] n=169
History 02437 s=[1. 0. 1. 3. 3. 0.] n=265
History 13547 s=[0. 1. 0. 3. 3. 1.] n=177

Fig. 4.7 Predicted processing of choices by the comparison-only agent. Top tree: the situation in which a decoy is posited close to A. Bottom tree: the situation in which a decoy is posited close to B. The figure of the decision tree maintains the most frequent process and only keeps the states in which the visit count n is greater than 150 (5% of 3000 choice tasks). Comparisons of value attributes are presented in red arrows, and comparisons of probability attributes are presented in blue arrows. Choosing A, B or C is shown in green, blue and red.

I also built two decision trees to analyse the record of each state and action in a total of 3000 tasks. Each node represents one state of the model and each arrow represents one action taken by the model. Each node contains 3 values which are the action history from the start state, current state identifier s, and visit count n. Each action is coded by a number from 0 to 5 indicating what information was gathered or what decision was made. For example, 'History 1320456; s=[1, 1, 3, 3, 1, 1]; n=85' means that the agent has taken the following actions in order: compare the probability of A and D – '1: PA?PD', compare the value of A and B – '3: VA?VB', compare the probability of B and D – '2: PB?PD', compare the probability of A and B – '0: PA?PB', compare the value of A and D – '4: VA?VD', compare the value of B and D – '5: VB?VD', choose option A –'6: choose A'; the value of s means the order relation between the attributes; the visit count 'n=85' means this state occurs 85 times. In order to make it easy to interpret, the figure of the decision tree maintains the most frequent process. The decision tree only keeps the states of which visit count n is larger than 150 (5% of 3000 choice tasks).

Figure 4.7 shows that on average the agent took $4 \sim 5$ comparisons between probabilities or values before making a choice and then chose the option proximal to the decoy. The amount of the most efficient perceiving actions are also $4 \sim 6$ which indicates that the model could learn an optimal way to gather the information without prior knowledge. The Figure also shows that the agent uses the comparison of probability attributes, which are blue arrows, more than the comparison of value attributes, which are red arrows, in the first few actions. The comparison of value attributes is used more before making a choice.

4.5 Discussion

I have proposed a novel explanation for how apparently irrational choice might emerge as a consequence of optimal sequential decision making under uncertainty. While this is not the first work to demonstrate the rationality of the preference reversal phenomena (Howes et al., 2016), nor the first work to use POMDPs to model humans (Daw et al., 2006; Rao, 2010), it is the first to formulate the contextual choice problem as a POMDP and demonstrate that a reinforcement learning agent that uses *comparison* observations generates higher reward than an agent that only makes independent assessments of value. These comparison actions, when deployed by people, have been thought by many to lead to violations of the independence axioms and they have been shown to underpin preference reversals in humans (Noguchi and Stewart, 2014). But, as has previously been pointed out (Howes et al., 2009), this seemingly paradoxical result makes sense when it is appreciated that the comparison of options reduces the uncertainty of option values.

A different RL model of preferences reversals is reported by Spektor et al. (2019). They explain context effects in a decisions-from-experience setting, where attribute values are not explicitly stated but have to be learned over many trials. My model in contrast, is based on decisions from description, where all attribute values are fully described. Unlike my model, their model does not acquire an explicit representation of different attributes and does not make attribute-based comparisons. Instead, it models a dynamic learning process during which the feedback on similar options is compared.

By extending Howes et al. (2016) I have demonstrated that the same pattern of behaviours that are thought to be irrational in humans, emerge from a process that attempts to maximize the cumulative reward of action. My results also show that comparison actions are increasingly *preferred* by the agent as observation noise increases. In addition, I have shown that higher information gathering costs can diminish the use of comparisons and reduce the preference reversal rate, thereby extending previous analysis to account for the economics of information gathering in contextual choice tasks. In contrast to previous models, where comparisons have been assumed, my model uses them preferentially depending on the structure of the task.

My model assumes that observations can be subject to noise and this assumption is worth further discussion given how easy it seems for people to make comparisons. I make three observations. First, noise helps explain the fact that people make more errors when under time pressure (Pettibone, 2012). These errors include choosing the distractor which is strictly dominated by one of the other choices. Comparison noise is one explanation for this error: If people select the distractor then they cannot have made correct comparisons. Second, the qualitative effects of context on preference reversal are not changed by the value of the comparison noise. All of the context effects reported in this chapter are also predicted by a model without comparison noise, as shown in Figure 4.4b. Third, the level of comparison noise in my fitted model is so low that it results in a decoy selection rate of about 2%. A decoy selection is the only type of error in the task. This rate exactly corresponds to the human rate.

The approach that I have taken in this chapter is an example of a broader class of analysis known as Computational Rationality (Howes et al., 2009; Lewis et al., 2014; Lieder and Griffiths, 2019). This approach starts from the assumption that people are approximately rational given the bounds imposed by the computation required for cognition. It then seeks to discover the computational limits that give rise to boundedly optimal (Russell and Subramanian, 1994) but apparently irrational behaviours. This aim demands that the analyst derive bounded optimal policies for well-formed decision problems. My results suggest an answer to the paradox of why it is worth motivating machine learning algorithms with apparently biased human decision making. While the behaviour appears biased, the underlying processes and heuristics (e.g., the use of option comparison) lead to gains in efficiency and therefore reward. Important directions for future research suggest that human irrationalities may offer a productive source of inspiration for improving the design of AI architectures and machine learning methods. As others have shown (Simsek et al., 2016) comparison observations are a particularly important avenue for exploration.

What is more, my results contribute to a growing body of work calling into question the long list of apparent irrationalities reported in the Economic literature. More may be amenable to POMDP, Meta-MDP, or MDP explanations and turn out to be rational adaptations to environmental and cognitive limits.

In conclusion, framing contextual choice problems as POMDPs reveals that apparently irrational choice reversals in behaviour are demonstrably rational under bounds imposed by uncertainty in the observation function.

Chapter 5

Deep Reinforcement Learning Models of the Fourfold Pattern (Separate Models)

5.1 Introduction

Kahneman and Tversky (1979) examined Expected Utility Theory as a model of human decision making under risk, and concluded that it was violated by a wide range of empirical biases in human behaviour. The model that they proposed in response, called Prospect Theory, has dominated scientific and popular views of human decision making for the intervening 40+ years. Recently, Kahneman (2016) has listed 36 biases that challenge the view of the human as a (bounded) rational actor – suggesting that the idea that people are biased is alive and well.

However, recent work in cognitive science has shown that some apparent biases, for example preference reversals, can be explained as emergent consequences of computational rationality (Lewis et al., 2014) or ecological rationality (Gigerenzer, 2018). In this thesis, I revisit the risky choice tasks that provided evidence for Prospect Theory and show, through computational modelling, that they can be explained as a consequence of a boundedly optimal process.

As I have said in Chapter 2, there is a disconnect in the risky choice literature between the models of risky choice based on expected utility theory (rational models) and models of cognitive processing (process models). These are two threads of risky choice modelling, each taking a different direction. As pointed out in Pachur et al. (2018) : “the disconnect between expectation models (e.g., CPT) and the process-tracing tradition is unfortunate: It may occlude how decision-making could be improved by influencing information processing”. So there arises a research question: is there another way to make the models of risky choice more realistic?

In my work I attempt to answer this question and bridge the gap. Accordingly, this Chapter reports models of risky choice developed within a single unifying framework, linking Marr’s level 1 and level 2 analysis (Love, 2015; Marr, 1982; Marr and Poggio, 1977). The model explains *why* people make the choices that they do in terms of the underlying information processes. Moreover, the underlying processes are rational adaptations to utility maximization. In the following, I will draw the two threads (rational and process) together by proposing a normative model of the computationally rational process in risky choice.

The contributions of this Chapter are as follows:

- The sequential models provide quantitative predictions of the information processes underlying risky choice.
- The models shed light on the way apparent human cognitive biases can emerge from computationally rational processing.
- The models show a novel way in which machine learning methods can be used in cognitive science.

5.2 Hypothetical Risky Choice Tasks

Kahneman and Tversky (1979) reported human data for a number of different types of hypothetical choice problems. These problems offer a choice between two options (A and B). First, the participants were asked to imagine that they were faced with the choice described in the problems which were presented in the questionnaire form. Then they indicated the option they would have chosen among two risky prospects in such a case. For example, in Problem

1 (Kahneman and Tversky, 1979) people prefer option B (2400 for certain) over option A (.33 chance of 2500, .66 chance of 2400, 0 with probability .01). Whereas in Problem 2, people prefer option C (0.33 chance of winning 2500), over option D (0.34 chance of winning 2400). This pattern of preferences violates the expected utility theory because each problem suggests a different rank ordering for the utility of 2400 and 2500. Problem 1 is a "certainty" problem, and Kahneman and Tversky (1979) designed it to provide evidence in support of the hypothesis that people overweight certain options. Problems 3 and 4 in (Kahneman and Tversky, 1979) also tested for the certainty effect, as do Problems 5 and 6.

Problems 7 and 8 illustrate that people will be risk seeking for low probability options and risk averse for high probability gains ("possible" gains). A key feature of the fourfold pattern of risky choice.

The next set of problems (Problems 3', 4', 7', and 8') demonstrate the *Reflection effect*. Here, the pattern of effects seen in 3, 4, 7, and 8 are reversed when the signs of the outcomes are reversed, and gains become losses – also required for the fourfold pattern.

The *Isolation effect* is illustrated by Problems 10, 11, and 12. In these problems, people are believed to disregard parts of the options that are shared and focus on the components that distinguish them, see page 271 (Kahneman and Tversky, 1979). For example, in stage 1 of problem 10 there is a probability of .75 to end the game without winning and .25 to proceed to the second stage. In the second stage there is a choice between (4000,.8) and (3000) – which is the same as Problem 4 above. Overall, the available choices – when integrated over stage 1 and stage 2 – are (4000,.2) and (3000, .25) which is the same as Problem 3 (above). Kahneman and Tversky's data indicate that on average participants treat Problem 10 similarly to Problem 3, rather than to Problem 4. They respond by assuming that participants ignore the shared (stage 1) information and they build this assumption into Prospect Theory. Problem 10 illustrates that the risky choices are altered by different representations of probabilities, which are the standard formulation and the sequential formulation. Furthermore, Problems 11 and 12 show that the preferences are altered by different representations of values that have been given a bonus. In Problems 11 and 12, they also assume that subjects ignore the shared information (the initial bonus). Prospect Theory (Kahneman and Tversky, 1979) explained that the additional information did not enter into the subjects' evaluation process. In other words, people ignored the sequential stage and the bonus.

Problems 13 and 13' demonstrate the *Value Function*, which is generally concave for gains and commonly convex for losses. The value function shows as S-shape and is steeper for losses than for gains.

Problems 14 and 14' demonstrate the *Weighting Function*. In these problems, people overweight the low probability that they would be willing to pay for both insurance and gambling.

5.3 Risky Choice as a POMDP

Here, I present a normative decision-making model based on a POMDP. It is an extension of the unifying framework described in Chapter 4, where a POMDP was used to build a computationally rational model of contextual choice. In this Chapter, I extend this approach to capture the fourfold pattern of risky choice and compare the predictions of the model to existing human data.

My approach is based on the intuition that the process of risky choice can itself be viewed as a sequential decision making problem. In other words, I describe risky choice as a process of noisy evidence accumulation from a stimulus. At each step of this task, the participant or agent chooses which information to gather or whether to terminate the task by making a decision. The goal is to obtain the maximum rewards given noisy evidence and the cost of information gathering actions. Gathering more evidence might increase the probability of a better, more rewarding, choice. Meanwhile, gathering more evidence also costs more time. Therefore, there is a trade-off between the information cost and the probability of making a more rewarding choice. Given this formulation, the problem of making a decision under risk is a sequential decision problem that can be modelled with a POMDP. The strategy is a policy for the POMDP, that is a function mapping the results of previous actions to the next action. Therefore, I use reinforcement learning methods to discover the approximately optimal policy and demonstrate that the fourfold pattern of risky choice is an emergent consequence of an optimal solution to a POMDP.

I formulated 17 choice problems as 4 POMDPs. The first POMDP is the basic model for all the others, named **basic model**. It is used for 11 problems, which are Problems 2, 3, 4, 7, 8, 14, 3', 4', 7', 8', and 14'. In each task, it presents a choice between two prospects or gambles ($X(v_X, p_X)$ and $Y(v_Y, p_Y)$). Each option is a contract that yields outcome v with probability p and outcome 0 with probability $1-p$. The second POMDP is a **trinary-outcome model**, corresponding to 3 problems, which are Problems 1, 13, and 13'. In these problems, one option has three outcomes, and another has two outcomes. They are represented as $X(v_{X1}, p_{X1}; v_{X2}, p_{X2})$ and $Y(v_Y, p_Y)$. For choosing option X, people have a p_{X1} chance to win v_{X1}, a p_{X2} chance to win v_{X2} and a $1-p_{X1}-p_{X2}$ chance to win 0, where $p_{X1}+p_{X2} \leq 1$. The third POMDP is a **two-stage model** and is used only for Problem 10, which is a two-stage game. In the first stage, there is a p_s chance to move into the second stage and a $1-p_s$ to end the task with nothing to win. In the next stage, people have a choice the same as the one in **basic model**. To be noted, people must choose an option before the game starts; that is when the outcome of the first stage is not released. The last POMDP is a **bonus model** and is used to formulate Problems 11 and 12. In these problems, the initial bonus was given to people before they made a choice as in **basic model**.

Basic model: Each risky choice task had two options $(X(v_X, p_X), Y(v_Y, p_Y))$ which were represented with two attributes: a randomly sampled probability p and a randomly sampled value v. I assumed that the probabilities p were sampled from a $\beta-$distribution ($a=1, b=1$) and the values v were sampled from a $t-$distribution ($location = 3000, scale = 1000$, $degrees\ of\ freedom = 100$). I note that the distributions of the two attribute values are the same as in Chapter 4, but with different parameters.

The state space $\mathcal{S}$ for each task was generated from a sampled risky choice task. A state represents visual attention, including fixations, saccades, and join fixations[1], on all attributes (v_X, p_X, v_Y, p_Y). There are saccades and join fixations within one option, that are (v_X, p_X) and (v_Y, p_Y). They correspond to 2 calculation actions (e.g., compute the expected value for

[1] Join fixation means that people look at two attributes at the same time. For example, commonly people look at v_X or v_Y separately. Join fixation means that people look at v_X and v_Y together (Stewart et al., 2016).

X given v_X and p_X). There are also attention within attribute dimension across two options, that are (v_X, v_Y) and (p_X, p_Y). They correspond to 2 comparison actions (e.g., compute the comparison relation for v_X and v_Y). In the end, the agent chooses an option with one of 2 choice actions (choose X, choose Y). The reward for comparison and calculation actions was negative c. The reward for a choice action was the outcome of the choice.

In summary, the state space $\mathcal{S}$ includes 4 elements. The action space $\mathcal{A}$ includes 2 comparison actions, 2 calculation actions and 2 choice actions. Further, for the **basic model**, the set of possible observations in the history $\mathcal{O}_{BAh}$ is the set of noisy encodings of partial orderings and calculations:

$$\mathcal{O}_{BAh} = \{f(p_X, p_Y), f(v_X, v_Y), E_X, E_Y\} \tag{5.1}$$

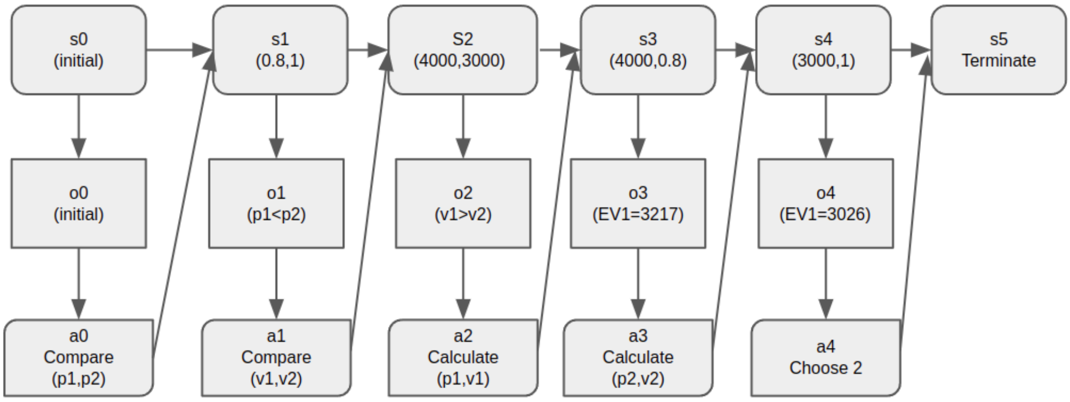


Fig. 5.1 Schematic illustration of the POMDP framework applied to the risky choice task. It takes Problem 3: $A : (v1 = 4000, p1 = 0.8)$ or $B : (v2 = 3000, p2 = 1)$ as an example.

As shown in figure 5.1, at each time step t there is a latent state $s_t \in \mathcal{S}$. The observation of this state $\mathcal{O}_{BAht}$ gives rise to "evidence", i.e the history of the partial and noisy observation of the latent state s_t. The latent state s_t captures the *external environment* in which decision-making takes place. And the observation $\mathcal{O}_{BAht}$ captures the *internal environment* of the cognitive process (calculation and comparison) that underlie the decisions. When the agent stops gathering information, it decides to choose an option by taking a choice action, at which point the action is taken in the external world and it receives a reward.

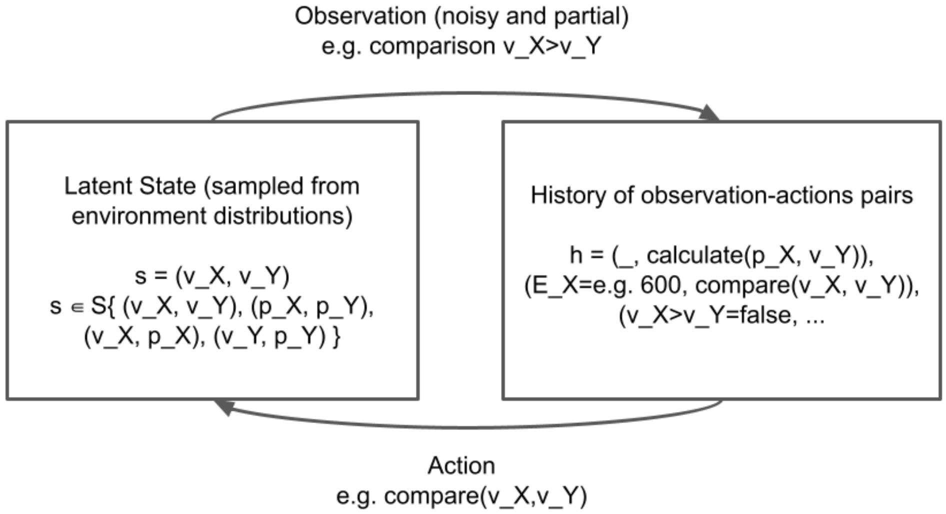

Observation (noisy and partial)
e.g. comparison v_X>v_Y
Latent State (sampled from environment distributions)
s = (v_X, v_Y)
s ∊ S{ (v_X, v_Y), (p_X, p_Y), (v_X, p_X), (v_Y, p_Y) }
History of observation-actions pairs
h = (_, calculate(p_X, v_Y)), (E_X=e.g. 600, compare(v_X, v_Y)), (v_X>v_Y=false, ...
Action
e.g. compare(v_X,v_Y)

Fig. 5.2 An example of the state and observation for the **basic model** at time step $t = 2$.

As shown in figure 5.2, at this time step $t = 2$, the agent takes the action (comparing v_X and v_Y) given the history of observations that only E_X is known. Then the agent computes the comparison relation for v_X and v_Y based on the information from state $s_2 = (v_X, v_Y)$. The result of this computation is the observation and is used to update the agent's history of observations.

Trinary-outcome model: In the trinary task, each choice again has two options $X($ $v_{X1}, p_{X1}; v_{X2}, p_{X2})$ and $Y(v_Y, p_Y)$. The option $X(v_{X1}, p_{X1}; v_{X2}, p_{X2})$ differs from the one $X(v_X, p_X)$ in the **basic model**. As a consequence, it has a bigger state space $\mathcal{S}$, observation space $\mathcal{O}$ and action space $\mathcal{A}$. The state space $\mathcal{S}$ includes $(v_{X1}, p_{X1}; v_{X2}, p_{X2})$, (v_Y, p_Y), (v_{X1}, v_Y), (v_{X2}, v_Y), (p_{X1}, p_Y) and (p_{X2}, p_Y). They correspond to 6 computation actions. The other components are the same as in the **basic model**.

In summary, the state space $\mathcal{S}$ includes 6 elements. The action space $\mathcal{A}$ includes 4 comparison actions, 2 calculation actions and 2 choice actions. So for the **trinary-outcome model**, the set of possible observations in the history $\mathcal{O}_{TOh}$ is:

$$\mathcal{O}_{TOh} = \{f(v_{X1}, v_Y), f(v_{X2}, v_Y), f(p_{X1}, p_Y), f(p_{X2}, p_Y), E_X, E_Y\} \tag{5.2}$$

Two-stage model: The components are the same as in the **basic model**, except for the state space $\mathcal{S}$ and observation space $\mathcal{O}$. Before the two-stage game starts, the probability p_s is known to the agent, which must choose an option. Thus, the probability p_s is represented in the state and the observation from the beginning to the end. The set of possible observations in the history $\mathcal{O}_{TSh}$ is:

$$\mathcal{O}_{TSh} = \{p_s, f(p_X, p_Y), f(v_X, v_Y), E_X, E_Y\} \tag{5.3}$$

Bonus model: Similar to the **two-stage model**, an initial bonus is represented in the state and the observation in all the processes of choice. The set of possible observations in the history $\mathcal{O}_{BOh}$ is:

$$\mathcal{O}_{BOh} = \{bonus, f(p_X, p_Y), f(v_X, v_Y), E_X, E_Y\} \tag{5.4}$$

The environment of the model was defined for all of the tasks used in Kahneman and Tversky (1979). All gambles were modeled within OpenAI gym (Brockman et al., 2016), which is an open source interface for reinforcement learning.

The RL models learn to gather two sources of information, expected values and order of feature values, and make a choice based on them. First, the models are trained on randomly sampled paired choices. Then, the trained RL models are tested on the critical tasks defined by Kahneman and Tversky (1979). The results are in the following sections.

5.4 Results

5.4.1 Learn an Approximately Optimal Policy

During the training process, for all four models, the probabilities p were sampled from a $\beta-$distribution $(a = 1, b = 1)$ and the values v were sampled from a $t-$distribution $(location = 3000, scale = 1000, degrees\ of\ freedom = 100)$. These distributions represented the ecological distributions experienced by participants in the human behaviour experiments reported by Kahneman and Tversky (1979). For all the experiments below, we used the same distributions.

The same as the settings in Chapter 4, I built the environments within OpenAI Gym (Brockman et al., 2016) and used the OpenAI Baselines [2] implementation of the deep RL algorithms: ACER. Hyperparameters were as follows:

- Policy Network: (64, tanh, 64, tanh, Linear) + Standard Deviation variable; Value Network (64, elu, 64, elu, linear)
- Number of timesteps = 10 million
- Batch size = 1024
- Learning rate for RMSProp = 7e-4
- Schedule of learning rate: linear annealed
- Actor-learner threads = 32
- Other hyperparameters are default as in OpenAI Baselines implementation

The fitted parameter values were: calculation noise $\sigma_{calc} = 0.3$, comparison error $P(error_f) = 0.1$, probability weighting parameters $\alpha = 1$, the perceived cost of comparison $C_{comparison} = -10$ and the calculation cost $C_{calc} = -10$. I do not claim to have achieved the best possible fit, nor a better fit than other models. The point of the fit was to show that the qualitative effects exhibited by humans were within the space of behaviours generated by

[2] https://github.com/openai/baselines

the agent. As a result, while this provides proof that the data can be modelled, it does not indicate how likely that result is, i.e., it could be in an obscure restricted region of parameter space. So I would now fit the new data with these parameters kept the same.

Kahneman and Tversky (1979) use a positive prospect to denote a gain and a negative prospect to denote a loss. For the training samples with the negative prospect in Problems 3', 4', 7', 8', 12, and 13', the *location* is correspondingly set to a negative value. For those problems, the values v were sampled from a $t-$distribution ($location = -3000, scale = 1000, degrees\,of\,freedom = 100$).

During the training process of the **two-stage model**, the probabilities p_s were also sampled from a $\beta-$distribution ($a = 1, b = 1$), the same as the distribution of the probability p. The initial bonus for the **bonus model** was drawn from a $t-$distribution ($location = 3000, scale = 1000, degrees\,of\,freedom = 1$), which is similar to the distribution of value v but with different $freedom$.

Therefore, seven models are trained to learn the approximately optimal policy for 17 problems. Reported results are averaged over 10 runs, each with a different random seed, after training on 10 million samples for each problem. In the following, I will report the performance of the trained models.

5.4.2 Certainty, Probability, Possibility, and The Reflection Effect

Kahneman and Tversky (1979) define the certainty effects as a preference for a certain gamble over a risky gamble with a higher or equal expected value. They claim that evidence indicates that people "overweight" outcomes that are considered certain. In this Section, I show that an optimal bounded model generates the certainty effect (and other effects) in the pursuit of maximizing utility – it does not overweight outcomes. I do so by demonstrating that the model predicts the human data for the problems in Kahneman and Tversky (1979).

The predictions in Tables 5.1 and 5.2 show that the utility maximizing models predict the reported human data (Kahneman and Tversky, 1979) in all problems. In Problems 1, 3, and 7, the data indicates risk aversion, whereas risk-seeking in Problems 2, 4, and 8. In each of the four problems (Problems 3', 4', 7', and 8') in Table 5.2, the results demonstrate the

Table 5.1 Risky preferences of the model and human. The > symbol denotes the most "prevalent" preference, which is the choice made by the majority of participants in the experiment. The percentage of subjects who chose each option is shown in brackets, e.g., 83 percent of the subjects chose the (2500, .33) option and 17 percent of the subjects chose the (2400, .34) option in Problem 2. The human data is from Kahneman and Tversky (1979).

Problem	Model	Human
1: (2500,.33;2400,.66) < (2400,1.0)	[0, 100]	[18, 82]
2: (2500,.33) > (2400,.34)	[96, 4]	[83, 17]
3: (4000,.80) < (3000,1.0)	[21, 79]	[20, 80]
4: (4000,.20) > (3000,.25)	[98, 2]	[65, 35]
7: (3000,.90) > (6000,.45)	[71, 29]	[86, 14]
8: (3000,.002) < (6000,.001)	[2, 98]	[27, 73]

Table 5.2 Predictions of the preference between negative prospects. The human data is from Kahneman and Tversky (1979).

Problem	Model	Human
3': (-4000,.80) > (-3000,1.0)	[95, 5]	[92, 8]
4': (-4000,.20) < (-3000,.25)	[23, 77]	[42, 58]
7': (-3000,.90) < (-6000,.45)	[3, 97]	[8, 92]
8': (-3000,.002) > (-6000,.001)	[70, 30]	[70, 30]

reflection effect, that is, the preference between negative outcomes is the mirror image of the preference between positive outcomes.

5.4.3 The Isolation Effect

Prospect Theory (Kahneman and Tversky, 1979) takes the isolation effect as evidence that the subjects ignore sequential information and bonuses (see above) and this assumption is built into the theory. However, the results of the current analysis show that the bounded optimal RL agent generates the isolation effect by maximizing rewards, as shown in Table 5.3.

Therefore, there is no need to make a specific theoretical commitment to ignoring particular types of information in order to model human choice; rather, information is used, or not used, to the extent that it helps maximize utility. As Kahneman and Tversky (1979)

Table 5.3 Predictions of the isolation effect. The human data is from Kahneman and Tversky (1979).

Problem	Model	Human
10: (4000,.80) > (3000,1.0), p = 0.75	[19, 81]	[22, 78]
11: (1000,.50) < (500,1.0), bonus = 1000	[28, 72]	[16, 84]
12: (-1000,.50) > (-500,1.0), bonus = 2000	[68, 32]	[69, 31]

assumed, people do not incorrectly ignore sequential information; rather, they do so correctly because it helps them perform the task better.

Unlike in Prospect Theory, the isolation effect is not a consequence of the theorist assuming that certain information will be ignored, rather the model learns to ignore irrelevant information. The model offers an explanation of the isolation effect by taking account of it from a normative point of view instead of a descriptive one, as Prospect Theory does.

5.4.4 The Value Function

The model predicts the qualitative direction of the value function effects, as shown in Table 5.4. These predictions of the preferences are in accord with the hypothesis made by Prospect Theory that the value function is concave for gains and convex for losses (Kahneman and Tversky, 1979).

Table 5.4 Predictions of the value function. The human data is from Kahneman and Tversky (1979).

Problem	Model	Human
13: (6000,.25) < (4000,.25;2000,.25)	[43, 57]	[18, 82]
13’: (-6000,.25) > (-4000,.25;-2000,.25)	[64, 36]	[70, 30]

I highlight an important feature of my models. The computationally rational model derives the value function from preferences even with an unbiased utility function. However, in the risky choice literature, much emphasis is put on different models of bias in the utility function. In contrast, my model uses expected utility instead of subjective utility – there is no bias in the utility function. The predictions made by the model explain human risky choice behaviour by virtue of bounded optimality, not by virtue of biased utility. The results explain the phenomena of risky choice in terms of policies that are optimized to the bounds. I use reinforcement learning methods to explore the implications of the cognitive and ecological bounds that limit risky choice behaviour. These bounds concern limited cognitive capacity and noisy, partial observations of the environment.

5.4.5 The Weighting Function

According to Prospect Theory, people commonly prefer what is, in effect, a lottery ticket (high value but very small probability) over the expected value of that ticket. However, the results in Table 5.5 show that the RL agents predict the different direction of this effect compared to human data. The models exhibit risk aversion for gains and risk-seeking for losses as they also do in “certainty” and “reflection” problems (Problems 3, 7, 3’, and 7’).

Table 5.5 Predictions of the weighting function with an unbiased utility function ($\alpha = 1$). The human data is from Kahneman and Tversky (1979).

Problem	Model	Human
14: (5000,.001) > (5,1.0)	[38, 62]	[72, 28]
14’: (-5000,.001) < (-5,1.0)	[64, 36]	[17, 83]

To model the properties of the weighting function, I introduce the assumption that the probability is weighted by an exponential parameter, therefore the observation of a calculation was computed using: $E_i = p_i^{\alpha} \times v_i + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2_{calc})$. The purpose of using parameter α was to model **subjective probability** following Savage (1972). The model assumes that the weighting function $w(p) = p^{\alpha}$, whereas Prospect Theory assumes the weighting function shown in:

$$w(p) = \frac{p^{\gamma}}{p^{\gamma} + (1 - p^{\gamma})^{1/\gamma}} \tag{5.5}$$

Obviously, the actual scaling of this assumption is considerably more complicated than in my model, which uses only an exponential parameter.

Table 5.6 Predictions of the weighting function when the parameter $\alpha = 0.6$. The human data is from Kahneman and Tversky (1979).

Problem	Model	Human
14: (5000,.001) > (5,1.0)	[97, 3]	[72, 28]
14’: (-5000,.001) < (-5,1.0)	[4, 96]	[17, 83]

The observation of a calculation overweights the low probability when the parameter $\alpha \in (0,1)$ is set between 0 and 1. Thus, it enhances the attractiveness of the option with a low probability for gains and the option with a high probability for losses. The results in Table 5.6 shows the predictions when the parameter $\alpha = 0.6$.

However, as I note in section 5.4.4, the model derives the fourfold pattern from preferences merely with an unbiased utility function. Here I maintain the parameter $\alpha = 1$ and fit the ecological distributions. In Problems 14 and 14’, the values (5000 and 5) have a larger variance than those in other problems. Therefore, during the training process, there is a larger $scale$ of t-distribution ($location = -3000\ or\ 3000, scale = 3000, degrees\ of\ freedom = 100$), from where the the values v are sampled.

Table 5.7 Predictions of the weighting function. The human data is from Kahneman and Tversky (1979).

Problem	Model	Human
14: (5000,.001) > (5,1.0)	[68, 32]	[72, 28]
14’: (-5000,.001) < (-5,1.0)	[22, 88]	[17, 83]

The RL model predicts the qualitative direction of this effect. The results are shown in Table 5.7. They show the properties of the weighting function for small values of the probability with an unbiased utility function. It is consistent with the assumption proposed by Prospect Theory that very low probabilities are generally over-weighted (Kahneman and Tversky, 1979). The results demonstrate that the computationally rational agent could capture the pattern of risky behaviour: risk-seeking for gains and risk aversion for losses of low probability.

Previous descriptive models have been fitted to this effect but did not explain why and how subjects perform in this way. In the following, I will offer an explanation for *why* the effect occurs in terms of the rational, decision theoretic, basis of the POMDP problem formulation and optimal policy.

5.4.6 Optimal Sequential Decision Making Under Uncertainty

While the above analysis suggests that the computationally rational model can capture four-fold pattern behaviour, the results provide only a preliminary suggestion of the implications of the model. These are explored further in this Section.

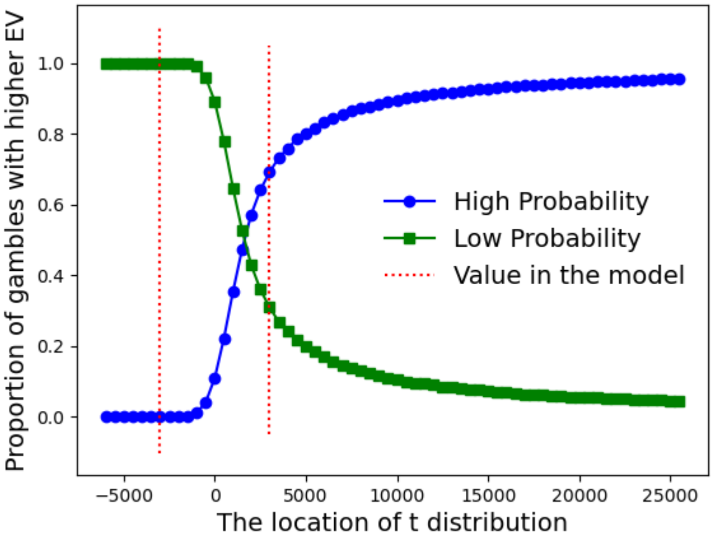


Fig. 5.3 The proportion of gambles with a high expected value against the *location* of the t-distribution of the value. These figures present the effect of the ecological distribution of probability and value on the proportion of options with higher expected value in two option risky choice problems. For each value of the *location*, the data is generated by 10^5 choice tasks which are sampled from the following distribution. The probabilities p are sampled from a $\beta-$distribution $(a = 1, b = 1)$ and the values v are sampled from a t-distribution $(location \in [-6000, 26000], scale = 1000, degrees\ of\ freedom = 100)$, where the $location$ ranges from -6000 to 26000. The red dot line indicates the value used in the training process of the model.

To understand why the model could predict the pattern of risky attitudes, I ran the numeric simulation shown in Figure 5.3 and Figure 5.4. These figures present the effect of the ecological distribution of probability and value on the proportion of options with higher expected value in two option risky choice problems. For each choice task, the probability of one option is greater than another option.

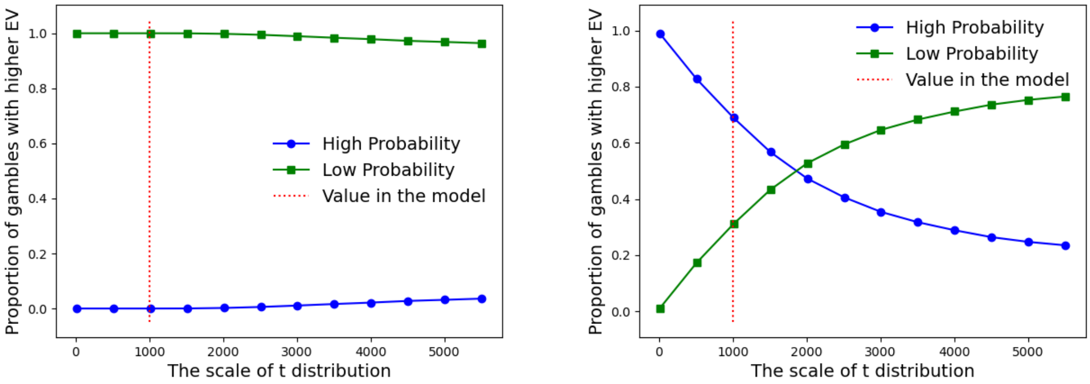


Fig. 5.4 The proportion of gambles with a high expected value against the *scale* of the t-distribution of the value. For each value of the *scale*, the data is generated by 10^5 choice tasks which are sampled from the following distribution. In both panels, the probabilities p are sampled from a $\beta-$distribution ($a = 1, b = 1$). In the left panel, the values v are sampled from a t-distribution ($location = -3000, scale \in [0, 6000], degrees\ of\ freedom = 100$), where the *scale* ranges from 0 to 6000. In the right panel, the values v are sampled from a t-distribution ($location = 3000, scale \in [0, 6000], degrees\ of\ freedom = 100$), where the *scale* ranges from 0 to 6000. The red dot line indicates the value used in the training process of the model.

The results in Figure 5.3 show that when $location > 1660$, options with high probability have, on average, a higher expected value than options with low probability. In contrast, when $location < 1660$, options with high probability have, on average, a lower expected value than options with low probability. In Figure 5.3, the *scale* is set to 1000. In the left panel of Figure 5.4, the $location = -3000$ is set to a negative value, which denotes the loss. The results show that, on average, options with high probability have a lower expected value. In the right panel of Figure 5.4, the $location = 3000$ is set to a positive value, which denotes the gain. The results show that when $scale < 1860$ options with high probability, on average, have a higher expected value.

In the training process of the above models, the *location* is set to -3000 or 3000, and the *scale* is set to 1000, indicated as red dot line in Figure 5.3 and Figure 5.4. Therefore, in the domain of losses, the risk seeking strategy that prefer the lowest probability option is approximately optimal. Since 98 percent of options with low probability have a higher expected value. In contrast, in the domain of gains, the risk aversion strategy that always selects the highest probability option is approximately optimal. Because 68% of options with high probability have a higher expected value. These findings explain why the computationally rational agent could capture the pattern of risk attitudes: risk aversion for gains and risk seeking for losses of high probability. The use of a risk aversion or risk-seeking strategy will not always lead to choosing the option with the highest expected value. However, over the long run, the strategy for the domain will more closely approximate the optimum.

According to the above analysis, the fourfold pattern phenomena in risky choice are computationally rational under bounds imposed by environmental and cognitive limits. Furthermore, I offer an explanation of the fourfold pattern in terms of the variation in the environmental distribution of probability and outcome. The results show that the phenomena could emerge from the expected value maximizing process given plausible features of the adaptation environment and cognitive mechanisms.

5.5 Discussion

The results presented in this Chapter demonstrate that, given reasonable assumptions about noisy information processing (bounds), an optimal agent will exhibit an identical pattern of choices to the majority of humans on tasks that were previously thought to reveal violations of rational choice theory. The results are therefore evidence against the claim, made by Kahneman and Tversky (1979), that people violate rational choice theory. Given the bounds imposed by the human ability to calculate expected value, people are likely doing the best that they can to make optimal choices and do not violate rational choice theory given the limitations of the information made available through noisy observation.

Moreover, there is evidence that across the reported models, each aspect of the fourfold pattern of choice is exhibited. The models exhibit risk seeking over low-probability gains, risk aversion over high-probability gains, risk aversion over low-probability losses, risk seeking over high-probability losses.

As I discussed above in the Chapter 2, there arises a strong desire to build a single unifying framework for modelling human decision making. Here, I would argue that computation rationality is a promising theoretical framework for integrating various approaches and explaining human behaviour. To illustrate the potential of this framework, I apply it to a set of risky choice tasks and drive the computationally rational processes underlying those phenomena that violate the normative principles of human rationality.

I offer an explanation of the fourfold pattern in terms of the variation in the ecological distribution of probability and outcome. For specific environmental distributions, the phenomena could emerge from the value maximizing processes. In the future, it would be of interest to train the RL model with choice tasks sampled from a broader range of distributions. Then, in the specific range of the distribution space, the RL agent could predict the quantitative preferences corresponding to human data. However, when the distributions are in the other specific space, the predicted preference reverses.

Chapter 6

A Deep Reinforcement Learning Model of Fourfold Pattern (a Unified Model)

In the previous Chapter, I explored the possibility that a wide range of risky choice phenomena emerge from a boundedly optimal adaptation of a decision-making agent with processing constraints. To accomplish this, 4 *separate* models were proposed to solve a range of risky choice problems following Kahneman and Tversky (1979). The 7 models were trained separately with different parameter settings. In this Chapter, I propose a *unified* model that solves all 17 risky choice problems. The single model demonstrates that all effects could emerge from a unified set of assumptions about the bounds on human cognition that lead to risky choice effects. In doing so, the model offers a unified explanation for the fourfold pattern of risk choice. Furthermore, all effects emerge from a single set of model parameters.

Moreover, another focus of Chapter 5 is on explanations of the fourfold pattern in terms of the variation in the ecological distribution and the cognitive limits. While the theory proposed in Chapter 5 explains *why* people make the choices that they do as rational adaptations to utility maximization, the model does not explain *how* subjects behave in terms of the underlying information processes. In addition, the nature of the underlying processes that take place in risky decision making has been a question of longstanding interest. Here, I take a normative account of the information processing mechanism that underlies risky choices. In the following, I will use a single *unified* model to predict the approximately optimal decision-making process. To the best of my knowledge, it is the first model to make such predictions from a normative rather than a descriptive perspective.

6.1 A Unified Model

As I discussed in the previous Chapter, 4 POMDPs were proposed to formulate 17 risky choice problems. And all the 7 models were trained separately to solve all the problems. Those models share the same framework. Thus, in this Chapter, I will combine the 7 models into a unified model.

Choice type c_{type} is recorded in the state. There are 7 choice types: basic gains (Problems 2, 3, 4, 7, 8, and 14), basic losses (Problems 3', 4', 7', and 8'), trinary gains (Problems 1 and 13), trinary losses (Problems 1' and 13'), two-stage gains (Problem 10), bonus with gains (Problem 11) and bonus with losses (Problem 12). The state space $\mathcal{S}$ includes c_{type}, p_s, $bonus$, $(v_{X1}, p_{X1}; v_{X2}, p_{X2})$ or (v_X, p_X), (v_Y, p_Y), (v_{X1}, v_Y) or (v_X, v_Y), (v_{X2}, v_Y), (p_{X2}, p_Y), and (p_{X1}, p_Y) or (p_X, p_Y). They correspond to 6 computation actions. In summary, the state space $\mathcal{S}$ includes 9 elements. The action space $\mathcal{A}$ includes 4 comparison actions, 2 calculation actions, and 2 choice actions. So for the unified model, the set of possible observations in the history $\mathcal{O}_{Uh}$ is:

$$\mathcal{O}_{Uh} = \{c_{type}, p_s, bonus, f(v_{X1}\ or\ v_X, v_Y), f(v_{X2}, v_Y), f(p_{X1}\ or\ p_X, p_Y), f(p_{X2}, p_Y), E_X, E_Y\} \tag{6.1}$$

6.2 Results

The results of Model Seven are the same as in the previous Chapter. Model One is the result of training the unified model on each type of problem with different parameters. Thus, there are seven sets of parameter settings for Model One. Whereas I train the unified model on each task using only one set of parameter settings. That is Model Zero. During the training process, the probabilities p were sampled from a $\beta-$distribution $(a = 1, b = 1)$ and the values v were sampled from a $t-$distribution ($location = 3000, scale = 1000, degrees\ of\ freedom = 100$). In each interaction with the environment, the agent randomly takes one type of choice task out of the seven types of choices. Reported results are averaged over 10 runs, each with a different random seed, after training on 100 million samples for all the problems. The other settings and hyperparameters were the same as in Chapter 5.

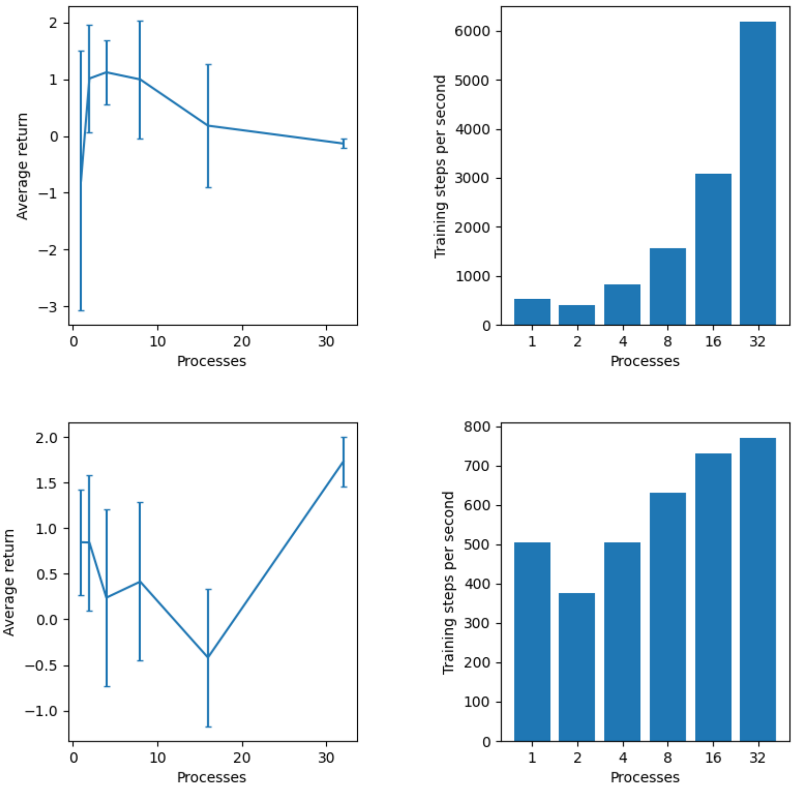


Fig. 6.1 The training performance of the Model Zero with each environment on separate process. The top is the performance of A2C and the bottom is the performance of PPO. The results are averaged over 10 runs with different seeds, and the error bars indicate the (95%) confidence interval.

To test the characteristics of the 17 choice tasks, I trained the Model Zero using the A2C and PPO algorithms with various copies of the environment per process, as shown in Figure 6.1. The results demonstrate that 4 copies of the environment per process are the optimal setting. The bottleneck of training speed is parallel training.

6.2.1 Risky Choice Predictions

The predictions show that a unified model predicts the reported human data (Kahneman and Tversky, 1979) in all problems, as shown in Tables 6.1, 6.2, 6.3, 6.4, 6.5.

Table 6.1 Risky preferences of the model and human. The > denotes the prevalent preference which is the choice made by the majority of participants in the experiment. The percentage of the subjects chose each option is shown in brackets, e.g., 83 percent of the subjects chose the (2500, .33) option and 17 percent of the subjects chose the (2400, .34) option in Problem 2. The human data is from Kahneman and Tversky (1979).

Problem	Model Seven	Model One	Model Zero	Human
1: (2500,.33;2400,.66) <(2400,1.0)	0, 100	0, 100	22, 78	18, 82
2: (2500,.33) >(2400,.34)	96, 4	96, 4	66, 34	83, 17
3: (4000,.80) <(3000,1.0)	21, 79	18, 82	27, 73	20, 80
4: (4000,.20) >(3000,.25)	98, 2	97, 3	75, 25	65, 35
7: (3000,.90) >(6000,.45)	71, 29	75, 25	65, 35	86, 14
8: (3000,.002) <(6000,.001)	2, 98	2, 98	18, 82	27, 73

Table 6.2 Predictions of the preference between negative prospects. The human data is from Kahneman and Tversky (1979).

Problem	Model Seven	Model One	Model Zero	Human
3’: (-4000,.80) >(-3000,1.0)	95, 5	77, 23	70, 30	92, 8
4’: (-4000,.20) <(-3000,.25)	23, 77	1, 99	23, 77	42, 58
7’: (-3000,.90) <(-6000,.45)	3, 97	31, 69	40, 60	8, 92
8’: (-3000,.002) >(-6000,.001)	70, 30	1, 99	84, 16	70, 30

Table 6.3 Predictions of the isolation effect. The human data is from Kahneman and Tversky (1979).

Problem	Model Seven	Model One	Model Zero	Human
10: (4000,.80) >(3000,1.0), p = 0.75	19, 81	16, 84	29, 71	22, 78
11: (1000,.50) <(500,1.0), bonus = 1000	28, 72	34, 66	36, 64	16, 84
12: (-1000,.50) >(-500,1.0), bonus = 2000	68, 32	73, 27	60, 40	69, 31

Table 6.4 Predictions of the value function. The human data is from Kahneman and Tversky (1979).

Problem	Model Seven	Model One	Model Zero	Human
13: (6000,.25) <(4000,.25;2000,.25)	43, 57	43, 57	64, 36	18, 82
13’: (-6000,.25) >(-4000,.25;-2000,.25)	64, 36	64, 36	41, 59	70, 30

Table 6.5 Predictions of the weighting function. The human data is from Kahneman and Tversky (1979).

Problem	Model Seven	Model One	Model Zero	Human
14: (5000,.001) >(5,1.0)	97, 3	78, 22	98, 2	72, 28
14’: (-5000,.001) <(-5,1.0)	4, 96	6, 94	2, 98	17, 83

6.2.2 Sequential Process Predictions

To explore the order in which computational information is gathered, the Model Zero agent is used to predict the information gathering process. In order to make it simple and straightforward, I only explore the sequence predictions for basic gains (Problems 2, 3, 4, 7, 8, and 14). I also built decision trees to analyse the record of each state and action in a total of 60000 tasks for each problem. In order to make it easy to interpret, the figure of the decision tree maintains the most frequent visit states. The decision trees only keep the states in which visit count *n* is larger than 9000 (15% of 60000 choice tasks). To make it easy to compare, option A is always the option with a higher probability, and option B is the one with a lower probability.

A novel contribution of the model is that, by virtue of the sequential decision process, it predicts how action sequences should vary with task type. As can be seen, the sequence predictions show the same pattern in Problems 3 and 7, where humans prefer the option with high probability. Furthermore, the results show the same pattern in Problems 2, 4, and 8, where over half of the subjects chose the option with low probability. In the following, I will reveal the underlying information process mechanism proposed by the proposed sequential model and analyze the two patterns of information gathering processes.

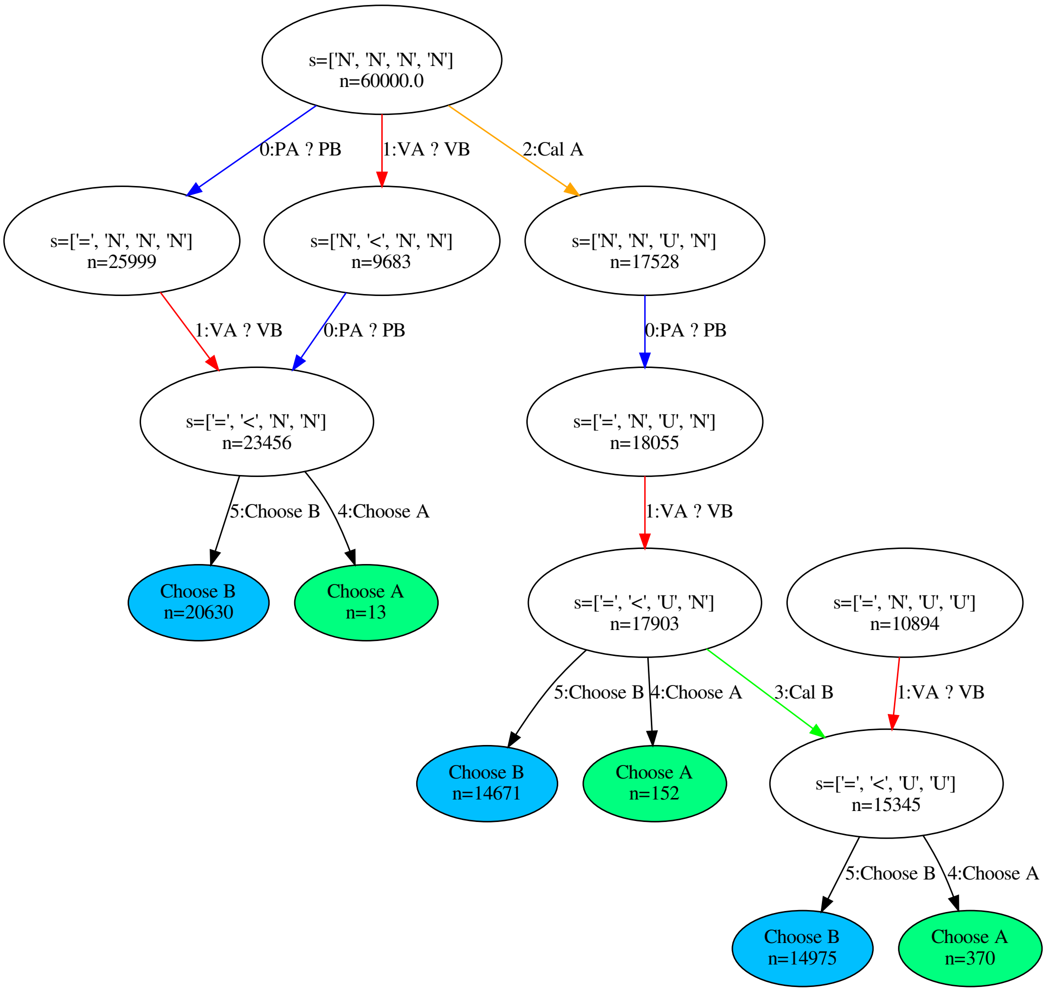


Fig. 6.2 Predicted the most frequent visit states of choices by the Model Zero agent for Problem 2. Comparisons of value attributes are presented in red arrows, and comparisons of probability attributes are presented in blue arrows. Calculations of the expected value are presented in yellow and green arrows for options A (2400, .34) and B (2500, .33), respectively. Finally, choosing A or B is shown in green and blue.

In Problems 2, 4, and 8, on average, the bounded optimal agent took $2 \sim 4$ steps before making a choice, as shown in Figure 6.2. The most frequently visited state before making a choice is $s = [=,<,N,N]$, which means that: the difference between the probability of A and B is within the tolerance value (tolerance $\tau_p = 0.06$), the value of A is less than B, and other information is unknown. The second most frequently visited state is $s = [=, <, U, N]$, which shows that the agents calculated the expected value of the option with high probability based on the most frequently visited state. Finally, the third most frequent visited state is $s = [>, N, U, U]$, which implies that all the computational information about the task is known to the agent before choosing one option. Then I will analyze the most frequent action trajectory. The number of observation actions of the most frequent sequence predictions is 2, as shown in Appendix B. In the first step, it compared the probabilities of two options. The resulting state is that the two options have an equal probability since the difference is within the tolerance value (tolerance $\tau_p = 0.06$). In the second step, it made the comparison in value attributes. The agent can correctly distinguish which option has a high value. In the last step, the agent chose the option with the highest value.

In Problems 3 and 7, on average, the model also took $2 \sim 4$ steps before making a choice, as shown in Figure 6.3. The most frequently visited state before making a choice is $s = [>, N, U, N]$, which means that: the probability of A is greater than B, the observation of the expected value of option A is known to the agent, and other information is unknown. The second most frequently visited state is $s = [>, <, U, U]$, which shows that all the computational information about the task is known to the agent before choosing one option. Finally, the third most frequently visited state is $s = [>, N, U, U]$, which demonstrates that only the ordinal features of the comparison between the values of A and B are unknown. The most frequent sequence predictions are that the agents take 2 observation actions before choosing an option and then terminate the task, as shown in Appendix B. Firstly, it compared the probabilities of two options (A and B). The agent perceives the order accurately since the difference is significantly greater than the tolerance $\tau_p = 0.06$. Secondly, the agent calculated the expected value of the highest probability option (A). Lastly, the agent chose the option with the highest probability. The predictions show that the agent looks at option A in both steps and at option B in only the first step. Then the agent ultimately chose option A more often. This simple pattern corresponds to human eye movements in the risky choice that people choose the gamble they look at more often (Stewart et al., 2016).

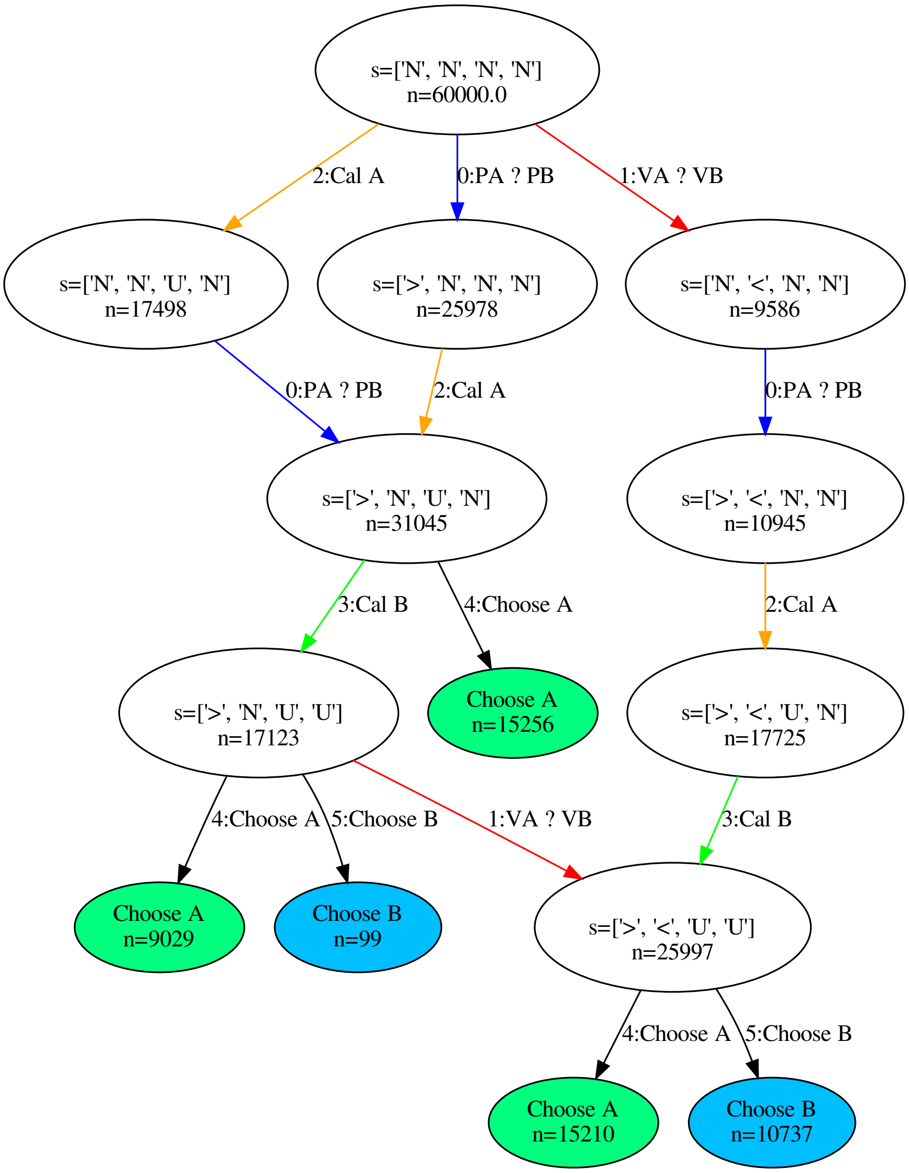


Fig. 6.3 Predicted the most frequent visit states of choices by the Model Zero agent for Problem 3. Option A is (3000, 1.0) and option B is (4000, .80).

The results show that major agents make decisions in all problems before all computational information is gathered. Moreover, the number of observation actions of the most frequent sequence predictions is 2 in all problems. It indicates that the model could learn an optimal way to gather only the information that helps it maximize utility without prior knowledge. The results also show that major agents learn to gather the ordinal features of attribute values before making a choice. These findings are consistent with the observed human behaviour (Noguchi and Stewart, 2014) that pair alternatives are compared on a single attribute dimension using eye-movement data.

The results indicate that the agent mainly compares the probability attributes between options in the first few steps. This prediction is consistent with the argument that people prefer comparisons to the calculation of expected value (Noguchi and Stewart, 2018; Ronayne and Brown, 2017; Stewart et al., 2006; Vlaev et al., 2011). The average proportions of each action type taken by the model in the first step are: comparing the probability of option A and B: 0.43, calculating the expected value of option A: 0.29, comparing the value of option A and B: 0.16, and calculating the expected value of option B: 0.12. For the first step, 59 percent of the actions involve comparing probabilities (43%) or values (16%). These findings are consistent with the phenomenon that people frequently ignore components that the alternatives share, and instead concentrate on the differences between the components (Tversky, 1972). The agent first compares the probability since it is a piece of vital information that would be efficient. Because the option with a higher probability is more likely to have a higher expected value in such an ecological distribution of attribute values, as concluded in Section 5.4.6. In most cases, the observation of the comparison of the probabilities can indicate the option with the highest expected value. Therefore, nearly half of the agents compare the probabilities first. It also demonstrates that the model learned an approximately optimal way to gather information for maximizing rewards.

Overall, in the first place, the model preferred the option with a high probability when there were noisy observations of expected value (e.g., problems 3 and 7). The agent does this because the order of the features indicates that the more probable options are more likely to have a higher expected value in the specific environmental distribution. However, in some scenarios (e.g., problems 2, 4, and 8), it is intractable or not convincing enough to accurately perceive the ordering of probabilities between two options. Then the agent turns to choose the option with a high value. It happens when the difference between the probabilities of the two options is too trivial to distinguish for humans. In other words, the probability difference is within the tolerance $\tau_p = 0.06$. Therefore, choosing the option with a high value is optimal when the probabilities are considered equal.

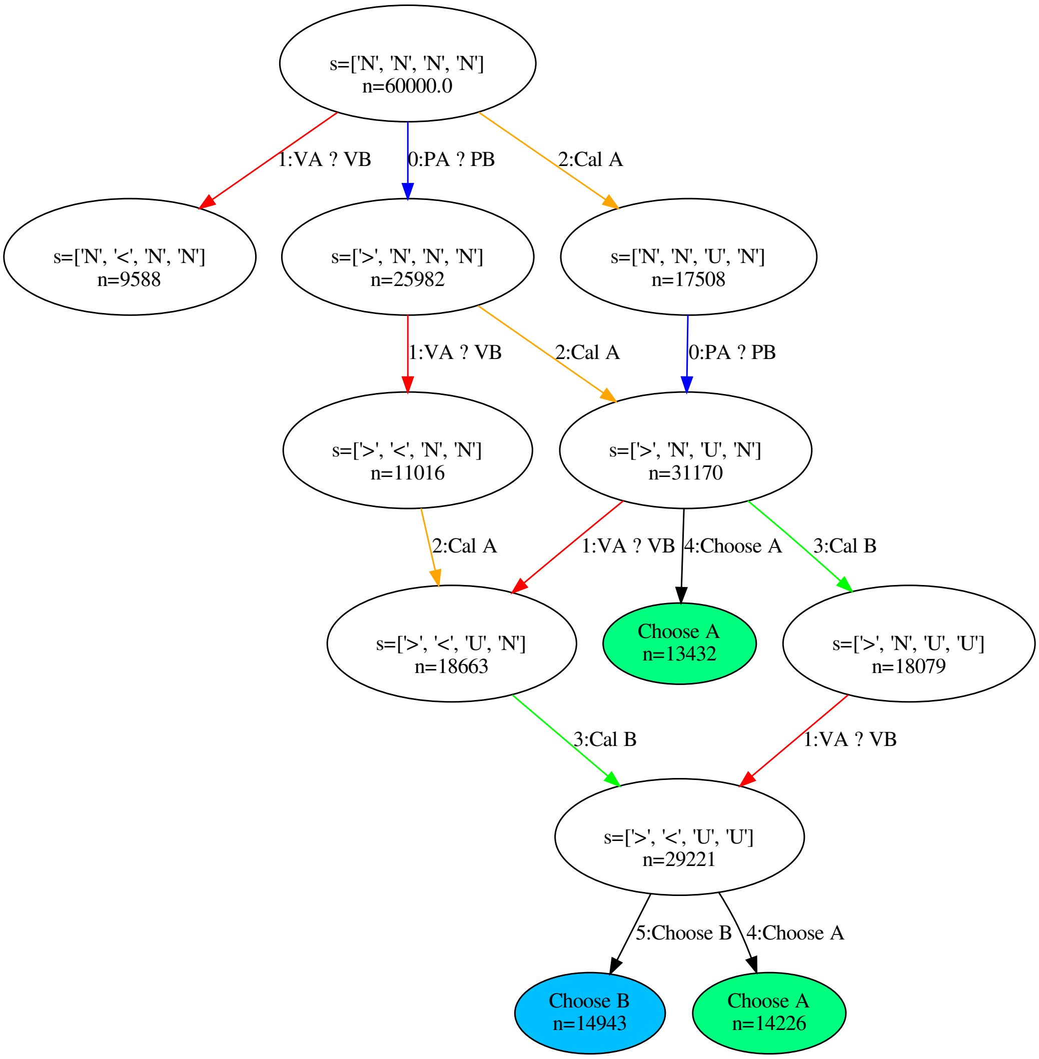


Fig. 6.4 Predicted the most frequent visit states of choices by the Model Zero agent for Problem 7. Option A is (3000, .90) and option B is (6000, .45).

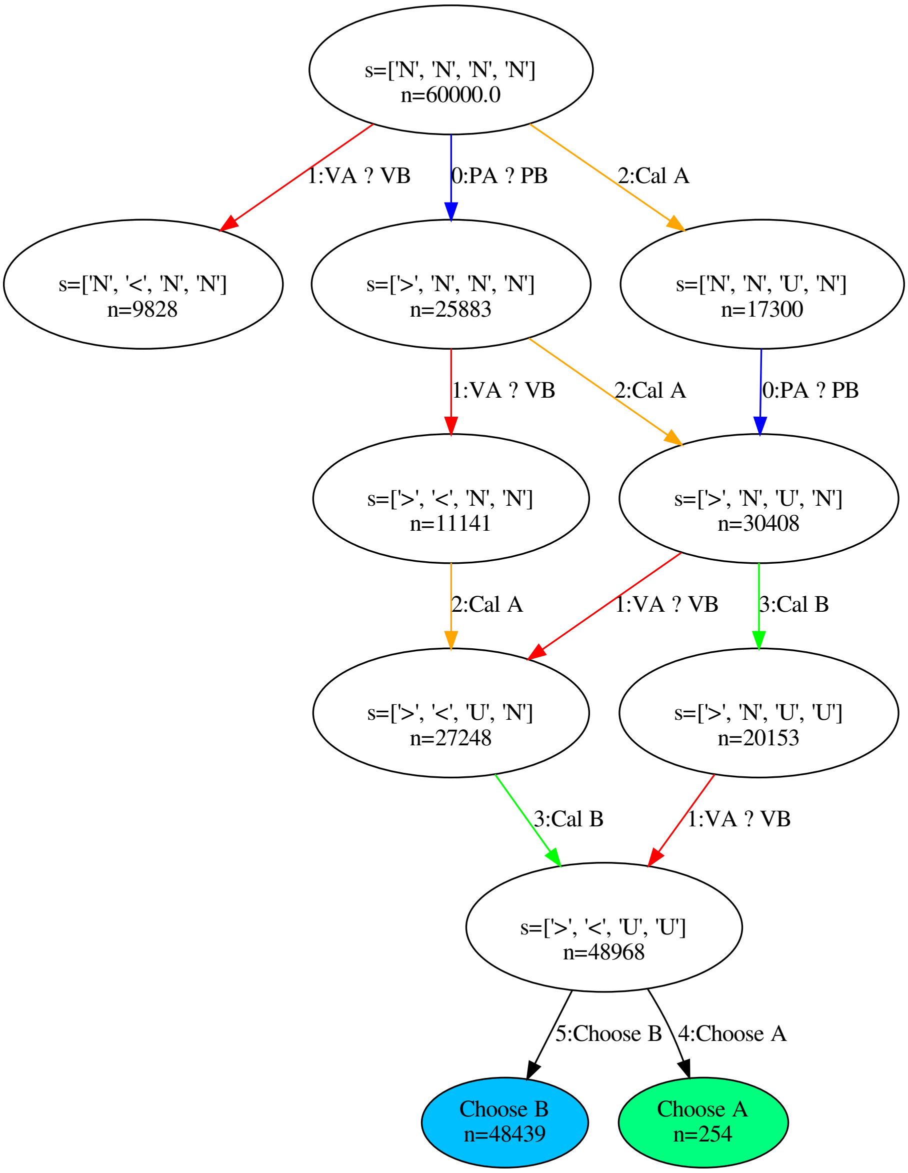


Fig. 6.5 Predicted the most frequent visit states of choices by the Model Zero agent for Problem 14. Option A is (5, 1.0) and option B is (5000, .001).

I note that the observations of the expected value of options also contribute to maximizing expected rewards, especially in hard choice problems (e.g., where the safe gamble is not a sure thing or when probabilities are similar). In Problem 7, when all the information is known ($s = [>, <, U, U]$), the agent did not show a marked preference for the two options, as shown in Figure 6.4. However, in the same state in problem 14, the agent prefers option B, which has a lower small probability, as shown in Figure 6.5. Because the agent learns to overweight the very small probability, as a human does. In such a state, the observations of the expected value of options play an important role in making the decision. The results show that the agents present different preferences in the two situations, although the state is the same before making a choice. It is also well supported that the model could learn to adapt to different environments.

In problem 14, 82 percent of the agents chose an option after gathering all the information. This prediction is not observed in other problems. It is consistent with the findings that people make more fixations or eye movements (and so have longer choice times) on harder choices (e.g., where the safe gamble is not a sure thing or when probabilities are similar) (Stewart et al., 2016).

6.3 Discussion

The results in this Chapter show that the phenomena that were used by Kahneman and Tversky (1979) to support Prospect Theory all emerge from a single computationally rational decision agent using an optimal policy with bounded observations. The single agent, with a single set of parameter values (e.g., noise) generates all of the effects reported in Kahneman and Tversky (1979). Moreover, collectively these effects show that the model generates the fourfold pattern of choice. It appears to be risk averse for gains, risk seeking for losses of high probability; risk seeking for gains and risk averse for losses of low probability.

This Chapter presents a single unifying model of risky choice tasks developed within the framework of computational rationality. The results demonstrate that computation rationality is a promising theoretical framework for integrating various approaches and explaining human behaviour. Various theories or studies have various explanations with different assumptions. Although these approaches succeed in showing that the apparent irrational behaviours are rational given the computational bounds, there are major theoretical questions about the nature of the bounds, which mean the constraints of the environment and the machine (brain or mind). Too often, different rational accounts invoke different bounds to explain behaviour. For example, DFT assumes that attribute differences are accumulated; the PCS model assumes that probabilities and outcomes are integrated; the DbS model assumes that favourable comparisons are accumulated. To avoid the problem arising from different theoretical assumptions, I provide a unifying framework to integrate these approaches and explain human risky behaviour. The approach is an application of computational rationality to the problem of human risky choice tasks. In such a framework, the analysis of human choice would not only provide a clear answer to the question of whether the cognitive biases are rational or not, but also offer insightful predictions about the cognitive processes.

Moreover, I show that the fourfold pattern of risk attitudes observed in economic choice tasks can be explained under simple and uncontroversial assumptions: (1) humans deliberate over options under bounds (e.g., limited time, limited cognitive resources and computation costs); (2) humans aim to maximize rewards; (3) humans make noisy calculation (e.g., by multiplication of a probability by a value) and noisy comparisons (e.g., comparing two probabilities to see which option is riskier) before acting (making a choice).

The focus of this Chapter is on explanations of the fourfold pattern of risky behaviour in terms of the underlying information processes. I asked how humans behave as if they maximize rewards given bounds on the environment and cognitive limits. The results show the influence of the process, i.e., the computationally rational deliberation process. The proposed models capture the apparent deviation in behaviour from normative Expected Utility theory and show that the decision making would be improved by influencing cognitive processing, which links to additional parameters. Furthermore, these new models offer a way to reduce that deviation by showing how deviations arise from adaptation to information processing constraints. Perhaps this is one of the most important contributions of the work.

The results demonstrate that the underlying information processing is organized in order to generate rational decisions. The RL agent is not pre-programmed to gather all information but learns to gather only that information that helps it maximize utility. The decision is reached optimally that the potential benefit of gathering information should be greater than the cost of gathering information. These findings in line with the phenomenon that human decision-makers only sample information from sources that are expected to provide relevant information because sampling all available information would be computationally impossible (Braunlich and Love, 2022). Therefore, the learned policy is an optimal process given the bounds rather than the stochastic process. This prediction conflicts with models that commonly assume that the information process is stochastic because of probabilistic attention switching (Busemeyer et al., 2019; Huang et al., 2012a).

I highlight that my explanations for the computationally rational nature of the fourfold pattern differ sharply from those proposed previously. In the risky choice literature, there is a disconnect between the models of risky choice based on expected utility theory (rational models) and models of cognitive processing (process models). These are two threads of risky choice modelling, each taking a different direction. Commonly, they explain the phenomena by taking a descriptive perspective. In this Chapter, I draw the two threads (rational and process) together by proposing a normative model of the computationally rational process in risky choice. This work suggests an answer to the longstanding questions of why and how humans make risky choices in an approximately optimal way. I shed light on the two questions by taking a normative account of the underlying information gathering processes.

Chapter 7

General Discussion

In this thesis, I have argued that a large range of human risky choice phenomena are a consequence of computationally rational processing. Specifically, I have shown that:

- contextual choice effects emerge in neural networks that are optimised to maximise correct choices for perceptual decision tasks that support comparison.
- contextual choice effects emerge in reinforcement learners that are optimised to maximise trade-offs between the points obtained from choice outcomes and the cost of information gathering.
- the fourfold pattern of choice emerges from reinforcement learner models that are optimised to maximise the accumulated rewards of the choice tasks.

These findings support the contention that *apparent* cognitive biases emerge from computationally rational processing. They also support the view that computational rationality, which defines bounded optimality problems faced by people, explored through deep RL, provides a unifying framework for modelling risky choice phenomena. Furthermore, deep RL has the potential to help discriminate between various explanations because it provides a means of computing approximately optimal policies given both ecological and cognitive bounds.

The results reported in the thesis demonstrate that the observations made by the model (comparisons and calculations) are vital for explaining the phenomena of risky choice. In the perceptual choice tasks, the different representation formats of the same symbol value led to different decisions with choice reversals only present with a representation that supported comparison. One plausible explanation is that the representation format affects the cognitive processes on which the estimations are based. In the inference choice tasks, the observation is the calculation or comparison. The agents are adapted to the various noise levels and time costs of the task environment. Therefore, observation has a strong influence on the learned information processes. As a consequence, the thesis provides a novel theory of risky choice. This theoretical commitment to the central role of the observation function in explaining risky choice connects the task representation to the utility function. The observation modulates the information gathering process. Meanwhile, evidence accumulation generates the observation. This interaction is a dynamic process.

The difference between my RL model of risky choice and Kahneman and Tversky's (K&T) prospect theory can be illustrated by considering the information flow. The information flow in K&T is: symbolic stimulus → heuristic editing to bias the utility function → choice. In contrast, the information flow in my RL models is: symbolic stimulus → bounded observation → optimal estimation → optimal choice where the optimality of choice is determined by learning to utility maximise through experience. My model of risky choice is the first to combine optimal estimation, optimal choice and optimal active information gathering in the service of choice. In this respect, the models reported in this thesis are a departure from previous models.

An important feature of my model concerns its use of an unbiased utility function. In the risky choice literature, much emphasis is put on different models of bias in the utility function. These theories can be viewed as descriptive models of human decision making. The assumption is that people make a choice that is *unbounded* except for a bias in a subjective utility; the subjective utility function plays a central role in explaining human behaviour. In contrast, my model uses expected utility instead of subjective utility – there is no bias on the utility function in any of the models reported in this thesis. The predictions made by these models explain human choice behaviour by virtue of bounded optimality, not by virtue of biased utility. The bounds concern limited cognitive capacity and partial observations of the environment.

Our results provide a new way to model risky choices. This approach emphasises the originally expected utility and the observation. This model also predicts the information gathering process of human choice.

For perceptual risky choice, the neural network model uses high-dimension input and could capture the effect of representation format. This is novel since previous models take low-dimension as the input, such as symbolic values.

In this work, information processing is organised in order to generate rational decisions. It contrasts with the hypothesis made by a list of models, which assume decision making as a stochastic process, e.g., random walk and diffusion process. The presented results show a computation rational process conforming to the environment and cognitive bounds. Consequently, the learnt policy is an optimal process given the constraints rather than a stochastic process.

7.1 Future work

While the models in this thesis have provided some evidence that risky choice phenomena emerge from computationally rational processes, much could be done to make a more robust argument.

A first priority might be to integrate the neural network models of perception with the RL models of sequential processing. While these are independently useful, both perceptual and symbolic risky choice phenomena arise through the processing of a single human information processing system; this remains to be explained. There are a number of approaches to integration, but the best might well be to take an end-to-end approach like that used in many current RL applications. This would involve providing a bit array input for all problems – not just rectangles and bars but also problems specified with numeric symbols. The policy network would be trained on the raw bit-array input. A further extension would involve the implementation of foveated vision so that partial observations are made of the bit-array at each time step.

Further testing of the predictions of the existing models reported in the thesis is also necessary. Tests should include testing an extended range of contextual phenomena, including, for example, phantom decoy effects and distance effects. Further sensitivity analysis is also needed. The models predict risky choice phenomena across a wide range of their parameter values but not across the entire range. These ranges could be more closely documented than they have been in this thesis, and the implications for when humans (and other species) exhibit risky choice phenomena and when they do not could be explored.

There are many other biases other than those typically associated with risky choices. These include confirmation bias, choice overload, base rate fallacy, availability, and anchoring. It is possible that the computational approach explored in this thesis could be applied to these additional biases in an effort to explain *why* these phenomena arise as a consequence of bounds rather than explain them away as irrationalities. For example, choice overload might simply be a rational response to the time cost of considering additional options. The base rate fallacy might be a response to uncertainty in the encoding of statistical information, much as appears to be the case with contextual choice effects.

7.2 Conclusion

In conclusion, this thesis has demonstrated that well-known phenomena, concerning how people make risky choices, can be explained as the consequence of utility maximisation given bounded observations of the choice problem. The conclusion is supported by a series of neural network and reinforcement learning models that find approximately optimal solutions to risky choice problems.

References

Acharya, A., Chen, X., Myers, C. W., Lewis, R. L., and Howes, A. (2017). Human visual search as a deep reinforcement learning solution to a pomdp. In *CogSci*, pages 51–56.

Acuna, D. and Schrater, P. R. (2009). Structure learning in human sequential decision-making. In *Advances in neural information processing systems*, pages 1–8.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14(3):471–485.

Anderson, J. R. et al. (1990). *The Adaptive Character of Thought*. Psychology Press.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38.

Ba, J., Mnih, V., and Kavukcuoglu, K. (2015). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Baker, C., Saxe, R., and Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 33 (33)*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10.

Barth-Maron, G., Hoffman, M. W., Budden, D., Dabney, W., Horgan, D., Muldal, A., Heess, N., and Lillicrap, T. (2018). Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*.

Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers Inc.

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

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bossaerts, P. and Murawski, C. (2017). Computational complexity and human decision-making. *Trends in cognitive sciences*, 21(12):917–929.

Bowman, H. and Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological review*, 114(1):38.

Braunlich, K. and Love, B. C. (2022). Bidirectional influences of information sampling and concept learning. *Psychological review*, 129(2):213.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Busemeyer, J. R., Gluth, S., Rieskamp, J., and Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in cognitive sciences*.

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.

Callaway, F., Jain, Y. R., van Opheusden, B., Das, P., Iwama, G., Gul, S., Krueger, P. M., Becker, F., Griffiths, T. L., and Lieder, F. (2022a). Leveraging artificial intelligence to improve people's planning strategies. *Proceedings of the National Academy of Sciences*, 119(12):e2117432119.

Callaway, F., Rangel, A., and Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology*, 17(3):e1008863.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., and Lieder, F. (2022b). Rational use of cognitive resources in human planning. *Nature Human Behaviour*.

Cataldo, A. M. and Cohen, A. L. (2018). Reversing the similarity effect: The effect of presentation format. *Cognition*, 175:141–156.

Cataldo, A. M. and Cohen, A. L. (2019). The comparison process as an account of variation in the attraction, compromise, and similarity effects. *Psychonomic bulletin & review*, 26(3):934–942.

Chater, N., Felin, T., Funder, D. C., Gigerenzer, G., Koenderink, J. J., Krueger, J. I., Noble, D., Nordli, S. A., Oaksford, M., Schwartz, B., et al. (2018). Mind, rationality, and cognition: An interdisciplinary debate. *Psychonomic Bulletin & Review*, 25(2):793–826.

Chater, N. and Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, 3(2):57–65.

Chen, X. (2015). *An optimal control approach to testing theories of human information processing constraints*. PhD thesis, University of Birmingham.

Chen, X., Bailly, G., Brumby, D. P., Oulasvirta, A., and Howes, A. (2015). The emergence of interactive behavior: A model of rational menu search. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4217–4226. ACM.

Chen, X., Starke, S. D., Baber, C., and Howes, A. (2017). A cognitive model of how people make decisions through interaction with visual displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1205–1216. ACM.

Choi, J., Chang, H. J., Fischer, T., Yun, S., Lee, K., Jeong, J., Demiris, Y., and Choi, J. Y. (2018). Context-aware deep feature compression for high-speed visual tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 479–488.

Choi, J., Chang, H. J., Jeong, J., Demiris, Y., and Choi, J. Y. (2016). Visual tracking using attention-modulated disintegration and integration. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4321–4330.

Choi, J., Chang, H. J., Yun, S., Fischer, T., Demiris, Y., and Choi, J. Y. (2017). Attentional correlation filter network for adaptive visual tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4828–4837.

Cichy, R. M. and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and brain sciences*, 26(2):139–153.

Cushman, F. and Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, 112(45):13817–13822.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural computation*, 18(7):1637–1677.

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. (2018). Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*.

Felin, T., Koenderink, J., and Krueger, J. I. (2017). Rationality, perception, and the all-seeing eye. *Psychonomic Bulletin & Review*, 24(4):1040–1059.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

Fiedler, S. and Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in psychology*, 3:335.

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., and Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature neuroscience*, 22(12):2066–2077.

Findling, C. and Wyart, V. (2021). Computation noise in human learning and decision-making: Origin, impact, function. *Current Opinion in Behavioral Sciences*, 38:124–132.

Fishburn, P. C. (1970). Utility theory for decision making. Technical report, Research analysis corp McLean VA.

Fontanesi, L., Gluth, S., Spektor, M. S., and Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review*, 26(4):1099–1121.

Forstmann, B. U., Ratcliff, R., and Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology*, 67:641–666.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., et al. (2018). Noisy networks for exploration. In *International Conference on Learning Representations*.

Frazier, P. and Angela, J. Y. (2008). Sequential hypothesis testing under stochastic deadlines. In *Advances in neural information processing systems*, pages 465–472.

Friedman, M. and Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of political Economy*, 56(4):279–304.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Garcez, A. d., Bader, S., Bowman, H., Lamb, L. C., de Penning, L., Illuminoo, B., Poon, H., and Gerson Zaverucha, C. (2022). Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342:1.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. *Blackwell handbook of judgment and decision making*, 62:88.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, 5(3-4):303–336.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, 62:451–482.

Gigerenzer, G., Todd, P. M., Group, A. R., et al. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Glöckner, A. and Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, 3(3):215–228.

Glöckner, A. and Herbold, A.-K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*, 24(1):71–98.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, 30:535–574.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., and Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30.

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

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258.

Hasselt, H. V. (2010). Double q-learning. In *Advances in Neural Information Processing Systems*, pages 2613–2621.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, L., Zhao, W. J., and Bhatia, S. (2020). An ontology of decision models. *Psychological Review*.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Horvitz, E. J. (1987). Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pages 429–447. AUAI Press.

Howes, A., Chen, X., Acharya, A., and Lewis, R. L. (2018). Interaction as an emergent property of a partially observable markov decision process. *Computational interaction design*, pages 287–310.

Howes, A., Lewis, R. L., and Singh, S. (2014). Utility maximization and bounds on human information processing. *Topics in cognitive science*, 6(2):198–203.

Howes, A., Lewis, R. L., and Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological review*, 116(4):717.

Howes, A., Warren, P. A., Farmer, G., El-Deredy, W., and Lewis, R. L. (2016). Why contextual preference reversals maximize expected value. *Psychological review*, 123(4):368.

Huang, K., Sen, S., and Szidarovszky, F. (2012a). Connections among decision field theory models of cognition. *Journal of Mathematical Psychology*, 56(5):287–296.

Huang, Y., Hanks, T., Shadlen, M., Friesen, A. L., and Rao, R. P. (2012b). How prior probability influences decision making: A unifying probabilistic model. In *Advances in neural information processing systems*, pages 1268–1276.

Igl, M., Zintgraf, L., Le, T. A., Wood, F., and Whiteson, S. (2018). Deep variational reinforcement learning for pomdps. In *International Conference on Machine Learning*, pages 2122–2131.

Janz, D., Hron, J., Mazur, P., Hofmann, K., Hernández-Lobato, J. M., and Tschiatschek, S. (2019). Successor uncertainties: exploration and uncertainty in temporal difference learning. In *Advances in Neural Information Processing Systems*, pages 4509–4518.

Jones, M. and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and brain sciences*, 34(4):169.

Juechems, K., Balaguer, J., Spitzer, B., and Summerfield, C. (2021). Optimal utility and probability functions for agents with finite computational precision. *Proceedings of the National Academy of Sciences*, 118(2):e2002232118.

Kacelnik, A. and Bateson, M. (1996). Risky theories—the effects of variance on foraging decisions. *American Zoologist*, 36(4):402–434.

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. (2016). 36 heuristics and biases. *Scientists Making a Difference: One Hundred Eminent Behavioral and Brain Scientists Talk about their Most Important Contributions*, page 171.

Kahneman, D., Rosenfield, A. M., Gandhi, L., and Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard business review*, 94(10):38–46.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Khalvati, K., Park, S. A., Dreher, J.-C., and Rao, R. P. (2016). A probabilistic model of social decision making based on reward maximization. In *Advances in Neural Information Processing Systems*, pages 2901–2909.

Khalvati, K. and Rao, R. P. (2015). A bayesian framework for modeling confidence in perceptual decision making. In *Advances in neural information processing systems*, pages 2413–2421.

Khaw, M. W., Li, Z., and Woodford, M. (2017). Risk aversion as a perceptual bias. Technical report, National Bureau of Economic Research.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13(10):1292–1298.

Krajbich, I. and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kruegera, P. M., Callawayb, F., Gulc, S., Griffithsa, T. L., and Liederd, F. (2022). Discovering rational heuristics for risky choice. *decision-making*, 10:14.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Lee, K., Ognibene, D., Chang, H. J., Kim, T., and Demiris, Y. (2015). STARE: Spatio-temporal attention relocation for multiple structured activities detection. *IEEE Transactions on Image Processing*, 24(12):5916–5927.

Lewis, R. L., Howes, A., and Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6(2):279–311.

Lieder, F., Griffiths, T., and Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems*, pages 2690–2798.

Lieder, F. and Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, 124(6):762.

Lieder, F. and Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, pages 1–85.

Lieder, F., Krueger, P. M., and Griffiths, T. (2017). An automatic method for discovering rational heuristics for risky choice. In *CogSci*.

Lin, L.-J. (1992). *Reinforcement learning for robots using neural networks*. PhD thesis, Carnegie Mellon University.

Littman, M. L. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521(7553):445.

Loomes, G. (2019). Taking process into account when modeling risky choice. *Handbook of Research Methods and Applications in Experimental Economics*.

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

Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J., and Uchida, N. (2020). Distributional reinforcement learning in the brain. *Trends in Neurosciences*.

Luo, X., Roads, B. D., and Love, B. C. (2021). The costs and benefits of goal-directed attention in deep convolutional neural networks. *Computational Brain & Behavior*, 4(2):213–230.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13132–13143.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin*, 15:470–488.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1):11–38.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., and Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8):348–356.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Milli, S., Lieder, F., and Griffiths, T. L. (2017). When does bounded-optimal metareasoning favor few cognitive systems? In *AAAI*, pages 4422–4428.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Navarro-Martinez, D., Loomes, G., Isoni, A., Butler, D., and Alaoui, L. (2018). Boundedly rational expected utility theory. *Journal of risk and uncertainty*, 57(3):199–223.

Niv, Y., Edlund, J. A., Dayan, P., and O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562.

Noguchi, T. and Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, 132(1):44–56.

Noguchi, T. and Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological review*, 125(4):512.

Oaksford, M. and Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M. and Chater, N. (2009). Précis of bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, 32(1):69–84.

Osawa, K., Swaroop, S., Khan, M. E. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. (2019). Practical deep learning with bayesian principles. In *Advances in Neural Information Processing Systems*, pages 4289–4301.

Oulasvirta, A., Bi, X., and Howes, A. (2018). *Computational Interaction*. Oxford University Press.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. (2018). The uncertainty bellman equation and exploration. In *International Conference on Machine Learning*, pages 3836–3845.

Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O., and Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. *Journal of Experimental Psychology: General*, 147(2):147.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., and Xu, F. (2011). A tutorial introduction to bayesian models of cognitive development. *Cognition*, 120(3):302–321.

Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, pages 1–22.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9617–9626.

Pettibone, J. C. (2012). Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgment and Decision Making*, 7(4):513.

Platt, M. L. and Huettel, S. A. (2008). Risky business: the neuroeconomics of decision making under uncertainty. *Nature neuroscience*, 11(4):398.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018). Machine theory of mind. In *International Conference on Machine Learning*, pages 4218–4227.

Rahnev, D. and Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, 41.

Rao, R. P. (2010). Decision making under uncertainty: a neural model based on partially observable markov decision processes. *Frontiers in computational neuroscience*, 4:146.

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

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

Ratcliff, R. and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, 111(2):333.

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

Roe, R. M., Busemeyer, J. R., and Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological review*, 108(2):370.

Ronayne, D. and Brown, G. D. (2017). Multi-attribute decision by sampling: an account of the attraction, compromise and similarity effects. *Journal of Mathematical Psychology*, 81:11–27.

Russell, S. J. (1997). Rationality and intelligence. *Artificial intelligence*, 94(1-2):57–77.

Russell, S. J. and Subramanian, D. (1994). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609.

Russell, S. J. and Wefald, E. (1991). *Do the right thing: studies in limited rationality*. MIT press.

Sanborn, A. N. and Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2020a). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., and DiCarlo, J. J. (2020b). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., and Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, 40:99–124.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *ICML*.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63(2):129.

Simon, H. A. (1978). Rationality as process and as product of thought. *The American economic review*, 68(2):1–16.

Simon, H. A. (1979). Information processing models of cognition. *Annual review of psychology*, 30(1):363–396.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simsek, O., Algorta, S., and Kothiyal, A. (2016). Why most decisions are easy in tetris—and perhaps in other sequential decision problems, as well. In *International Conference on Machine Learning*, pages 1757–1765.

Song, M., Wang, X., Zhang, H., and Li, J. (2019). Proactive information sampling in value-based decision-making: Deciding when and where to saccade. *Frontiers in human neuroscience*, 13:35.

Spektor, M. S., Gluth, S., Fontanesi, L., and Rieskamp, J. (2019). How similarity between choice options affects decisions from experience: The accentuation-of-differences model. *Psychological review*, 126(1):52.

Spektor, M. S., Kellen, D., and Hotaling, J. M. (2018). When the good looks bad: An experimental exploration of the repulsion effect. *Psychological science*, 29(8):1309–1320.

Steiner, J. and Stewart, C. (2016). Perceiving prospects properly. *American Economic Review*, 106(7):1601–31.

Stewart, N., Chater, N., and Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, 53(1):1–26.

Stewart, N., Hermens, F., and Matthews, W. J. (2016). Eye movements in risky choice. *Journal of behavioral decision making*, 29(2-3):116–136.

Stocker, A. A. and Simoncelli, E. P. (2006a). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578.

Stocker, A. A. and Simoncelli, E. P. (2006b). Sensory adaptation within a bayesian framework for perception. In *Advances in neural information processing systems*, pages 1289–1296.

Sturm, T. (2012). The "rationality wars" in psychology: Where they are and where they could go. *Inquiry*, 55(1):66–81.

Sun, R. and Giles, C. L. (2001). Sequence learning: from recognition and prediction to sequential decision making. *IEEE Intelligent Systems*, 16(4):67–70.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Todd, P. M. and Gigerenzer, G. E. (2012). *Ecological rationality: Intelligence in the world.* Oxford University Press.

Trueblood, J. S. (2012). Multialternative context effects obtained using an inference task. *Psychonomic bulletin & review*, 19(5):962–968.

Trueblood, J. S., Brown, S. D., and Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, 121(2):179.

Trueblood, J. S., Brown, S. D., Heathcote, A., and Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision making. *Psychological science*, 24(6):901–908.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., and Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, 113(11):3102–3107.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323.

Tversky, A., Kahneman, D., et al. (1986). Rational choice and the framing of decisions. *The Journal of Business*, 59(4):251–278.

Tversky, A. and Simonson, I. (1993). Context-dependent preferences. *Management science*, 39(10):1179–1189.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3):550.

Van Gerven, M. (2017). Computational foundations of natural intelligence. *Frontiers in computational neuroscience*, 11:112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vlaev, I., Chater, N., Stewart, N., and Brown, G. D. (2011). Does the brain calculate value? *Trends in cognitive sciences*, 15(11):546–554.

Von Neumann, J. and Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton university press.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. (2016a). Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., and Freitas, N. (2016b). Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1995–2003.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, King's College, Cambridge.

Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(4):767.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pages 5–32. Springer.

Wollschlaeger, L. M. and Diederich, A. (2020). Similarity, attraction, and compromise effects: Original findings, recent empirical observations, and computational cognitive process models. *The American Journal of Psychology*, 133(1):1–30.

Wyble, B., Bowman, H., and Nieuwenstein, M. (2009). The attentional blink provides episodic distinctiveness: sparing at a cost. *Journal of experimental psychology: Human perception and performance*, 35(3):787.

Wyble, B., Potter, M. C., Bowman, H., and Nieuwenstein, M. (2011). Attentional episodes in visual perception. *Journal of Experimental Psychology: General*, 140(3):488.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Appendix A

Parametric Details of the Choice Tasks Used in Each Simulation

The values used to model the choice tasks were those used by Trueblood et al. (2013) and presented in Supplemental Material [1]. I produce the values of options in Table A.1, A.2, and A.3. The locations of the options for the task 6 are presented in Figure A.1, A.2, and A.3.

Table A.1 The values of the choice task set for attraction effect. X and Y are the two options. Rx and Ry are the range decoys for X and Y, respectively, subtracting a random number in the interval [7, 9] from the appropriate attribute value for the target option. Fx and Fy are the frequency decoys for X and Y, respectively, calculated in the same way as range decoys. RFx and RFy are the range-frequency decoys for X and Y, respectively, subtracting a random number in the interval [4, 5] from both attribute values for the target option.

Task	X	Y	Rx	Ry	Fx	Fy	RFx	RFy
1:	45, 75	75, 45	37, 75	75, 37	45, 67	67, 45	40, 70	70, 40
2:	46, 76	76, 46	38, 76	76, 38	46, 68	68, 46	41, 71	71, 41
3:	47, 77	77, 47	38, 77	77, 38	47, 68	68, 47	43, 73	73, 43
4:	48, 78	78, 48	41, 78	78, 41	48, 71	71, 48	44, 74	74, 44
5:	49, 79	79, 49	41, 79	79, 41	49, 71	71, 49	44, 74	74, 44
6:	50, 80	80, 50	41, 80	80, 41	50, 71	71, 50	46, 76	76, 46
7:	51, 81	81, 51	42, 81	81, 42	51, 72	72, 51	46, 76	76, 46
8:	52, 82	82, 52	45, 82	82, 45	52, 75	75, 52	48, 78	78, 48
9:	53, 83	83, 53	45, 83	83, 45	53, 75	75, 53	49, 79	79, 49
10:	54, 84	84, 54	45, 84	84, 45	54, 75	75, 54	49, 79	79, 49

[1] https://journals.sagepub.com/doi/suppl/10.1177/0956797612464241/suppl_file/DS_10.1177_0956797612464241.pdf

Table A.2 The values for compromise effect. The compromise decoy values are calculated by subtracting a random number in the interval [15, 25] from the appropriate attribute value for the target option, then calculating the remaining attribute value that makes the multiplication of the two attributes the same as the target option.

Task	X	Y	Cx	Cy
1:	57, 53	75, 40	36, 84	93, 32
2:	58, 54	76, 41	35, 89	98, 32
3:	59, 55	77, 42	40, 81	98, 33
4:	60, 56	78, 43	44, 76	96, 35
5:	61, 57	79, 44	37, 94	100, 35
6:	62, 58	80, 45	46, 78	101, 36
7:	63, 59	81, 46	48, 77	102, 36
8:	64, 60	82, 47	40, 96	102, 38
9:	65, 61	83, 48	45, 88	105, 38
10:	66, 62	84, 49	49, 84	101, 41

Table A.3 The values for similarity effect set. For the Y option, there are two different X options (X1 and X2). The similarity decoy values are calculated in the same way as range decoys, but with the random number in the interval [3, 5].

Task	X1	Y	Sx1	Sy1	Sx2	Sy2
1:	70, 43	55, 55	75, 40	51, 59	38, 79	59, 51
2:	71, 44	56, 56	80, 39	53, 59	40, 78	60, 52
3:	72, 45	57, 57	79, 41	53, 61	42, 77	62, 52
4:	73, 46	58, 58	82, 41	55, 61	41, 82	62, 54
5:	74, 47	59, 59	81, 43	56, 62	42, 83	63, 55
6:	75, 48	60, 60	82, 44	57, 63	45, 80	64, 56
7:	76, 49	61, 61	81, 46	56, 66	45, 83	66, 56
8:	77, 50	62, 62	84, 46	58, 66	47, 82	66, 58
9:	78, 51	63, 63	85, 47	58, 68	47, 85	68, 58
10:	79, 52	64, 64	86, 48	61, 67	49, 84	68, 60

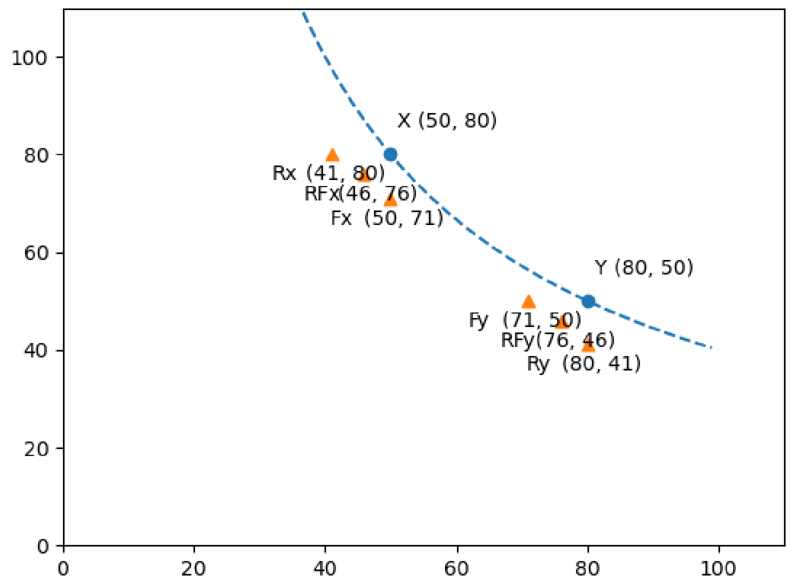


Fig. A.1 The locations of the options in the attraction effect set.

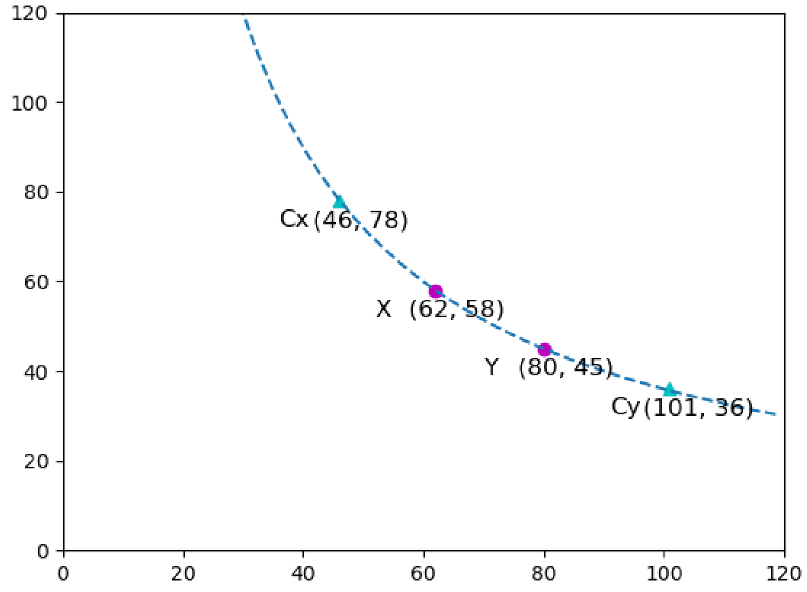


Fig. A.2 The locations of the options in the compromise effect set.

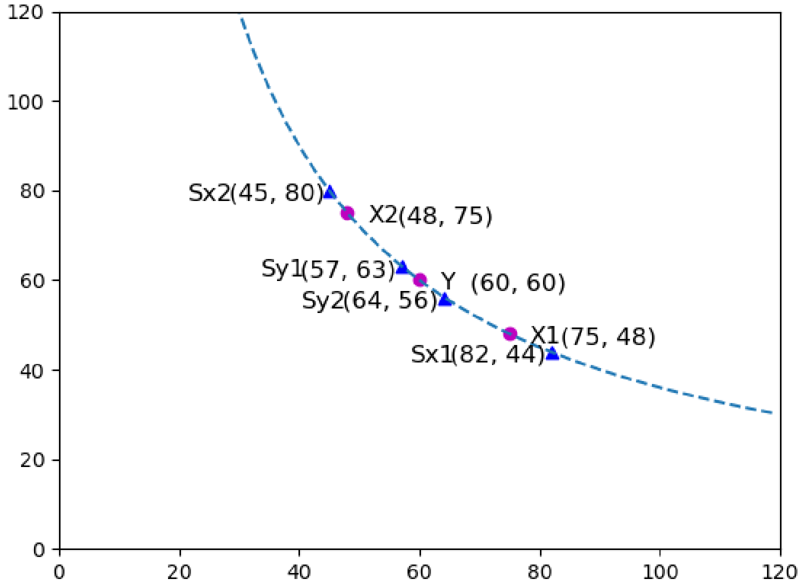


Fig. A.3 The locations of the options in the similarity effect set.

Following the experiment in Cataldo and Cohen (2019), the values presented in the bars are listed in Table A.4. In the first row, the values are used in the paper, and we generate the values according to them. The locations of the options for task 1 are presented in Figure A.4.

Table A.4 The values are given in pixels for the length of the bars in the experiment. X and Y are the two options. Ax and Ay are the attraction decoys for X and Y, respectively. Cx and Cy are the compromise decoys for X and Y, respectively. Sx and Sy are the similarity decoys for X and Y, respectively.

Task	X	Y	Ax	Ay	Cx	Cy	Sx	Sy
1:	6, 10	10, 6	5, 9	9, 5	2, 14	14, 2	5, 11	11, 5
2:	4, 8	8, 4	3, 7	7, 3	1, 11	11, 1	3, 9	9, 3
3:	4, 10	10, 4	3, 9	9, 3	1, 13	13, 1	3, 11	11, 3
4:	5, 9	9, 5	4, 8	8, 4	2, 12	12, 2	4, 10	10, 4
5:	6, 8	8, 6	5, 7	7, 5	3, 11	11, 3	5, 9	9, 5
6:	7, 9	9, 7	6, 8	8, 6	3, 13	13, 3	6, 10	10, 6
7:	5, 11	11, 5	4, 10	10, 4	1, 15	15, 1	4, 12	12, 4

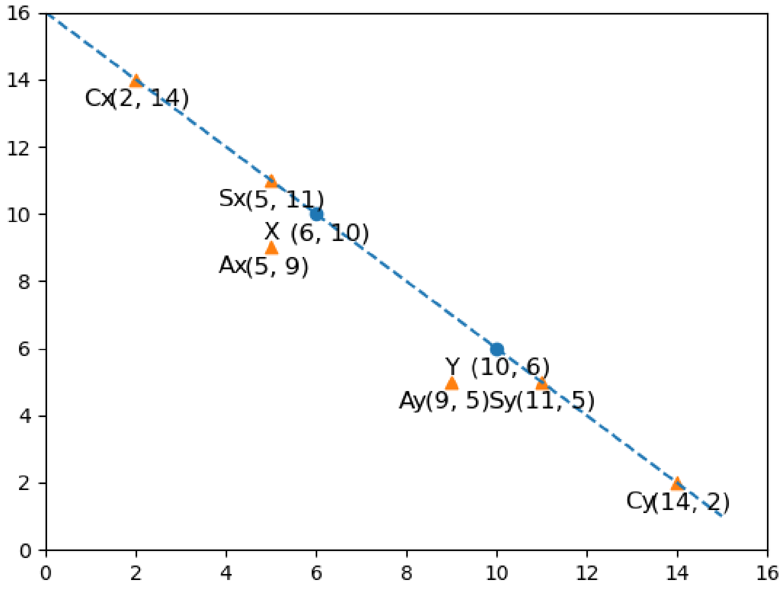


Fig. A.4 The locations of the options in the bars set.

Appendix B

Information Gathering Process Predictions

In order to make it easy to interpret, the figure of the decision tree maintains the most frequent process. The decision trees only keep the states of which visit count *n* is larger than 6000 (10% of 60000 choice tasks).

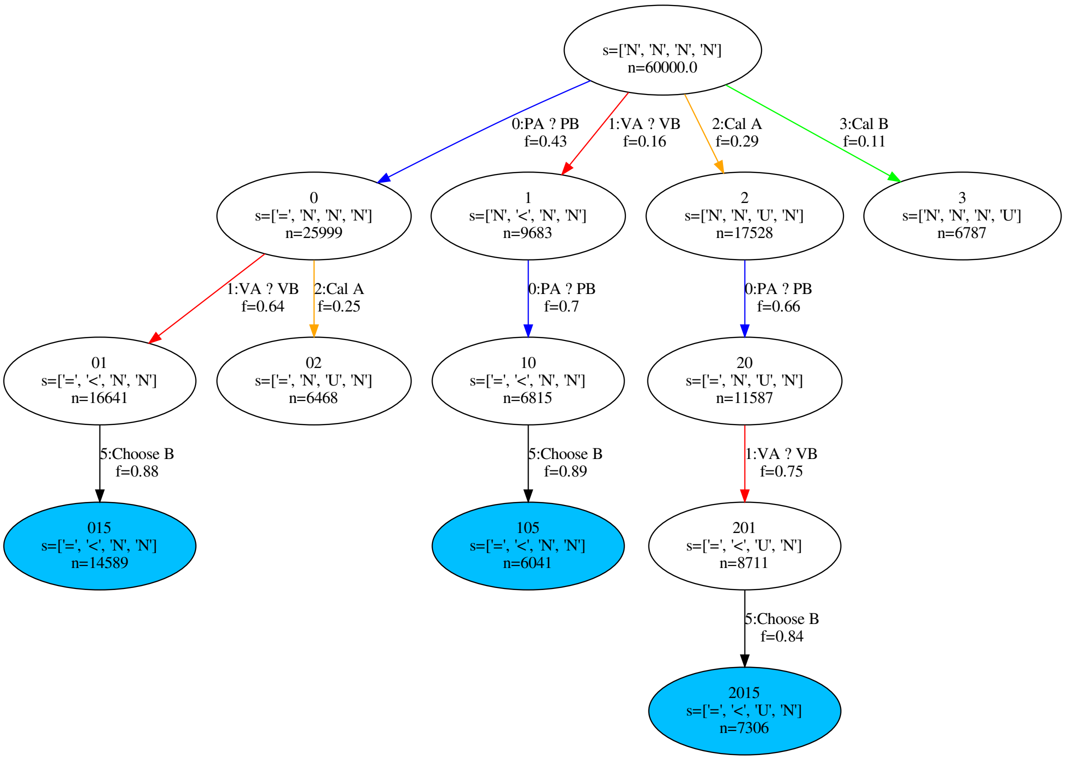


Fig. B.1 Predicted the most frequent sequence of choices by the Model Zero agent for Problem 2. Option A is (2400, .34) and option B is (2500, .33).

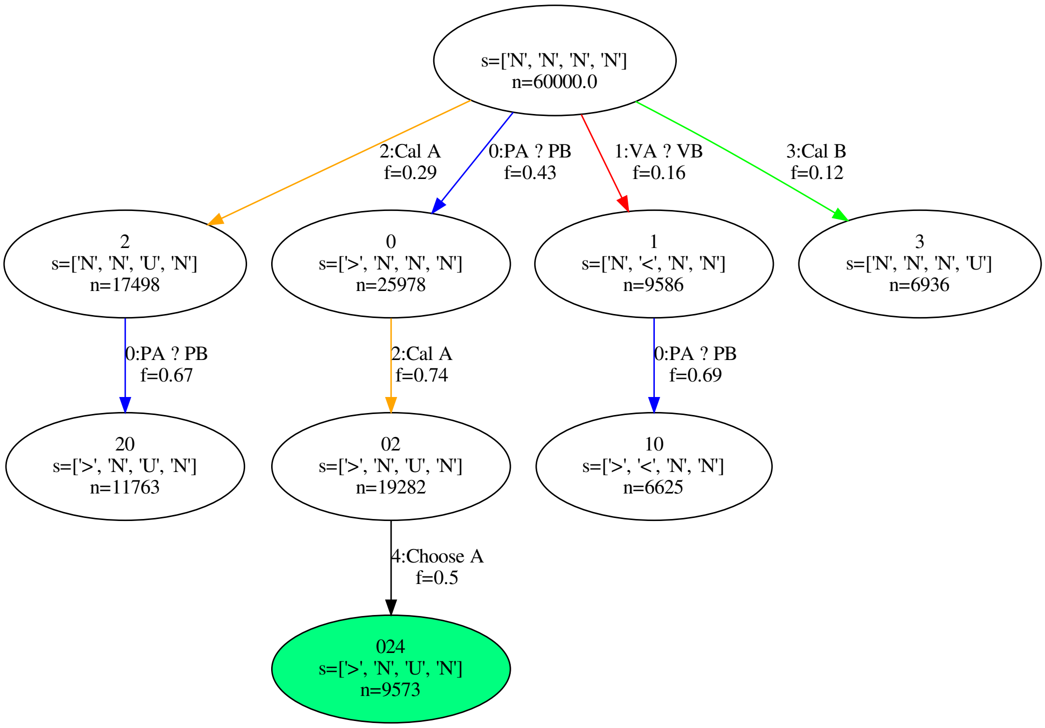


Fig. B.2 Predicted the most frequent sequence of choices by the Model Zero agent for Problem 3. Option A is (3000, 1.0) and option B is (4000, .80).

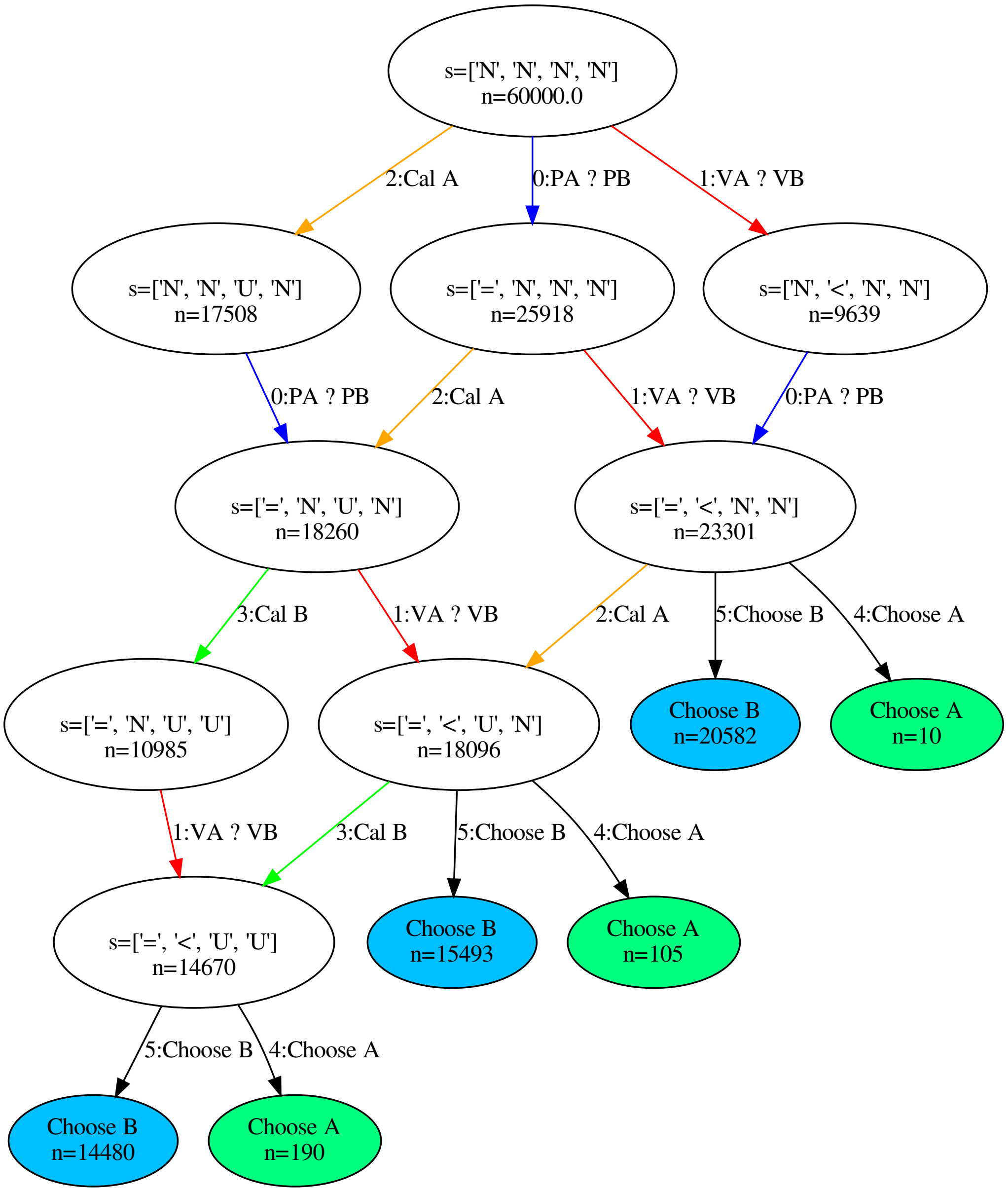


Fig. B.3 Predicted the most frequent visit states of choices by the Model Zero agent for Problem 4. Option A is (3000, .25) and option B is (4000, .20).

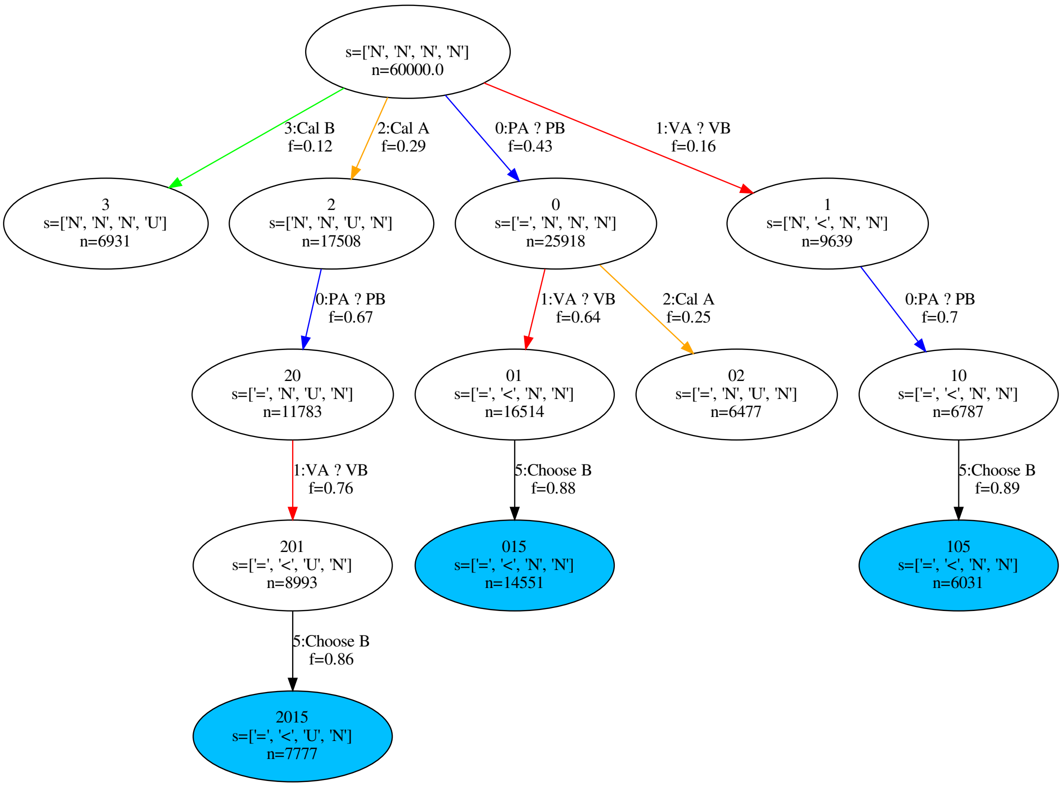


Fig. B.4 Predicted the most frequent sequence of choices by the Model Zero agent for Problem 4. Option A is (3000, .25) and option B is (4000, .20).

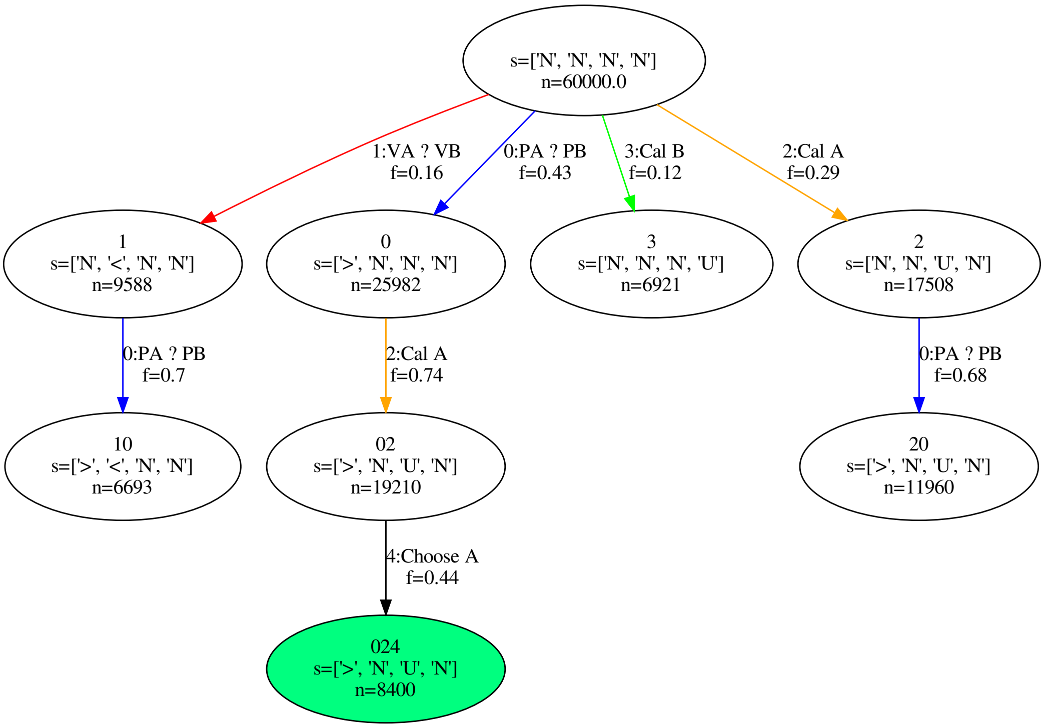


Fig. B.5 Predicted the most frequent sequence of choices by the Model Zero agent for Problem 7. Option A is (3000, .90) and option B is (6000, .45).

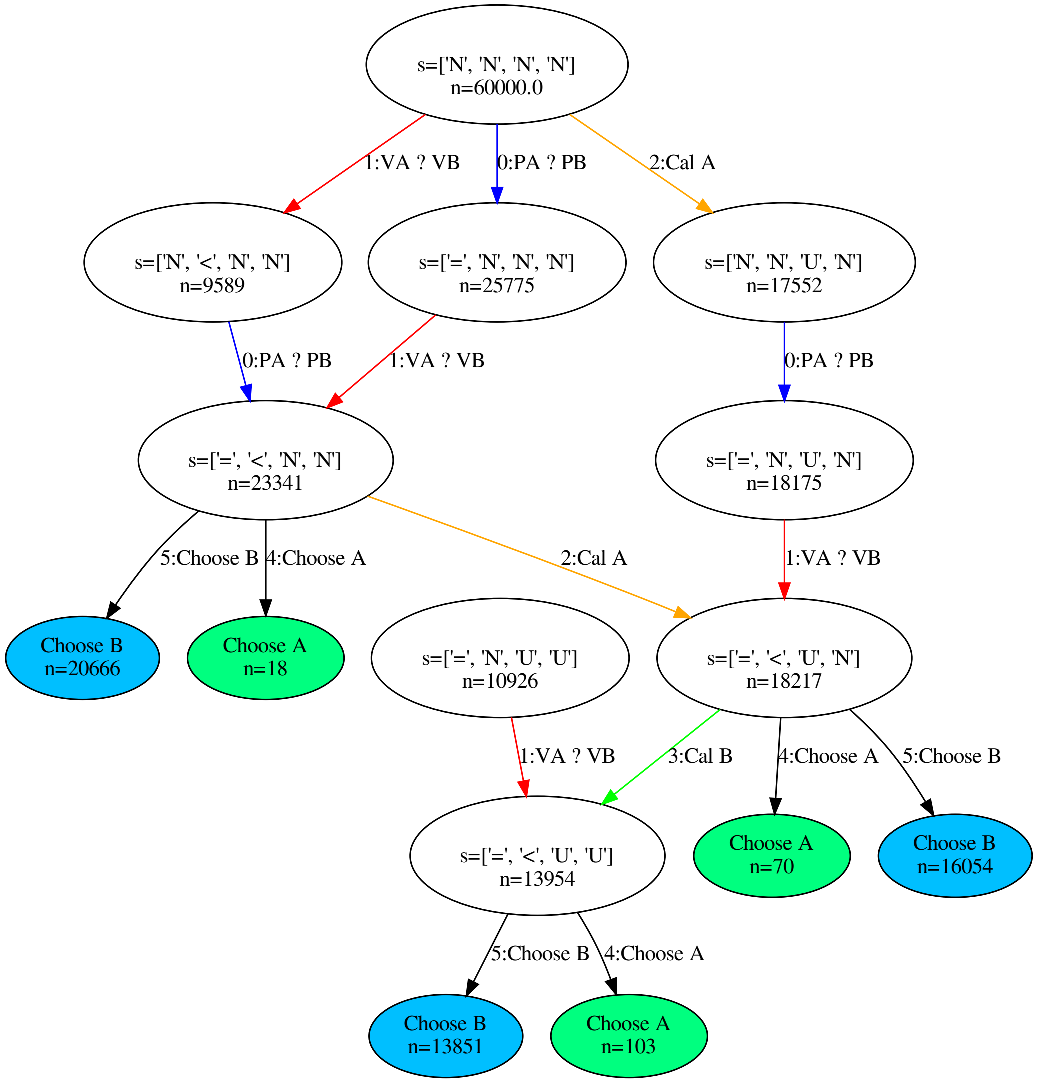


Fig. B.6 Predicted the most frequent visit states of choices by the Model Zero agent for Problem 8. Option A is (3000, .002) and option B is (6000, .001).

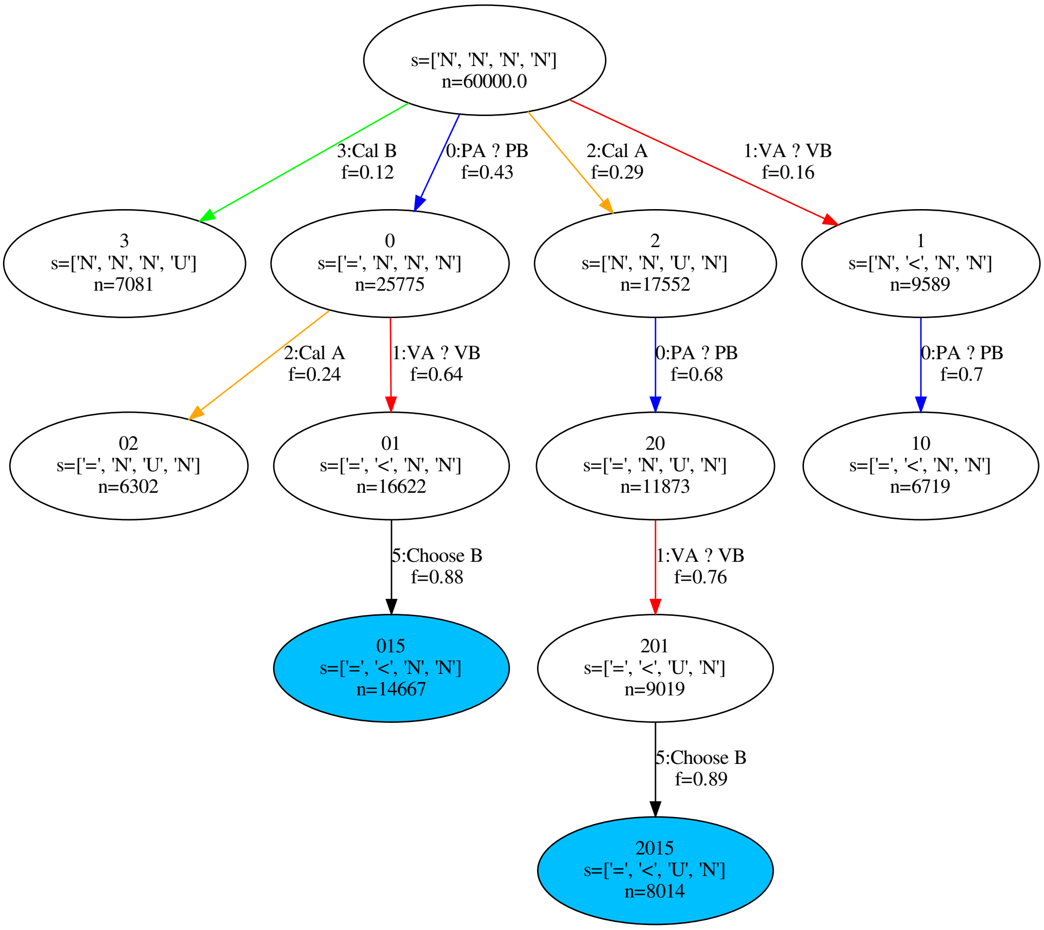


Fig. B.7 Predicted the most frequent sequence of choices by the Model Zero agent for Problem 8. Option A is (3000, .002) and option B is (6000, .001).

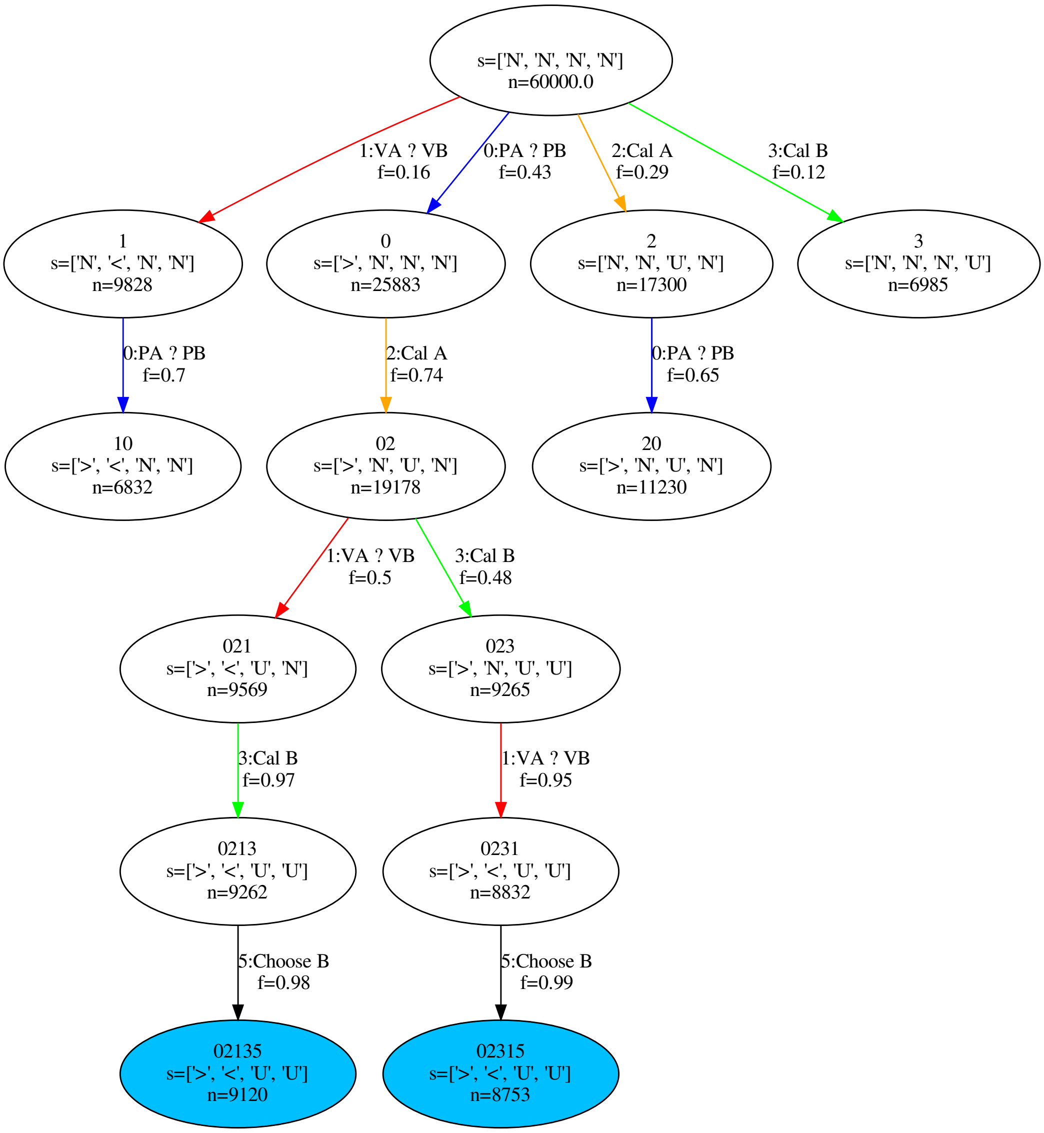


Fig. B.8 Predicted the most frequent sequence of choices by the Model Zero agent for Problem 14. Option A is (5, 1.0) and option B is (5000, .001).

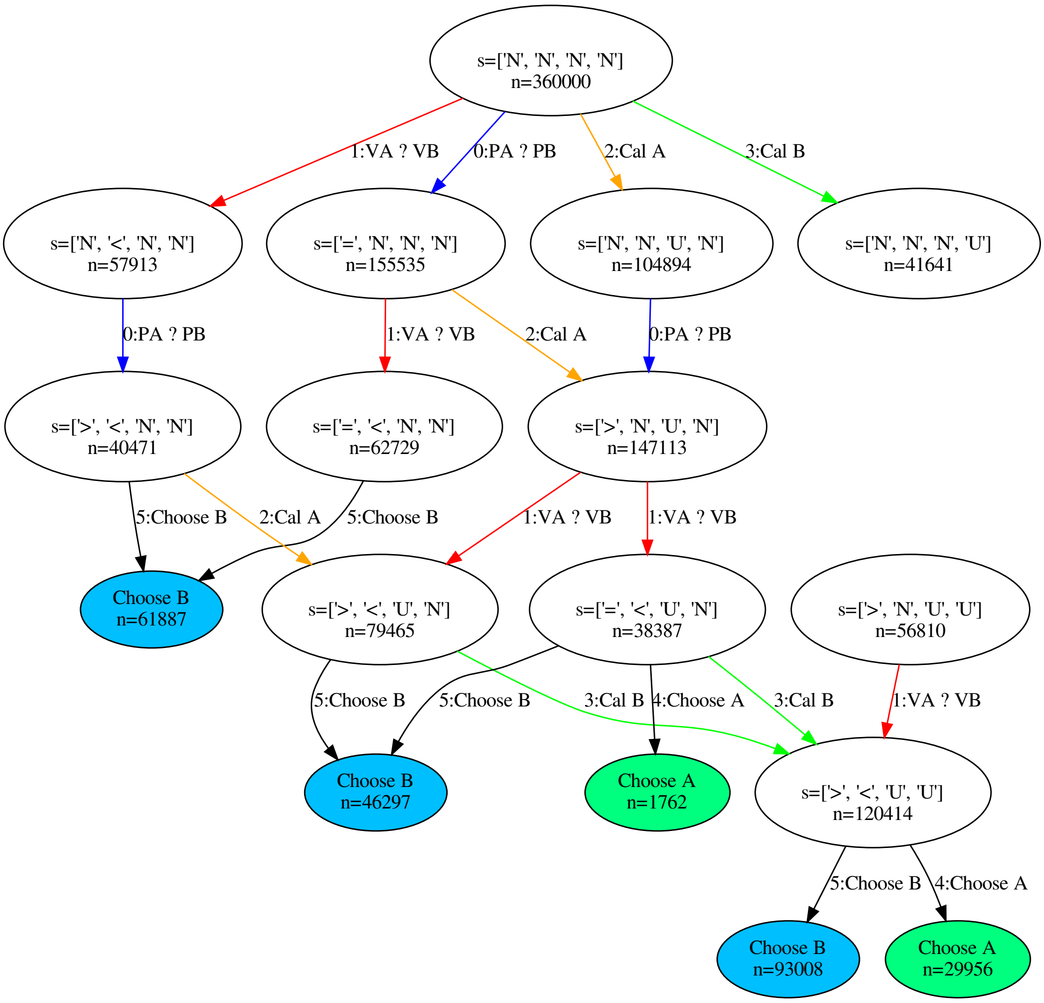


Fig. B.9 Predicted the most frequent visit states of choices by the Model Zero agent for all the problems. Option A has a higher probability and option B has a lower probability.

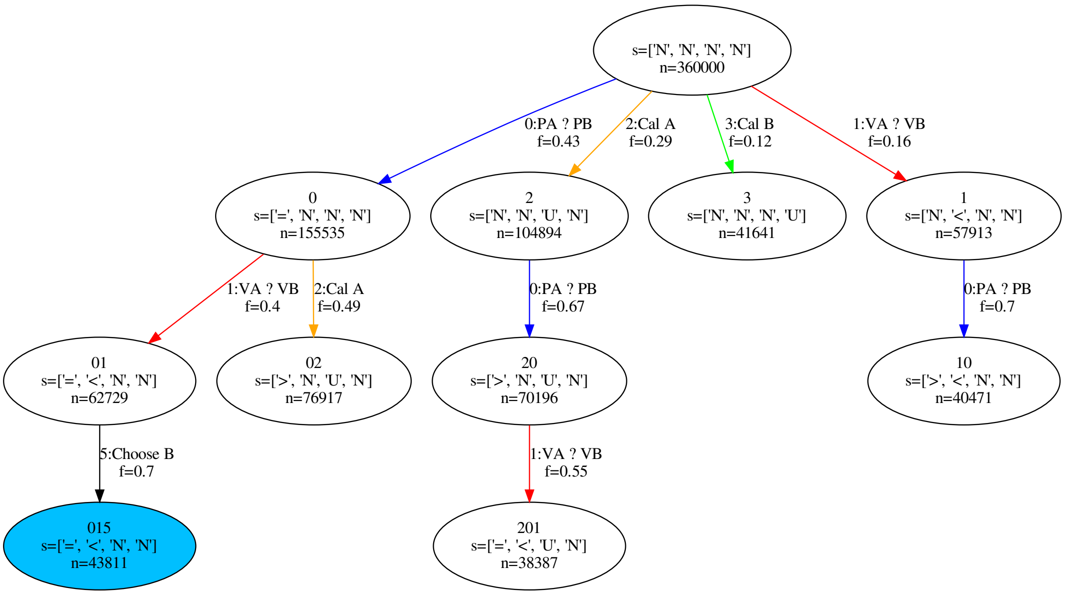


Fig. B.10 Predicted the most frequent sequence of choices by the Model Zero agent for all the problems. Option A has a higher probability and option B has a lower probability.