MODELLING LEARNING AND DECISION MAKING UNDER INFORMATION PROCESSING CONSTRAINTS

Tyler Malloy

Submitted in Partial Fullfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Approved by:
Chris R. Sims, Chair
Bram Van Heuveln
Brett Fajen

Department of Cognitive Science
Rensselaer Polytechnic Institute
Troy, New York

[December 2020]
Submitted October 2020

CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENT

Thank you to my advisor Chris R. Sims for mentorship and advice throughout the completion of my degree. To the members of my Master's Thesis committee, Brett Fajen and Bram van Heuveln. To the members of our Adaptive Cognition lab, Rachel A. Lerch and Zeming Feng, as well as the Perception and Action lab, Nathan Powell, Scott Steinmetz, and Grace Roessling. To my research partners at IBM, Tim Klinger, Miao Liu, Matt Riemer, and Gerry Tesauro as well as the entire RPI-IBM AI Horizons Network for their support and collaboration. Finally I would like to thank my family, Richard, Nanette, Meredith, and Ryan, as well as my loving partner Jamie for all the care and support they have provided me.

ABSTRACT

This thesis explores the impact of information processing constraints on models of human learning and decision making. This is achieved through altering existing methods within the fields of economic decision making and reinforcement learning, with inspiration from information theory and the rational inattention economic framework. The result is two models, one of human decision making and one of human learning, which seek to represent the way that differences in individual information processing abilities impact learning and decision making. Data from human responses in a learning task is used to compare the accuracy of this model against existing methods. Results from experimentation show that these models achieve a high degree of accuracy while accounting for the impact of differences in information processing capacity utilized by participants during the learning task. These results further the understanding of how cognitive limitations impact human learning and decision making, and suggest ways in which this type of model could potentially benefit current approaches in artificial intelligence, by incorporating more human-like learning strategies.

CHAPTER 1
INTRODUCTION

This work introduces 'capacity-limited' models of learning and decision making that incorporate limitations on information processing to explain sub-optimal performance observed in humans. There is a long history of methods that seek to explain this sub-optimally such as Heuristics (Tversky & Kahneman, 1974), Rational Inattention (Simon, 2019), and Sampling based probability approximation (McKenzie, 1994). These models and others will be detailed more fully in the following chapter, taking care to describe the way that these methods relate to the capacity-limited approach. These approaches are all similar in that they model sub-optimal behaviour as the result of reducing the informational complexity of the learning and decision making tasks. As a result of this reduction in task complexity, these resulting models predict some aspects of sub-optimal behaviour that is observed in humans.

The basis of the two methods introduced in this paper are Expected Utility Theory (EUT) in the case of decision making, and Reinforcement Learning (RL) in the case of learning. Information theory will provide the tools necessary to investigate learning and decision making, and quantify the difficulty of these tasks. The result is two models related to EUT and RL which seek to represent the way that differences in individual information processing abilities impact learning and decision making.

Information theory is a broad field that has influenced wide ranging disciplines such as communication (Shannon, 1948), lossy compression (Ahmed et al., 1974) and cryptography (Menezes et al., 1996). But recently it has been applied to the field of economic decision making in the rational inattention framework (Caplin et al., 2019). The specific purpose that it serves within this thesis is in allowing the proposed models to define the complexity of a decision or a learning task based on the information that would be required to represent it. These models are described as 'capacity-limited', because they have a constraint on the amount of information used by the model to represent behaviour.

Models of economic decision making seek to represent the way that rational decision makers choose between different options that are available to them. Traditionally, these models represent differences between outcomes based on the so-called utility associated with them, as well as the probability of that outcome taking place given the action the agent performs (Resnik, 1987). The rational decision maker is thus represented using Maximum

Expected Utility (MEU) by choosing the action that maximizes the expected utility. This expected utility is calculated by weighing the utility of each outcome of an action by the probability of that outcome occurring given the action performed. Key aspects of this theory are the definitions of 'utility', 'outcome' and 'action', with a history of disagreement regarding their precise meaning (Briggs, 2019). To avoid an overly detailed description of all possible conceptualizations of these terms, the following will provide working definitions.

In this thesis, utility will refer to a quantification of any preference that a rational decision maker has over possible outcomes. This allows utility to define differences in preference over monetary rewards, time, emotional states, etc. and is a common definition used in economic decision modelling (Kahneman & Tversky, 2013). For the remaining terms, I will adopt Savage's definitions which represent actions as behaviour that the decision maker has independent control over, and outcomes as the immediate change on the environment that are relevant to the preferences of the agent (Savage, 1972).

When we move from decision making under EUT into learning, the 'utility' will be discussed with an analogous term called reward, which for the purposes of this paper can be thought of as essentially the same as utility. Reinforcement learning extends the basis formed by EUT into the domain of learning by modelling the way that agent's learn the probabilities of outcomes used by the agent to calculate the expected utility (Sutton & Barto, 2018). These outcome probabilities are learned and updated through experience in a learning environment, based on the actions taken and the outcome observed. This allows for modelling the learning and decision making of human participants in learning tasks by predicting the actions that they agents will perform in a learning task based on the RL model.

A theoretical decision making task will be used to demonstrate the differences between the novel decision making model and related methods. For the learning model, data from human participants engaging in a learning task is used for comparison to existing methods. Results from theoretical analysis and real-world data demonstrate the benefit of applying information processing constraints onto these models. Critically, they demonstrate the possibility of quantifying an individual participant's information processing capacity in a way that is useful for modelling their behaviour across two closely related learning tasks.

CHAPTER 2
LITERATURE REVIEW

2.1 Explaining Sub-Optimal Behaviour

The motivation for using capacity-limits in modelling human learning and decision making is to account for the impact of individual agents' information processing capacities on learning and decision making. This is motivated by observations in studies of human learning and decision making of behaviour that would be considered sub-optimal or irrational from the perspective of utility maximization. To better understand this motivation and position the capacity-limited method within a broader family of related approaches, this chapter will detail existing methods that seek to account for sub-optimal behaviour in modelling learning and decision making. Additionally, when possible, the capacity-limited approach will be related to these methods to demonstrate where they would make similar predictions, and whether or not they are based on the same assumptions regarding human learning and decision making.

Within the field of heuristics there are many different descriptions of sub-optimal behaviour resulting from a decision maker using a heuristic to simplify the task they are presented with. One example that is highly relevant for economic decision making and reinforcement learning are 'Misconceptions of Chance', such believing that the specific coin flip record H-H-H-H-H-H is less likely than H-T-T-H-H-T because of the uniformity of the first outcome (Tversky & Kahneman, 1974). This type of misconception would have a significant impact on learning and decision making models, as they are grounded in estimating the probability of an event occurring. Sampling based methods of probability approximation were originally described to account for the mismatch between human ability in updating beliefs based on evidence and optimal Bayesian Inference (McKenzie, 1994). This is similarly relevant to RL and EUT due to the importance of assessing outcome probabilities and updating these based on evidence. The rational inattention framework has been used to explain observations in real-world economic systems such as the effect of 'price stickiness' whereby price fluctuations tend to be slower than would be theoretically optimal for the individual agents making decisions in that system (C. A. Sims, 2003). This failure to update utility estimates is relevant to the types models discussed in this paper, as it has a significant effect on the calculation of the optimal action to take. The remainder of this chapter will

more thoroughly describe these models of sub-optimality, and the assumptions that they make about human learning and decision making.

2.2 Bounded Rationality and Heuristics

Many decision making tasks are too complex to be performed optimally by rational agents with high consistency. Bounded rationality considers the 'limits of rationality' that occur in these complex tasks in which decision makers must reduce the complexity of the task they are presented with due to their inability of perfectly representing the task (Simon, 2019). When taken as a high level description of human decision making, this approach is agnostic to the precise mechanisms by which the complexity of the task is reduced. Because of this, methods which similarly reduce the complexity of the decision maker's task can be considered as compatible with this approach.

One of the most computationally difficult aspects of decision making is determining the probability that an outcome will occur if a certain action is taken. One common method for reducing the complexity of approximating outcome probabilities is the use of heuristic based approaches. These methods avoid the direct calculation of outcome probabilities by relying on heuristics such as judging an outcome as more likely if it is easily recalled, judging an outcome as more likely if it has happened recently, or not considering outcomes with very low probability when making a decision (Tversky & Kahneman, 1974). This list is far from complete, but it represents the types of simplifications that people tend to make when presented with a decision that is too complex to perfectly represent.

The motivation for a boundedly rational analysis of human decision making is both observations of human's sub-optimal behaviour as well as a simple observation that many real world environments have too many relevant features to consider. In developing the theory of bounded rationality Herbert Simon noted that "optimization is an ideal that can be realized only in extremely simple worlds... and worlds having strong and simple mathematical structures" which do not reflect the reality of most real world learning (Simon, 1992). Thus, modelling behaviour must necessarily take into account 'side conditions' that deviate from optimization of performance, and are caused by a wide range of factors such as limitations in ability, attention and heuristics used by the decision maker. This thesis will later describe how the capacity-limited approach can be considered as compatible with the main goal of bounded rationality, by describing one possible method for reducing task complexity.

2.3 Sampling Based Methods

The origins of sampling based methods are in explaining differences between human performance and so-called 'Bayesian Ideal Observers'. In some tasks, humans have been shown to update their beliefs about the likelihood of an outcome in correspondence with Bayes' Theorem based on the evidence presented to them, such as the relative likelihood of an outcome occurring based on a related event occurring (McKenzie, 1994). Bayes' Theorem allows for the precise calculation of an outcome based on the base-rate or prior probability as well as the probability given another event occurring. However, these original experiments were conducted with only one or a few outcomes, whereas many decisions in the real world have dozens or hundreds of potential outcomes, raising the question of how well humans could make these calculations in these more complex decisions.

Rather than viewing humans as making decisions as if they are Bayesian Ideal Observers, sampling based methods attempt to represent human decision making as resulting from a limited number of samples from the posterior distribution (Vul et al., 2014). Thus, a decision maker is thought of as choosing actions by simulating experience based on the possible outcomes, and determining which action to take based on these samples. This results in a model of behaviour that can account for certain observations in non-optimal behaviour from humans. One example being the existence of stochasticity or randomness in human decision making when a single deterministic choice would result in higher utility.

One issue that is introduced with sampling based methods is the question of when the model should stop sampling, as in real world decision making continuing to deliberate can result in lower utility. Some models that rely on sampling account for this issue by incorporating the cost of utility in sampling another simulated experience into the objective function of the decision maker, thus allowing them to determine the optimal number of simulated experiences to consider in order to maximize expected utility (Lieder & Griffiths, 2015). Because the capacity-limited method is concerned with the complexity of behaviour, it could be possible to use it as a tool for determining when a model should stop sampling. This could be done by representing the informational complexity of the agent's behaviour, and stopping sampling once the complexity is higher than the agent can represent. In this way the capacity-limited approach could be related to the techniques described in sampling based models, though this method will not be described fully in this thesis.

Some issues with applying sampling methods to understanding human decision making

are the effects of unlikely outcomes with either very high or very low utility (Lieder et al., 2018). A common example given to explain this phenomenon is a game of 'Russian Roulette' in which a player has a 1 in 6 chance of dying (Vul et al., 2014). Although the odds of dying are not extremely high, it may require many simulated experiences of playing the game to properly model the high risk of playing the game. This means that an agent who relies on simulated experience to determine the expected utility of an action may under represent the risk in this and similar cases. Sampling based methods can attempt to account for these and other issues by altering the sampling strategy away from a given outcome distribution towards more consideration on the outcome of extreme events (Lieder et al., 2018).

2.4 Attention and Information

A well-studied phenomenon in the field of economic decision making is that of 'limited attention' under which a decision maker uses a subset of all available information when making a decision. An extensive overview provided in (Lleras et al., 2017) details work that connects limited attention to several heuristics and behaviors of human decision makers. These heuristics typically consist of taking a difficult decision such as choosing a car to buy, and reducing the number of features to consider based on some preference over the features themselves. Examples given include 'Top N' whereby a decision maker only considers the top N options according to a limited number of qualities of their choices, such as only choosing from a set of cars that are your favourite color or preferred number of cup-holders (Rubinstein & Salant, 2006). 'Top on each' describes a decision maker as considering only the optimal options for each of the features when considered alone (Gourville & Soman, 2005) such as choosing between two cars, one with your preferred number of miles driven, and one with your preferred number of doors. 'Rationalization' is described by a decision maker subjectively selecting one or a few features to consider, and rationalizing their preference by selecting the option that optimizes only these features, such as choosing the single car that has your favourite color and number of doors, and not taking into consideration any other car (Cherepanov et al., 2013).

These phenomena have been further connected to the conceptualization of decision makers as Shannon information channels which attempt to communicate information with a limited capacity for processing information (Caplin et al., 2019). This is done by considering the expected decrease in utility associated with gathering information about the probability

of outcomes given the actions under consideration. This is different from the traditional economic decision making paradigm where all relevant information is available to the decision maker. Instead, this agent must incur a penalty on utility by spending resources to gain information that is relevant to determine which action to take. An intuitive example of this can be thought of as spending time and monetary resources to study the stock market before choosing a stock.

This is closely related to the motivation of the capacity-limited method applied to expected utility theory. However, there are clear differences between the rational inattention domain and our interests in modelling human learning and decision making. In this economic rational-inattention domain the 'cost of information' does not represent cognitive costs related to determining the optimal action as is the case with the capacity-limited approach. Instead, this cost of information is related to the concept within economics of treating information as a resource that can be gathered through expending utility, and attempting to maximize expected utility subtracted by any utility cost incurred by gathering information (Mackowiak et al., 2018).

2.5 Working Memory in RL

The capacity-limited method applied to the domain of RL is one approach among many that attempt to account for the sub-optimal performance of human participants in learning tasks. One such method is Collins' RL-Working Memory (RLWM) model which incorporates the impact of limitations on working memory in RL, and has been shown to better predict phenomenon observed in human learning compared to traditional RL methods (Collins & Frank, 2012) (Collins et al., 2014). Results from experimentation demonstrated that the RLWM model is able to better predict differences in learning tasks caused by a set-size effect in a 'n-armed bandit task' in which participants determine which action is associated with the highest expected reward by taking actions and observing rewards (Collins, 2018). This effect is observed when comparing performance of a single participant in learning tasks with different requirements on working memory. The presence of a set-size effect in the RLWM model differs from traditional RL methods which would not perform notably worse in a task with a slight increase in the number of options. In these experiments, this difference was caused by changing the number of options that were available to a participant in the n-armed bandit task. The RLWM model posits that the decrease in performance when increasing

the number of options available is partially due to the increased requirements on working memory during learning.

The Capacity-Limited approach when applied onto the RL setting (CLRL) is closely related to the RLWM method in that both attempt to account for supposed sub-optimal performance of human participants in learning tasks. However they achieve this goal using different methods and are based on different assumptions. One difference is that RLWM is a 'mixture method' which functions by simultaneously training two models of learning, one based on RL and one on working memory, and predicting performance based on a mixture of the two predictions. The CLRL approach differs from this by altering the mechanism of the RL model at the level of the learning objective used by the model. For this reason, CLRL can be used to model behaviour on any learning task that may or may not be well represented by a working memory model. Additionally, because CLRL is based on modelling the informational complexity of a learning task, it is compatible with the observed decrease in performance due to the set size effect. As we will see demonstrated in the chapter on capacity-limited methods, increasing the number of options available increases the information processing requirements of a learning task in a way that is captured by the CL approach.

CHAPTER 3
CAPACITY-LIMITED METHODS

3.1 Introduction

The main contribution of this work is in describing two 'capacity-limited' methods, one for modelling human learning and one for decision making. These methods alter traditional approaches by limiting the amount of information that is used to represent the behaviour of decision makers. This is done by representing the difficulty of a learning or decision making task using information theory in terms of mutual information or KL-Divergence. With this representation of task difficulty, the capacity-limited method applies an artificial bound onto the model, to represent the same information processing constraints that would exist naturally in biological agents. Additionally, the capacity-limited approach builds off of similar methods that seek to explain the sub-optimal performance of humans in these types of tasks, such as bounded rationality and rational inattention. The result is a model of learning and decision making that incorporates the information processing limitations that can vary between decision makers. Ideally, this measure of information processing capability can be determined from human participant's behaviour and used across different tasks. This is demonstrated in the chapter on capacity-limited RL by fitting an information processing capacity to individual human participants' behavioural results and then using a capacity-limited model to predict their choices across two different learning tasks.

The first method described in this paper is a Capacity-Limited form of Expected Utility Theory (CL-EUT), using a simple example of choosing between different actions with different reward structures and expected utilities. This simple decision making task is used to describe how we derive the measure of the complexity of a task and apply the capacity to the information processing capacity of the model. Additionally, this example is further used to describe the expected behaviour of agents with more or less information processing abilities. This expected behaviour shares many common features with similar approaches in explaining the sub-optimal behaviour of real agents. As the information processing ability of the model varies, the predicted behaviour can have more or less expected utility, ideally in a similar manner as would be expected of a human performing a task that requires more information processing resources than they have available.

The second method discussed in this paper is an alternative to the traditional rein-

forcement learning model. By applying the same general method of limiting the amount of information used by our model as in CL-EUT, we can develop a new learning objective that applies the same information-theoretic limitation. This capacity-limited method allows for representing differences in information processing capacity between the different participants in the learning task. Because this learning objective is based on a regularization of the traditional RL objective, it can be substituted in any existing RL method. Previously, this approach has been applied onto the domain of Artificial Intelligence to improve the generalizability of learned behaviour in these artificial agents (Lerch & Sims, 2019 July 7-10). In this thesis this general approach is extended to applications in models that predict human learning and decision making. In the chapter on Capacity-Limited Reinforcement Learning, this learning objective will be applied onto an existing RL model.

The following chapters will describe how the capacity-limited approach is applied onto existing methods, and the features of these models that are different from existing approaches. After each method, a short discussion is given on the broader implications of these capacity-limited approaches, including their relation to topics outside of the scope of modelling human learning and decision making, such as human-inspired artificial intelligence. Some of the aspects of human learning that we are interested in modelling are difficult to represent in current AI techniques, and therefore could benefit from the capacity-limited method.

3.2 Information Theory

This section will briefly describe some terminology from the field of Information Theory, which is incorporated into the capacity-limited. In all decision making agents, there is a natural capacity for storing and processing information and Information Theory give us the tools to analyze the difficulty of a learning task. Once this difficulty is quantified, we will be able to model agents as acting in accordance with their individual abilities. As mentioned, there are many use cases for information theory, each with different assumptions which could impact how the information requirements of a task are quantified (Adriaans, 2020). For the purpose of modelling human learning and decision making using utility maximization and reinforcement learning, the most useful conceptualization of the quantities described in the following sections is in terms of random variables with equations given in terms of probability density functions. This is done to conform to the conceptualization of information based on Shannon Information that defines information in terms of the negative log of the

probability: $I(A) = -\log P(A)$ (Shannon, 2001). The result is an information-theoretic measure of complexity that will be be applied onto the capacity-limited versions of EUT and RL.

3.2.1 Entropy

Within the field of information theory, entropy is the fundamental measure of information that is applied onto probability distributions, and can be represented in terms of bits of information. Entropy can be thought of as a measure of the difficulty, in an information theoretic sense, of representing a probability distribution by quantifying the amount of information that is required to represent it. Intuitively, if the entropy of a probability distribution is 0 then the random variable always takes on the same value. Conversely, the entropy of a random variable is maximized by applying an equal probability to all possible outcomes. We calculate entropy by:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)) \tag{3.1}$$

Outcomes of actions in learning and decision making tasks are represented by probability distributions and are used to determine which action has the highest expected utility. In this way probability distributions are a fundamental aspect of economic decision modelling and RL. The goal of the capacity-limited approach is to investigate how the probabilities that are learned or observed by humans are represented differently based on information processing limitations.

3.2.2 KL-Divergence

KL-Divergence is an information-theoretic measure of the difference between two probability distributions that can be represented in terms of bits of information. An important quality of this measure is that it is not symmetric, meaning that the KL-Divergence from P to Q is not necessarily the same as from Q to P. The KL-Divergence from probability Q to P defined over the set X is as follows:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \frac{P(x)}{Q(x)} \tag{3.2}$$

This quantity can be related to decision making tasks by describing the difference between two action distributions that describe an agent's behaviour. For instance if one agent selects from two choices with the action distribution [1,0] then they will deterministically select the first option. When we compare this distribution to another agent such as one that uses the action distribution [0,1], we can see that the value of the KL-Divergence will be maximized for this action distribution. On the other hand, if the second agent has the same action distribution as the first, then the KL-divergence between the two will be minimized. Although a KL-Divergence of zero does not imply in all cases that two distributions are identical, it can be assumed for the purpose of economic decision making here. This quantity will be used in the Capacity-Limited decision making model that will be described later, to compare the information cost of different decisions.

3.2.3 Mutual Information

Mutual information represents the amount of information that is gained about a random variable after observing the value of another random variable. For instance, considering two random variables X and Y, if knowledge about the value of X completely determines the value of Y, then the mutual information between them is maximized. On the other hand, if knowing the value of X gives no additional information about what the value of Y, then their mutual information is zero. Mutual information can be calculated for these random variables in terms of probability mass functions as follows:

$$I(X;Y) = \sum_{y\ inY} \sum_{x\ inX} p_{(X,Y)}(x,y) \log \left(\frac{p_{(X,Y)(x,y)}}{p_X(X) p_Y(y)} \right) \tag{3.3}$$

In the context of information processing analysis this measure can be used to describe the fidelity of a channel that processes information. If the mutual information is zero then it means that the input source provides no information about the output source and they are random relative to each other. On the other extreme, if knowledge of the input gives you complete knowledge of the output then mutual information would be maximized. This quantity will also be related to Reinforcement Learning through the capacity-limited method. Mutual information will represent the information requirements of an agent's behaviour in terms of the information conveyed about an agent's beliefs by their actions.

3.2.4 Rate-Distortion Theory

Rate-distortion theory describes the performance of an information channel that processes an input and produces an output but must do so while utilizing a limited information capacity for performing this operation. This is highly relevant for our goal of considering the learning agent to be constrained by their capacity for processing information. The result of applying rate-distortion theory onto the goal of minimizing the amount of distortion in an information channel that has a limited capacity for processing information is the Blahut-Arimoto algorithm (Blahut, 1972).

$$p_{t+1}(\hat{x}|x) = \frac{p_t(\hat{x}|x))exp(-Bd(\hat{x},x))}{\sum_x p_t(\hat{x}|x))exp(-Bd(\hat{x},x))} \tag{3.4a}$$

$$p_{t+2}(\hat{x}|x) = \frac{p_{t+1}(\hat{x}|x))}{\sum_x p_{t+1}(\hat{x}|x))} \tag{3.4b}$$

Where $d(\hat{x}, x)$ is the distortion value between the input signal $\hat{x}$ and the output x, which is to be minimized in order to minimize the amount of distortion or error in the channel. This algorithm is an iterative method which is applied repeatedly until convergence and results in an informationally compressed function that takes in the input $\hat{x}$ and gives the output x. With this tool we will define an alternative meaning of 'distortion' and apply the above algorithm onto the problem of maximizing performance under some limited capacity for processing information. This will form the basis of the 'capacity-limited' methods which are introduced in this paper. Further analysis of this method used in the RL setting is given in (Lerch & Sims, 2019 July 7-10) which applies it onto a different learning domain. This paper will apply rate-distortion theory in RL to predict the behaviour of human participants in a learning task that will be further discussed later.

CHAPTER 4
CAPACITY-LIMITED EXPECTED UTILITY

4.1 Maximum Expected Utility

Maximum Expected Utility (MEU) theory defines rational decision making as choosing actions that maximize the expected utility associated with those decisions (Resnik, 1987). Using an abstract conceptualization of utility is useful in that it allows us to describe decisions made in any domain, from playing a game of chess to driving a car or choosing a stock portfolio. In each decision making environment, utility is defined based on the preferences that that agent should have to preform well in that environment. More concretely, MEU can be formulated as defining the rational decision for an agent to choose as x from the set all available actions X that maximizes the expected reward:

$$\mathbb{E}[u(x)] = \sum_{i=1}^{n} p_i u(x_i) \tag{4.1}$$

Where x_i represents an outcome with utility $u(x_i)$ that has probability p_i of occurring, with n possible outcomes of the action x. Thus, we can think of expected utility as the sum of all possible outcome utilities, weighted by the probability of that outcome occurring dependent on performing the action x.

Although it is adequate in explaining some ideal features of decision making, the actions of real-world agents rarely conform to such a stringent maximization on utility. As a result there are several cases where human and animal decision makers choose 'sub-optimal' actions that have less expected utility than other available options (Tversky & Kahneman, 1974). As previously described, there are many cases of decision making where agents perform sub-optimally due to many different factors that could potentially impact their decisions. To give a concrete example we consider an agent that is choosing between rolling two fair dice with the following reward structure:

$$\begin{aligned} &\text{die one: } \{\tfrac{1}{6}, 0; \tfrac{1}{6}, 2; \tfrac{1}{6}, 0; \tfrac{1}{6}, 4; \tfrac{1}{6}, 0; \tfrac{1}{6}, 6; \} \\ &\text{die two: } \{\tfrac{1}{6}, 1; \tfrac{1}{6}, 0; \tfrac{1}{6}, 3; \tfrac{1}{6}, 0; \tfrac{1}{6}, 5; \tfrac{1}{6}, 0; \} \end{aligned} \tag{4.2}$$

Where for each result pair $(\frac{1}{6}, 0, \frac{1}{6}, 2\ ...)$ the first value represents the probability of that outcome occurring, which in this case is $1/6$ for all values, and the second value represents the utility outcome of that value being rolled.

From equation 4.1, we can see that the expected utility of the first die is slightly higher (2) than the second die (1.5), meaning that Expected Utility Theory would model a rational agent as always selecting the first die to roll. This action selection would be deterministic, meaning that if the agent was repeatedly presented with the same problem, it would always make the same decision with an action probability distribution defined by $A = [1, 0]$. However, this model ignores the potential cognitive limitations of our rational agent. The next sections describe how the capacity-limited approach can provide an account for explaining sub-optimal behaviour in this task that is grounded in information processing limitations.

4.2 Sampling Estimated EUT

As mentioned in the section on sub-optimal performance in learning and decision making tasks, there is evidence from experimentation that humans often make sub-optimal decisions. An example of this is the Allais' paradox in which human participants can be more likely to select an option that has a lower expected utility than an alternative when it has a lower chance of receiving very low utility (Machina, 1987). The way this is typically observed is by agents preferring a smaller guaranteed utility over a small chance at a larger utility. Oftentimes human participants select the guaranteed outcome over a chance at a larger outcome in a phenomenon referred to as risk-aversion that is observable in both real monetary rewards as well as participants' judgement of what they would do if they were given the opportunity (Myagkov, Plott, et al., 1997). One way to account for these observed phenomena would be to represent a human decision maker as taking a limited sampling from the possible outcomes. If an agent were to do this, they may have an altered conception of the true expected utilities of the different outcomes due to the limited number of samples taken. Specifically, taking the Allais' paradox as an example, a limited sampling of the outcomes of both the risky and safe bets might suggest that the safe bet is better to take. This estimated expected utility would in turn lead to sub-optimal performance on this task. In the following sections, the specific mechanism of this sampling based expected utility approximation method will be more fully detailed.

4.3 Capacity Limited EUT

The basis for capacity-limited methods is an altered version of a sampling based expected utility approximation method. This alteration attempts to incorporate differences is cognitive ability in representing the way humans learn. The way this is done in relation to expected utility in sampling methods is by using the KL-Divergence $D_{KL}(P||Q)$ as a metric of the cognitive requirements of policies learned through sampling outcomes in decision making problems. The exact method of applying KL-Divergence in this way is done will be detailed explicitly in the following section. To compare the performance of models with different information processing constraints the simple die roll betting decision described in section 4.1 will be used to describe how this capacity alters the behaviour the model predicts.

4.3.1 Capacity-Limited Sampling

The goal of the capacity-limited approach is to quantify the difficulty of a task in information-theoretic terms, and then limit the amount of information that is used to represent this behaviour. The way that this is done in the Capacity-Limited Expected Utility Sampling (CL-EUS) method is by representing the difficulty of a decision maker's behaviour as the KL-Divergence between a sub-optimal stochastic decision and an optimal deterministic decision. The sub-optimal stochastic decision is the result of the traditional sampling based method which defines behaviour as a soft-max of the expected utility approximations based on the sample of possible outcomes. The result is a model that predicts an agent as selecting an action a with probability:

$$p_s(a) = \frac{\exp(E'[a])}{\sum_{a_i \in A} \exp(E'[a_i])} \tag{4.3}$$

Where $E'_s[a]$ represents the estimated expected utility of the action a based on the sample.

An alternative to taking this soft-max would be to deterministically select the option with the highest expected utility based on the sampled outcomes. However because we are taking a limited sampling, this deterministic policy isn't guaranteed to be optimal. Additionally, the difference in terms of information between the stochastic and deterministic policies as measured by KL-Divergence may be large. This motivates the CL-EUS approach which takes the stochastic policy as a prior probability, and shifts this prior towards the deterministic policy as far as possible until the KL-Divergence between the two is too high

for an agent with a given information capacity.

4.3.2 Capacity-Limited Sampling Algorithm

The CL-EUT Sampling algorithm consists of two internal action distributions, a stochastic action distribution and a deterministic action distribution, that are used to define the behavioural action distribution that represents the behaviour of the agent that is being modelled. The stochastic action distribution $p_s(a)$ is defined based on an running estimate of expected utility of the different actions the agent can select $E[u(a)]$. This utility estimate defines the stochastic action distribution as selecting an action proportionally to its expected utility, with higher utilities selected with higher probability. The second internal policy is a deterministic probability distribution $p_d(a)$ that deterministically selects the action with the highest expected utility $E[u(a)]$. The final action distribution is the output of the model, the behavioural action distribution $p_b(a)$, which is used to represent the behaviour of an agent that has the information processing capacity C.

Algorithm 1: Capacity-Limited Expected Utility Sampling (CL-EUS)

Initialize: Information capacity C; Outcomes O; Actions A
Initialize: Behavioural Policy Learning Rate α
Initialize: Expected Utility Estimate $E[a] = \emptyset$
Initialize: Deterministic Action $p_d(a) = U\{a, b\}$
Initialize: Stochastic Action $p_s(a) = U\{a, b\}$
Initialize: Behavioural Action $p_b(a) = U\{a, b\}$
while ***not converged*** **do**
  $u(a) \leftarrow$ simulated outcomes
  for *each action a in A* **do**
    $E[a] \leftarrow E[a] + ((u(a) - E[a]/n + 1)$
    $p_d(a) \leftarrow \max E'[a]$
    $p_s(a) \leftarrow \exp(E[a]) / \sum_{a_i \in A} \exp(E[a_i])$
    while $D_{KL}(p_b(a)||p_d(a)) < C$ **do**
      $p_b(a) \leftarrow p_b(a) - \alpha(p_d(a) - p_b(a))$

After drawing the random sample and updating the estimate of the expected utilities, the deterministic policy is updated to always select the action with the highest estimated expected utility with the update $p_d(a) \leftarrow \max E'[a]$. The stochastic policy is updated to a soft-max of the current estimated expected utility with the update $p_s(a) \leftarrow \exp(E[a]) / \sum_{a_i \in A} \exp(E[a_i])$. The behavioural policy is updated in with the learning rate α to iteratively shift from the

stochastic policy towards the deterministic policy in a small increment based on the difference between the two distributions at that action $(p_d(a) - p_b(a))$. This is done until the divergence between the behavioural policy and the deterministic policy $D_{KL}(p_b(a)||p_d(a))$ is less than the capacity C. This algorithm views the discrete policy as too informationally different from the stochastic policy, and alters the stochastic policy to be as similar to the deterministic under the given capacity constraint.

When modelling a decision maker's behaviour, these three policies are all stored so that they can be efficiently updated. However this does not necessitate that a real agent would store all of these policies independently, rather that the resulting behavioural policy of the model is representative of the information-constrained behaviour of the agent. This views the deterministic action probability $p_d(a)$ to be optimal based on the simulated experience, but too complex to represent with complete fidelity, necessitating the less informationally complex probability distribution $p_b(a)$. This algorithm should be viewed as an 'as-if' method, with only the behavioural policy actually being represented in the mind of the decision maker, and the algorithm representing the way that the behaviour of the agent is updated through deliberation over a decision making task. In the following sections we will see examples of decision making tasks and the behaviour that the CL-EU sampling method would represent.

4.3.3 Modelling CL-EUS

In modelling CL-EUS it is possible to use any decision making task, such as the die rolling task mentioned earlier in the chapter, and choose an information processing capacity that is lower than what is required by the task. This allows us to compare the performance of CL-EUS models that use different information processing capacities. The following results compare the expected utility of the behavioural policy as the number of simulated experience iterations increases. Normally these simulations would be cut off once the behavioral policy stabilizes, as it represents the best possible performance of that agent on the task, but here it is drawn out to demonstrate that it does indeed converge.

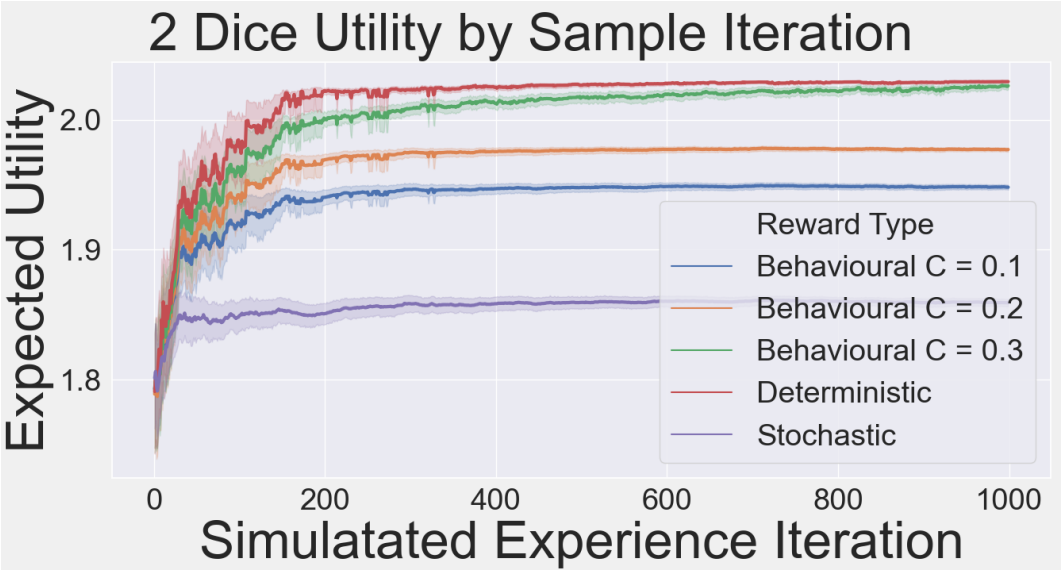

Figure 4.1: Expected utility of each policy by the step in simulated experience. Purple line represents the internal stochastic action distribution while red represents the internal deterministic action selection. Green, orange, and blue lines represent the behavioural probability distribution of an agent with 0.3, 0.2, and 0.1 bits of information capacity respectively.

These results demonstrate several important features of modelling decision making of real world agents. Firstly, they give an account for the source of sub-optimality in decision making that is grounded in information processing capability. The highest capacity model at the end of sampling has the highest expected utility, with utility decreasing as capacity decreases. Additionally the higher capacity model has a faster increase, demonstrating the lower deliberation requirements of agents that can process more information. Finally the model gives a natural definition for when sampling should end, as it continues until the behavioural policy no longer improves.

4.3.4 Set-Size Effects in CL-EUS

The set-size effect demonstrated in Collins' RLWM model (Collins, 2018) showed that humans can underperform in these tasks dependent on the number of options available to them. In modelling decision making, we can similarly increase the number of options that are available to an agent. If an agent uses the same capacity to represent both decision making tasks, we would expect the CL-EUS model to predict worse performance. This can

be demonstrated by increasing the number of sides for each die as follows:

$$\begin{aligned}
\text{die one: } &\{\frac{1}{6},0;\frac{1}{6},2;\frac{1}{6},0;\frac{1}{6},4;\frac{1}{6},0;\frac{1}{6},6;\} \\
\text{die two: } &\{\frac{1}{6},1;\frac{1}{6},0;\frac{1}{6},3;\frac{1}{6},0;\frac{1}{6},5;\frac{1}{6},0;\} \\
\text{die three: } &\{\frac{1}{6},1;\frac{1}{6},2;\frac{1}{6},3;\frac{1}{6},0;\frac{1}{6},0;\frac{1}{6},0;\} \\
\text{die four: } &\{\frac{1}{6},0;\frac{1}{6},0;\frac{1}{6},0;\frac{1}{6},4;\frac{1}{6},5;\frac{1}{6},5;\}
\end{aligned} \tag{4.4}$$

With this new decision task, the optimal choice is the 4th die with a expected utility of 2.5. The decrease in performance from adding more options to this task is noted by considering the agent using a capacity of 0.3 bits of information. This agent had a high enough information capacity to represent the deterministic action distribution, but in the case of 4 dice they cannot. Here the capacity required to represent the optimal deterministic probability distribution is approximately 0.5 bits. This corresponds with what we would expect from the previous results, as the optimal probability distribution requires roughly twice as much information to represent, and the decision has twice as many options.

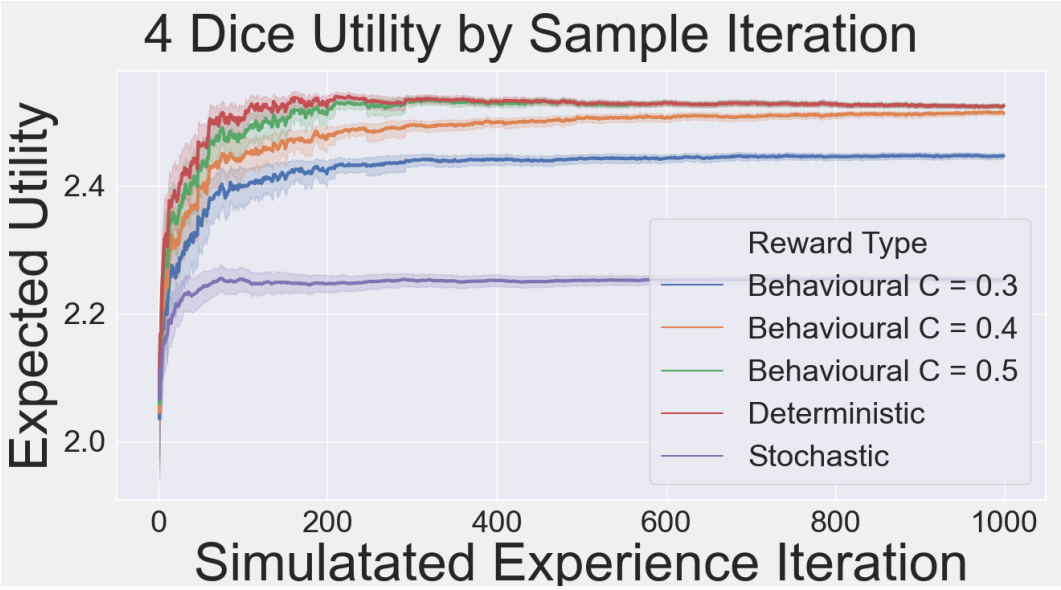

Figure 4.2: Expected utility of each policy by the step in simulated experience in the 4 die task. Purple and red lines represent the same internal distributions as earlier. Green, orange, and blue lines represent the behavioural probability distribution of an agent with 0.3, 0.4, and 0.5 bits of information capacity respectively.

The reason for the stark contrast between the performance of an agent with an information capacity of 0.3 of 2 and 4 dice is the fact that the simplicity of this decision making task means that slight differences can have a large impact on the amount of information required to represent the task. In practice, it would be likely that a human would have a large enough information processing capacity to represent the optimal performance for both of these tasks. However these results demonstrate the decrease in performance that would be expected if an agent did have too low an information capacity.

4.3.5 Discussion

The task of choosing between two or four different dice to roll with slightly different reward payoffs may not be very complex, but even with this example we can begin to understand how agents acting under limitations for processing information may act differently than ones with perfect knowledge representations. Although artificial systems have processing power that dwarfs humans, any physical system is necessarily limited in its processing of information. In this way, investigating the behaviour of lower information capacity agents such as humans can inform us on how we define optimal behaviour, and how we train artificial intelligence agents.

Additionally, there are some features of the capacity-limited decision maker that make it better suited for the types of decisions that people make in the real world. Humans and other agents acting in the real world rarely make decisions in a vacuum, and previous experience can influence our perception of novel environments. The capacity-limited method gives an account for sub-optimal stochastic behaviour as arising from a limited capacity for processing information.

However in some real-world cases it may be advantageous to adopt a stochastic policy. For instance if the reward structure changes or the agent experiences a brand new set of dice with an unknown reward structure, deterministically selecting one option may not be a good strategy. For the same reason, training an artificial agent that acts in the real world may benefit from the capacity-limited approach. This example of generalization is simple due to the simple nature of the decision making task, but we can see that in more complex tasks with more features relevant for defining good behaviour, this approach would have some desirable features.

CHAPTER 5
CAPACITY-LIMITED REINFORCEMENT LEARNING

5.1 Reinforcement Learning

We are interested in modelling not only the decisions of rational agents, but also the mechanism by which they learn the outcomes and probabilities that are required to act as optimally as possible. Many reinforcement learning algorithms exist with subtle differences depending on the specific task that they seek to preform, but the ultimate goal of any reinforcement learning agent is to learn an optimal policy $\pi^*(a|s)$ [2] which produces a probability of performing an action a dependent upon the environment state s (Sutton & Barto, 2018). This optimal policy is typically defined as that which maximizes the expected utility, referred to as reward [3] in reinforcement learning, observed by the agent. This gives the reward maximization learning objective over the time horizon $\{0, 1, ..., T\}$:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\pi}[r(s_t, a_t) \tag{5.1}$$

In this way RL is related to MEU theory, with the added interest in maximizing reward throughout all experience in the environment, where actions that are performed by the agent have an impact on the environment. Additionally, since this RL environment has unknown outcome probabilities for the actions the agent can perform, these must be approximated and updated through the learning process. For a more complete description of the RL environment see (Sutton & Barto, 2018). The state of the environment is thought of as all of the relevant parts of the environment that are necessary to determine which action should be taken. For

Portions of this chapter previously appeared as
Malloy, T. J., & Sims, C. R. (2020 July 25-31). Modelling human information processing limitations in learning tasks with reinforcement learning [Conference presentation]. Virtual Conference. https://www.youtube.com/watch?v=MCRYCRTEyfA
Malloy, T., Sims, C. R., Klinger, T., Liu, M., Riemer, M., & Tesauro, G. (2020). Deep rl with information constrained policies: Generalization in continuous control.arXiv preprintarXiv:2010.04646. https://arxiv.org/abs/2010.04646

[2] The asterisk in the policy function $\pi^*(a|s)$ represents the optimal mapping from the state of the agent onto the action they should preform to maximize expected long-term reward.

[3] Although there are some slight differences in the use of the terms reward in RL and utility in economic decision modelling, for our purposes they can be considered as essentially interchangeable.

instance, in a game of chess the state would be represented as the position of all of the pieces on the board, while a car driving agent represents its state by the visual scene around the car it is driving. The agent performs some action A, moving a chess piece or turning the car wheel, which impacts the environment. After the action is performed, the agent observes a change in the state and a reward signal from the environment. Like in MEU, the goal of the RL agent is to maximize the reward that is observed from this environment.

Reinforcement Learning (RL) has seen significant success in modelling the way that humans learn and make decisions (Niv et al., 2015, Collins and Frank, 2012), as well as advancements in artificial intelligence by controlling agents that act in both simulated and real world environments (Haarnoja et al., 2018, Mnih et al., 2016). RL is based on the same reward maximization concepts as in EUT, without the requirement that the agent already knows the probability and reward associated with each outcome from their actions. Instead, these values are learned through the reinforcement learning algorithms that form the basis of this learning model. In RL, the learning objective is refereed to as reward maximization, which performs essentially the same role as utility in EUT.

As in EUT, we are interested in determining an optimal action to take based on a state, and to that end the RL agent learns a policy function $\pi(a|s)$ that defines the action to take a dependent on the current state s. All RL agents will learn a policy function, and the way that they go about this can vary from method to method, as RL defines a broad family of approaches. Within this broad family, the most common methods of improving the policy utilize two additional functions, the state value function $v(s)$ and the state-action value function $q(s, a)$. These functions represent the learned behaviour of the agent in approximating the expected utility of a state $v(s)$ and performing an action in a state $q(s, a)$, and they are related to each other through the following equations (Sutton & Barto, 2018):

$$q_*(s, a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1})|S = s, A = a] \tag{5.2a}$$

$$v_*(s) = \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1})|S = s, A = a] \tag{5.2b}$$

Q-learning is one simple approach of reinforcement learning which uses the Bellman equations to update the q-function based on the reward observed by the agent's experience (Sutton & Barto, 2018). After learning the expected reward of each action in each state of the environment, the agent can maximize the expected reward by selecting the action a

which is associated with the highest expected reward in the state that the agent is currently in. Q-learning results in a set of values called a Q-table, which consists of one value for each state-action pair, with these values being updated throughout experience as follows:

$$Q(s,a) = Q(s,a) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \tag{5.3}$$

The inner term $[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$ is referred to as Temporal-Difference (TD) error, and is the basis of neuro-physiological connections between reinforcement learning and the learning that takes place in the human brain (Niv et al., 2005). Specifically, studies of human dopamine response levels during learning tasks show that the amount of dopamine released mirrors the magnitude of the TD-error that would be observed by an RL agent (Glimcher, 2011). This connection serves as part of the justification for using RL methods in modelling human learning and decision making (Niv, 2009). However, clear differences between RL models and human learning exist, such as the speed at which humans learn and the level at which they are able to generalize past experience into new learning tasks, resulting in newer models of RL such as the one described in the next section that try to tackle these issues (Niv et al., 2015).

5.2 Rational Inattention Model

In behavioral economics, 'rational inattention' (C. A. Sims, 2010) has been proposed as a theory of human decision-making that maximizes expected utility subject to a cost of acquiring information. This theory hypothesizes that decision-makers act so as to optimize a trade-off between the utility of their behavior, and the information processing effort required to reach a good decision. Shannon information has been proposed as a means of quantifying this information processing cost. According to theories of rational inattention, human decision-makers seek to maximize the following objective (Jung et al., 2019):

$$\max \mathrm{E}[U(X,Y)] - \lambda I(X,Y), \tag{5.4}$$

Where where $U(X,Y)$ describes the utility of choice Y in state X, and $\lambda I(X,Y)$ is a measure of the utility cost that is associated with gathering the information that is required to determine that Y is the action to perform in state X. This objective can be expanded as

follows:

$$\max \sum_{\omega \in \Omega} \mu(\omega) \left(\sum_{a \in A} P(a|\omega) u(a, \omega) \right) \tag{5.5a}$$

$$- \lambda \left[\sum_{\omega \in \Omega} \mu(\omega) \left(\sum_{a \in A} P(a|\omega) \ln P(a|\omega) \right) - \sum_{a \in A} P(a) \ln P(a) \right] \tag{5.5b}$$

The first term 5.5a is the traditional reward maximization method as described in the formula for EUT 5.4. The second term 5.5b represents the 'cost of information' which is equal to the mutual information between outcomes and actions, multiplied by a regularizer λ that captures the marginal cost of information. As we will see in the definition of the objective for the capacity-limited approach, these objectives share a close mathematical connection.

This objective incorporates these information gathering costs instead of assuming that the decision maker has the information required to make an optimal decision available, or can attain it at no cost of expected utility. The parameter λ represents the 'cost of information' and balances the scale of the expected utility with the mutual information between the choice Y and state X. This approach is mathematically very similar to the capacity-limited, and resembles the learning objective described later in Eq. 5.9. The main difference between the two is that the capacity-limited method limits information in terms of an internal constraint on cognitive information processing as opposed to an external constraint on the utility cost incurred by gathering information.

The result of incorporating this cost of information in maximizing utility is that decisions made by such a model do not necessarily result in the highest expected utility when not accounting for the costs associated with acquiring information. This difference is proportional to the scale of the parameter λ; as λ increases, behaviour that requires less information to gather will be preferred. In the extreme, a decision-maker would act randomly or else choose the same action regardless of his or her state, because such behaviour requires no information to define.

In this work we seek to extend the understanding of the learning objective described in Eq. 5.4 onto the behaviour of an agent that has a limit on its capacity for processing information, rather than an agent with unlimited capacity and a utility penalty associated with gathering information. This will first done both in terms of economic decision making as was initially described in the field of rational inattention. However, the decision making

domain does not take into account learning dynamics, as it assumes all available knowledge relevant to the agent is available instead of being learned through experience. For this reason we additionally extend this approach into reinforcement learning, which seeks to model such learning dynamics.

5.3 Feature Reinforcement Learning

Now that we have a basic understanding of RL, we can investigate one application onto the domain of modelling human learning and decision making. In (Niv et al., 2015), Niv et al. describe a method of predicting human learning and decision making in a multidimensional environment using a method called Feature Reinforcement Learning (FRL). This task consisted of an altered form of the classic 2-armed bandit learning challenge (Robbins, 1952), augmented by adding a context clue as to which option is optimal, as well as 3 total options. See Niv et al., 2015 for a more complete description of the task stimulus and diagrams. The goal of the task was to determine which of the 9 potential features (3 colors, shapes and textures) is more likely to be associated with a higher probability of observing a reward. During each set of ~ 20 stimuli, one of the nine features was randomly selected to be associated with a 75% chance of reward when selecting the option with that feature in it. The two remaining options that do not contain the feature of interest were associated with a 25% chance of reward.

In traditional reinforcement learning, each state of this learning task would be represented by the specific combination of option features and their location, and an artificial agent would learn which action is optimal for any given state. This would result in learning a Q-table of at least 256 states and 3 actions, for a total of 768 values to learn for an optimal policy. However, this method lacks the generalization present in human learning, and would be a poor model of how humans would learn and make decisions in this environment. Specifically, we know that humans would be able to leverage their experience in any state of the environment in determining which action is optimal, whereas a traditional reinforcement learning agent needs to perform several actions in every possible state before beginning to learn which actions have the highest expected reward.

In order to better account for the way that humans generalize based on their experience in this learning task, Niv et al. developed Feature Reinforcement Learning (FRL) as a variant of Q-learning (Niv et al., 2015). Firstly, instead of learning values associated with states of the environment, it learns 9 values associated with each of the 9 features that make up the 3

options. The algorithm defines the value of an option $V(S)$ in the contextual n-armed bandit learning task to be the sum of these values of the features that make up that option:

$$V(S) = \sum_{f \in S} W(f), \tag{5.6}$$

where the weights of each feature are updated based on the selection that was made by the participant and the reward that was observed as follows:

$$W^{\text{new}}(f) = W^{\text{old}}(f) + \eta[R_t - V(S_{\text{chosen}})] \; \forall f \in S_{\text{chosen}}. \tag{5.7}$$

Additionally, a decay parameter decreases the values of each feature that was not selected after each feature weight update:

$$W^{\text{new}}(f) = (1 - \delta)W^{\text{old}}(f) \; \forall f \notin S_{\text{chosen}} \tag{5.8}$$

This FRL-decay model was the best performing method among all of the methods that were tested in the original paper that described Feature Reinforcement Learning (Niv et al., 2015). Other methods tested included a Bayesian model, traditional reinforcement learning, a hybrid model, as well as a 'serial hypothesis' that assumed participants iterate through each of the 9 feature before continually selecting whichever was associated with a reward.

Although this FRL-decay model takes into account several features of human learning that are important for modelling this task, such generalization, and maximizing expected utility, it does not attempt to explicitly account for differences in the cognitive abilities for different participants. Specifically, we would like to develop a model of human learning that can account for the effect of differences in the capacity for processing information on learning and making decisions. In chapter 3, this desired feature will be achieved by incorporating the capacity-limited approach into the FRL model.

5.4 Capacity-Limited FRL

5.4.1 Capacity-Limited Learning Objective

Insight from an information theoretic perspective inspires the goal of maximizing reward obtained subject to some constraints on policy complexity (as measured by its mutual

information). One potential method of limiting the amount of information that the agent uses to represent its policy is done by applying a penalty to the reward based on this value. This allows us to define a learning objective that regularizes the observed reward:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\pi}[r(s_t, a_t) - \beta \mathcal{I}(\pi(\cdot|s_t)]. \tag{5.9}$$

The key difference with the standard RL objective is the added penalty to the reward observed based on the amount of information that would be required to represent the policy. Policies with higher mutual information values have a greater complexity, in an information-theoretic sense, and this weighted value is used to discourage policies that would require a high information capacity channel. Thus, this learning objective will directly encourage the development of policies that are simple (use low information to represent) but have high utility. Additionally, if there are multiple policies that achieve the same performance, this objective will naturally favor the simplest among them. Higher values of β skew the learning objective to prefer policies with less required information, much in the same way that the 'cost of information' parameter λ impacts the rational inattention model.

The limitation that is imposed on the information capacity of the agent's policy is introduced by this learning objective. Because this learning objective is used to update the agent's policy throughout training, the information capacity of the learned policy will be dependent on the value of β. In the extreme, very high values of β will train an agent to prefer a policy that requires as little information to represent as possible over any improvement in the reward. Because of the nature of information capacities, this policy could be either uniform in all states of the environment and perform actions randomly everywhere, or deterministic in all states or perform the same action everywhere. Conversely, setting the value of β to zero results in the traditional learning objective of maximizing the reward with no limitation on the amount of information utilized by the policy.

This learning objective is useful for demonstrating how the capacity-limited approach could be applied onto any method of reinforcement learning. This is important since it is not always computationally efficient to compute the mutual information of the policy if the state space is large or continuous. However due to the simplicity of calculating the mutual information of the policy in the FRL method, the next section will use a hard constraint on the mutual information, and update the policy until that constraint is met.

Algorithm 2: Capacity-Limited FRL

Initialize: Feature weights $W(f) = \bar{0}$;
Initialize: Hyper-parameters: α, β, η, δ, C ;
for *each participant selection* S **do**
  Predict choice with probability distribution $\pi(A|S)$
  for *each feature* f *in selection* S **do**
    $W^{\text{new}}(f) = W^{\text{old}}(f) + \eta[R_t - V(S_{\text{chosen}})]$
  for *each feature* f *not in selection* S **do**
    $W^{\text{new}}(f) = (1-\delta)W^{\text{old}}(f)\ \forall f \notin S_{\text{chosen}}$
  while $I(\pi(a|s)) > C$ **do**
    for *each* f *in* $W(f)$ **do**
      $f = f - \alpha(f - \sum_{f \in F} W(f)/|W(f)|)$

5.4.2 CL-FRL Algorithm

Applying the learning objective defined in (5.4) onto the domain of reinforcement learning results in a novel [4] algorithm that allows us to define a capacity for the amount of information that is used to represent our agent's policy. Here, instead of impacting the reward that is actually observed by our agent, we achieve the same difference in learned behaviour by setting a hard constraint on the amount of information used to represent behaviour. In this way the CL-FRL model is more similar to the original capacity-limited reinforcement learning objective described in equation 5.9. The two additional hyper-parameters are the capacity-limit C, which is determined for each participant individually using the same method as described in (Niv et al., 2015), as well as the feature weight adjustment learning rate $\alpha = 1e-3$ for all participants.

The constraint on the amount of information used to represent performance is determined by the magnitude of the capacity parameter C, which performs the same function as the parameter λ in Eq (5.4). However, instead of being a regularizer used during optimization, this method uses a hard constraint on the amount of information used to represent the policy. Decreasing the value of C results in a more and more strict limitation on the amount of information that is used by the model to represent the performance of the participant. The algorithm iteratively updates the RL Q-table to decrease the mutual information until it is below the bound. In the next section, we fit this parameter to each of the individual

[4] This algorithm was described previously in (T. J. Malloy and Sims, 2020 July 25-31, T. Malloy et al., 2020)

participants' performance in the contextual n-armed bandit learning environment. This algorithm demonstrates that the mutual information regularized expected utility maximization approach that is described in Eq (5.4) is applicable into the domain of reinforcement learning.

5.4.3 Modelling CL-FRL

The original experiment design described in (Niv et al., 2015) includes 2 different speed trials, fast (500ms) and slow (1.5s) response times, with the slow response times used during trials to allow for a fMRI scanner enough time to capture data for a separate analysis that is not discussed further. Hyper-parameters were originally fit by minimizing negative log posterior individually for both the slow and fast trials. However, one potential benefit of the capacity-limited approach is that the information capacity parameter C could be the same across different tasks for the same participant, as long as factors such as motivation and attention remain consistent enough across the different tasks. To support this, we instead fit both models [5] to the entire data set for individual participants, and compare the performance of the FRL and CLRL methods. These results indicate that it is possible to determine the information capacity that is used by a participant in a learning task, even across tasks with slightly different cognitive requirements such as the different time constraints shown here.

The high predictive accuracy of the CLRL model when fit to the entire data set demonstrates a similarity of participant's information processing capacities across different tasks. Although the individual sources of these capacities can be varied, from attention and motivation to differences in cognitive abilities, this model determines the amount of information required to represent participants' learned behaviour. This difference represents one possible explanation for less than optimal performance on learning and decision making tasks that is observed with human participants. By connecting the information-constrained maximum utility with reinforcement learning, this algorithm expands the application into learning tasks. In developing this algorithm, these results further support the conceptualization of rational decision makers as Shannon information channels with a limited capacity for storing and processing information that is efficiently allocated to maximize reward when learning and making decisions.

[5] Using the Python Scipy minimization package (Virtanen et al., 2020) with the same negative log-loss hyper-parameter optimization method described in the original FRL paper (Niv et al., 2015). This was done using a Leave-One-Out method where model parameters were fit based on all decisions besides one game, trained on that game, and then repeated for all games.

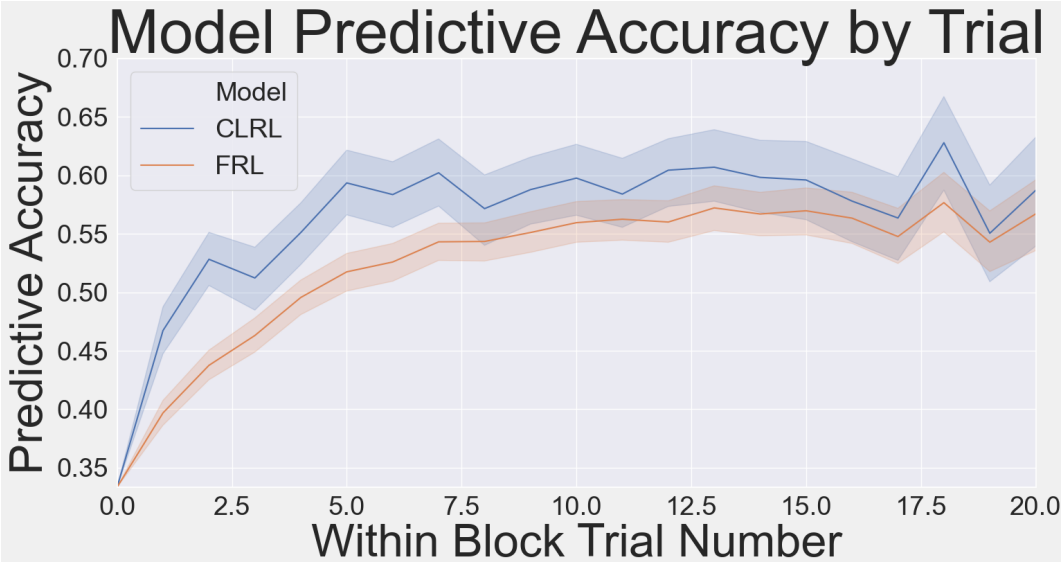

Figure 5.1: Mean predictive accuracy of CLRL and FRL models based on parameters fit to minimize negative log loss across both fast (500ms) and slow (1.5s) response times. Error bars represent 99% confidence intervals. Predictive accuracy is the value assigned by the model to select the option that was ultimately selected by the human participant.

Figure 5.2 details the predictive accuracy of each type of model (FRL and CLRL) additionally broken down into models trained on shared parameters across the different trial types (fast and slow), and unique parameters. For the shared parameter models, a single set of model parameters are found using a optimization of the negative-log loss of the human decisions, selecting the parameters which maximize the probability that is assigned to the decisions the human participants made. The unique parameters models optimize their parameters independently on either of the two learning conditions, fast and slow responses, and use those parameters for the model that predicts performance only on the responses corresponding to their optimization.

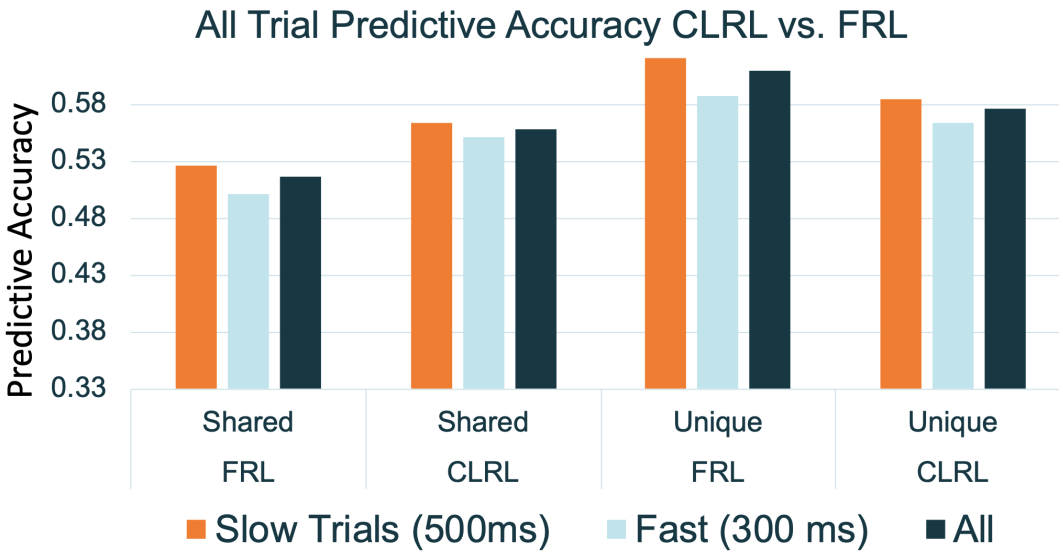

Figure 5.2: Mean predictive accuracy of CLRL and FRL models. Results are reported for models trained and with parameters fit across both fast and slow tasks (left 2 columns), as well as models that are individually trained and with parameters fit on exclusively fast and slow trials (right 2 columns).

From these results we can see that the CLRL model has little difference between model performance when sharing parameters across both trial conditions (55% vs. 57%). Meanwhile, the FRL model has significantly decreased predictive accuracy 51% compared with the model trained on two different sets of parameters 61%. This result is to be expected, since the CLRL model attempts to capture a feature of human learning that should be relatively static across different tasks. This differs from the FRL approach which simply seeks to maximize predictive accuracy. Although the FRL model with unique parameters has a marginally improved predictive accuracy compared to the best performing CLRL model, it lacks the additional benefit of representing the information processing capacity of the human participants.

5.4.4 Discussion

For this learning task, the capacity-limited version of the FRL model provided a modest improvement when fitting model hyper-parameters across the two different task instances. This suggests that the capacity-limited model is a good candidate for use is other areas of reinforcement learning, such as training artificial agents in tasks that require human-like generalization that is displayed in this learning task. However, this improvement in model performance is not the only key finding of the success of this model. Another important aspect

is that the model learned a single parameter that represented each participant's capacity for storing and processing information and used it across two different tasks, with a high degree of accuracy in predicting performance. This later finding has implications more closely related to the goal of modelling how humans learn and make decisions, and particularly for how differences in cognitive abilities can impact performance.

There have been many significant advancements in the field of reinforcement learning from board games (Tesauro, 1995) to video games (Mnih et al., 2015) (Silver et al., 2016), to controlling robots designed to walk and drive cars (Lillicrap et al., 2015). However, a significant remaining issue is in developing systems that learn in a similar manner to humans in their ability to learn quickly and flexibly through generalization (Packer et al., 2018) (Cobbe et al., 2019), which has been an interest of cognitive science since its foundation (Shepard, 1987) as well as in recent years (C. R. Sims, 2018). We see this work as furthering the collaborative effort of both cognitive scientists and artificial intelligence researchers to connect the learning mechanisms present in human learning in solving issues in AI research. In particular, the results presented in this final chapter represent a clearer understanding of how humans' natural capacity for storing and processing information impacts the way they learn in a difficult task that requires a high degree of generalization over a short period of time.

CHAPTER 6
CONCLUSIONS

6.1 Limitations

Both sections on Capacity-Limited EUS and RL discussed the implications that can be drawn from the experiments discussed. However these experiments cannot give a complete picture of all aspects of these models. This section will introduce some limitations associated with the experiments and results drawn from them in this thesis. In the following section, possible future directions will be discussed that may be able to address these limitations. This will include suggestions of potential additional studies that could be conducted to account for the limitations that were present in this work.

One shared limitation that exists between both the CL-RL and CL-EUS models is that the sources of limitations in information processing are difficult to determine from observing behaviour alone. Many factors could all impact the resulting information processing capacity of an agent such as cognitive ability, motivation, attention, or familiarity with a task, among others. However there may be additional experiments that could be performed with human participants to account for these issues.

The die rolling task allowed for a simple and clear example of how capacity-limited sampling can impact the behaviour of expected utility models. However, this task is simple enough that the majority of human participants would likely perform it optimally without much deliberation. This means that it would be unlikely to expect any difference in behaviour among human participants, even if they had a difference in their information processing capacity. An experiment method that could potentially address this issue by increasing the informational difficulty of the task will be discussed in the next section.

Results from modelling human learning in the multi-armed contextual bandit task did allow for a slight difference between tasks as a result of the different time requirements (fast and slow trials). Although we might expect a participant to behave in relation to a single information processing capacity between these tasks, it may be the case that the different time requirement changed this capacity slightly. A clearer method of testing this might be to use environments that have the same time constraint but different amounts of information to process. This ideally would show the same or a similar capacity across more diverse tasks.

6.2 Future Directions

Modelling human learning and decision making in more disparate and complex tasks may help to better account for the impact of information processing capacities across tasks compared to the relatively similar tasks used in this thesis. This could be done to either determine a single shared information processing capacity across all tasks, or a baseline capacity that can be slightly varied depending on the motivation and attention of the agent. Doing these types of experiments across many different tasks with a standardized learning environment and monetary reward might help to reduce the impact of other factors beyond just differences in cognitive abilities.

A simple extension of the die rolling task using more dice with more sides, and a more complex reward system may allow for a decision task where human participants would perform sub-optimally. Additionally because the CL-EUS model relies on sampling which makes some claims on the differences in deliberation, a time limit difference in this task could be used to analyze how well the model compares to limited-deliberation util decisions.

To account for the lack of diversity between the two types of trials in the CL-RL modelling section, it may be possible to preform a similar analysis on more diverse environments. This could most easily be done in relation to the experiment described earlier by simply altering the number of options and features in the learning environment. Alternatively, we could change aspects of the learning task like the features that are used to represent options, or the probabilities associated with the different options in the game. These types of experiments could reveal interesting aspects of the capacity-limited RL model in relation to participants capacity for processing information.

The models described in this paper and the results from theoretical and real-world experimentation demonstrate a wealth of future research in the area of modelling learning and decision making in humans. As noted throughout the paper, there also exists a connection between models that more accurately reflect human learning, and advances in artificial intelligence. As models of learning and decision making become more human like, they would hopefully be more able to perform well in the types of tasks that humans excel at related to artificial agents.

BIBLIOGRAPHY

Adriaans, P. (2020). Information, In *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, *100*(1), 90–93.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, *18*(4), 460–473.

Briggs, R. A. (2019). Normative Theories of Rational Choice: Expected Utility, In *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Caplin, A., Dean, M., & Leahy, J. (2019). Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, *86*(3), 1061–1094.

Cherepanov, V., Feddersen, T., & Sandroni, A. (2013). Rationalization. *Theoretical Economics*, *8*(3), 775–800.

Cobbe, K., Klimov, O., Hesse, C., Kim, T., & Schulman, J. (2019). Quantifying generalization in reinforcement learning, In *International conference on machine learning*. PMLR.

Collins, A. G. (2018). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432.

Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*(41), 13747–13756.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15647–15654.

Gourville, J. T., & Soman, D. (2005). Overchoice and assortment type: When and why variety backfires. *Marketing Science*, *24*(3), 382–395.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Jung, J., Kim, J. H., Matějka, F., & Sims, C. A. (2019). Discrete actions in information-constrained decision problems. *The Review of Economic Studies*, *86*(6), 2643–2667.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk, In *Handbook of the fundamentals of financial decision making: Part i*. World Scientific.

Lerch, R. A., & Sims, C. R. (2019 July 7-10). *Rate-distortion theory and computationally rational reinforcement learning* [Conference presentation]. Montreal, Canada.

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem., In *Cognitive science*.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lleras, J. S., Masatlioglu, Y., Nakajima, D., & Ozbay, E. Y. (2017). When more is less: Limited consideration. *Journal of Economic Theory*, *170*, 70–85.

Machina, M. J. (1987). Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, *1*(1), 121–154.

Mackowiak, B., Matejka, F., & Wiederholt, M. (2018). Rational inattention: A disciplined behavioral model. *CEPR Discussion Papers*, *13243*.

Malloy, T. J., & Sims, C. R. (2020 July 25-31). *Modelling human information processing limitations in learning tasks with reinforcement learning* [Conference presentation]. Virtual Conference. https://www.youtube.com/watch?v=MCRYCRTEyfA

Malloy, T., Sims, C. R., Klinger, T., Liu, M., Riemer, M., & Tesauro, G. (2020). Deep rl with information constrained policies: Generalization in continuous control. *arXiv preprint arXiv:2010.04646*. https://arxiv.org/abs/2010.04646

McKenzie, C. R. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and bayesian inference. *Cognitive Psychology*, *26*(3), 209–239.

Menezes, A. J., Katz, J., Van Oorschot, P. C., & Vanstone, S. A. (1996). *Handbook of applied cryptography*. CRC press.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning, In *International conference on machine learning*.

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), 529–533.

Myagkov, M., Plott, C. R. Et al. (1997). Exchange economies and loss exposure: Experiments exploring prospect theory and competitive equilibria in market environments. *American Economic Review*, *87*(5), 801–828.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neurosci*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and td learning. *Behavioral and Brain Functions*, *1*(1), 6.

Packer, C., Gao, K., Kos, J., Krähenbühl, P., Koltun, V., & Song, D. (2018). Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*.

Resnik, M. D. (1987). *Choices: An introduction to decision theory*. U of Minnesota Press.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Rubinstein, A., & Salant, Y. (2006). A model of choice from lists. *Theoretical Economics*, *1*(1), 3–17.

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science*, *3*(3), 150–161.

Simon, H. A. (2019). *The sciences of the artificial*. MIT press.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690.

Sims, C. A. (2010). Rational inattention and monetary economics, In *Handbook of monetary economics*. Elsevier.

Sutton, R. S., & Barto, A. G. (2018). *Introduction to reinforcement learning (2nd ed.)* MIT press Cambridge.

Tesauro, G. (1995). Temporal difference learning and td-gammon. *Communications of the ACM*, *38*(3), 58–68.

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. https://doi.org/https://doi.org/10.1038/s41592-019-0686-2

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.