WARWICK
THE UNIVERSITY OF WARWICK

A Thesis Submitted for the Degree of PhD at the University of Warwick

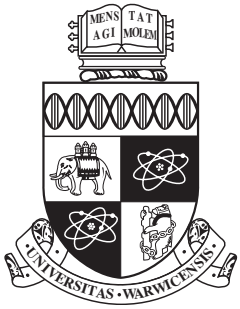
Permanent WRAP URL:
http://wrap.warwick.ac.uk/132094

Copyright and reuse:
This thesis is made available online and is protected by original copyright.
Please scroll down to view the document itself.
Please refer to the repository record for this item for information to help you to cite it.
Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

The Sampling Brain



by

Jian-Qiao Zhu

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Psychology

Department of Psychology

University of Warwick

December 2018

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am fortunate to have many people to thank.

My primary advisor, Elliot Ludvig, for the opportunity to conduct researches in psychology. My interests in computational modelling were kindled at his neuroeconomics course. Later, I spent a summer building cognitive models using reinforcement learning algorithms, and since then, he has been instrumental in guiding me and demonstrating the power of reverse-engineering the mind.

My second advisor, Adam Sanborn, for the right balance of freedom and guidance, for all the late night email chains, for his careful and critical reading of manuscripts, and for the most intensive and productive collaborations at Warwick. Adam brings excitements to all the mathematical and research problems, and his works inspired me to explore rational analysis and Bayesian models.

I would like to acknowledge a debt to Nick Chater, certainly an unofficial advisor, for his inspiration all along the way, for his helpful advises on research and career, and for showing me, along with a generation of students and readers, the power of theory.

I also owe a great deal to Samuel Gershman and his students, Ishita Dasgupta, Eric Schulz, and Rahul Bhui, for their friendly reception at Harvard and for many enlightening discussions.

My experience as an international graduate student would have been poorer but for the fellow students at Warwick and the friends around the world. For their friendships and for many selfless supports, I would like to particularly thank Andong Chen, Victoria Collard, Ahuti Das, Tianshi Feng, Alina Gutoreva, Daniel Gunnell, Anita Lenneis, Chao Li, Ling Liu, Si Li, Wenfeng Li, Danielle Norman, Max Rodriguez, Divya Sukumar, Wonnie Sim, Alexandra Surdina, Mengran Wang, Wendi Xiang, and Shengnan Zhang.

Finally, those to whom this thesis is dedicated. My parents have generously supported me every step of the way here, and their encouragements animate these pages.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below.

Inclusion of Published Works

Parts of this thesis have been published or preprinted by the author.

Chapter 2 includes the following publication:

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2018). Mental Sampling in Multimodal Representations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 5749-5762). Montréal, Canada.

JQZ, ANS, and NC designed research; JQZ and ANS derived the models and analysed the data; JQZ carried out model simulations and data analyses; JQZ and ANS wrote the manuscript in consultation with NC. ANS supervised the project.

Chapter 3 includes the following preprint:

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2018, November 28). Bayesian inference causes incoherence in human probability judgments. https://doi.org/10.31234/osf.io/af9vy.

JQZ, ANS, and NC conceived of idea and developed the theory; JQZ performed the data analysis and model simulations; All authors discussed the results and contributed to the final manuscript. ANS supervised the project.

Chapter 4 includes the following preprint:

Ludvig, E. A., Zhu, J. Q., Mirian, M. S., Kehoe, E. J., & Sutton, R. S. (under review). Associative learning from replayed experience. On *bioRxiv*, 100800.

EAL, JQZ, MSM, EJK, RSS conceived of idea and developed the theory; EAL, JQZ, and MSM performed model simulations; EAL and RSS encouraged JQZ to investigate new memory model and hippocampal lesion data; All authors discussed the results and contributed to the final manuscript. EAL supervised the project.

Chapter 5 includes the following publications:

Zhu, J. Q., Xiang, W., & Ludvig, E. A. (2017). Information seeking as chasing anticipated prediction errors. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 3658–3663). Austin, TX: Cognitive Science Society.

JQZ and EAL conceived of idea and developed the theory and experiment; WX conducted experiments and collected data; JQZ and WX analysed behavioural data; JQZ performed model simulations; All authors discussed the results and contributed to the final manuscript. EAL supervised the project.

Rodríguez-Cabrero, J. A. M., Zhu, J. Q., & Ludvig, E. A. (2019). Costly curiosity: People pay a price to resolve an uncertain gamble early. *Behavioural Processes*.

JAMR, JQZ and EAL conceived of idea and experiment; JAMR conducted experiments and collected data; All authors analysed behavioural data; JAMR wrote the manuscript in consultation with JQZ and EAL. EAL supervised the project.

Abstract

Understanding the algorithmic nature of mental processes is of vital importance to psychology, neuroscience, and artificial intelligence. In response to a rapidly changing world and computational demanding cognitive tasks, evolution may have endowed us with brains that are approximating rational solutions, such that our performance is close to optimal. This thesis suggests one instance of the approximation algorithms, sample-based approximation, to be implemented by the brain to tackle complex cognitive tasks. Knowing that certain types of sampling is used to generate mental samples, the brain could also actively correct for the uncertainty comes along with the sampling process. This correction process for samples left traces in human probability estimates, suggesting a more rational account of sample-based estimations. In addition, these mental samples can come from both observed experiences (memory) and synthesised experiences (imagination). Each source of mental samples has unique role in learning tasks and the classical error-correction principle of learning can be generalised when mental-sampling processes are considered.

Abbreviations

Abbrev.	Explanations
MCMC	Markov chain Monte Carlo
MC^3	Metropolis-coupled Markov chain Monte Carlo
IRI	Inter-response interval
DS	Direct Sampling
KL	Kullback-Leibler
AR	Auto-regressive
RF	Relative frequency
PTN	Probability Theory plus Noise
BS	Bayesian Sampling
ESS	Effective sample size
MSE	Mean squared errors
CS	Conditioned stimulus/stimuli
US	Unconditioned stimulus/stimuli
MDP	Markov decision process
RL	Reinforcement learning
IL	Initial link
TL	Terminal link
TD	Temporal difference
APE	Anticipated prediction errors

This page is intended to be blank.

Chapter 1

Why Sampling?

"The first thoughts and attempts I made to practice [the Monte Carlo method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practice method than 'abstract thinking' might not be to lay it out say one hundred times and simply observe and count the number of successful plays" (Stan Ulam, 1983)

1.1 Intuitions for Sampling

The spirit of sampling is captured by Stan Ulam's interest in estimating the probability of winning in solitaire — from the simple to the sublime. When the analytical solution is difficult to obtain, numerical approximations are often treated as desirable alternatives. Another classic example that demonstrates the power of sampling is the calculation of the value of π (the ratio of a circle's circumference to its diameter). The sample-based approximation to π takes three simple steps. First, we inscribe a circle within a square. Second, randomly scatter a number of points over the square. Third, count up the number of points bounded inside the circle. Given that the ratio of areas of circle and square is $\pi/4$, the value of π can be approximated from number of points inside the circle and the total number of points:

$$\pi \approx 4 \times \frac{\text{no. of points within circle}}{\text{total no. of points}} \tag{1.1}$$

As illustrated in the Figure 1, on average, the sample-based approximation improves in accuracy to the true value of π as more samples are generated.

Indeed, the history of scientific development resembles an everlasting sample-based approximation to truth. Generations of scientists are constantly drawing samples from nature, through experimentation, in order to perform better estimates on the probabilities of possible theoretical models. Though the number of samples or experiments is always finite, by continually sampling, we slowly build up a picture of all of the probabilities.

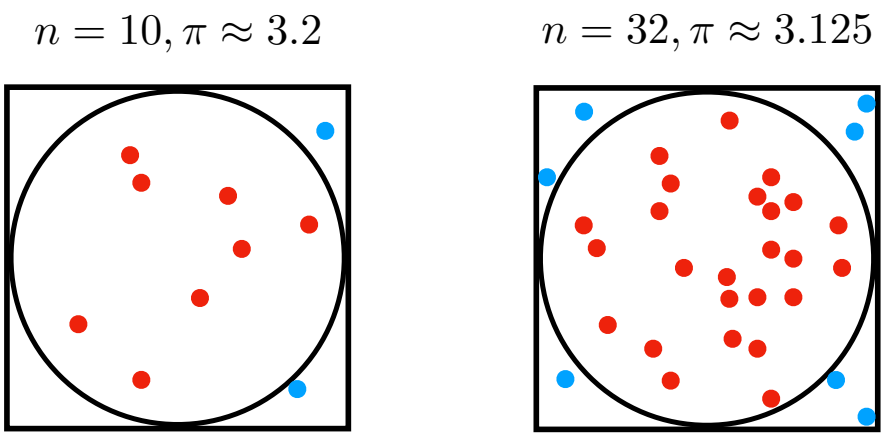


Figure 1. Approximating the value of π through sampling. Points that placed inside the circle are marked as red and those outside are blue. The accuracy of sample-based approximation increases as more samples are scattered in the square, on average.

1.2 A Rational Model of Sampling Brain

Similar challenges, at least in principle, are also imposed on the brain: the world is a highly uncertain place, and we want our brain to be able to generate good estimates of these uncertainties. In fact, there are many sources of uncertainty the brain has to deal with, including the sensory system, the motor apparatus, one's own knowledge, and the data-generation process from the world. To process noisy data efficiently to make judgments and guide choices, the brain must represent and use information

about uncertainty in its computations. One normative and ecologically rational method is for the brain to adopt a Bayesian approach because it provides an optimal way of reasoning about these uncertainties — the *Bayesian brain hypothesis* (Knill & Pouget, 2004; Doya, Ishii, Pouget, & Rao, 2007; Friston, 2012; Sanborn & Chater, 2016). Indeed, over the past few decades, Bayesian approach has spawned an enormous range of applications in cognitive science from perception (Knill & Richards, 1996; Yuille & Kersten, 2006; Gershman, Vul, & Tenenbaum, 2009; Shams & Beierholm, 2010), memory (Anderson & Milson, 1989; Gershman, 2017), intuitive physics (Sanborn, Mansinghka, & Griffiths, 2013; Battaglia, Hamrick, & Tenenbaum, 2013), and animal learning (Courville, Daw, & Touretzky, 2006; Gershman, Blei, & Niv, 2010). Moreover, a growing body of neuroscience evidence suggests a complementary explanation of Bayesian models of cognition in that the brain could encode information probabilistically with neural computations that follows Bayes rule (e.g., Knill & Pouget, 2004; Berkes, Orban, Langyel, & Fiser, 2011; Savin & Deneve, 2014).

Yet, the large literature on judgment and decision-making has emphasised irrationality and identified an array of replicable systematic biases in cognition (e.g., Peterson & Beech, 1967; Tversky & Kahneman, 1973; 1974; 1983; Gigerenzer & Gaissmaier, 2011; Hilbert, 2012). This research tradition apparently argues against normative Bayesian principles and advocates heuristic approximations of various kinds (e.g., Gigerenzer, 2001; Ariely, 2009; Marcus, 2009; McRaney, 2011), downplaying any systematic coherence in how the brain deals with uncertainty.

In addition, many everyday cognitive tasks, such as recognising a cat from a photo or identifying your mother's voice from phone calls, may look too trivial to be solved by a Bayesian approach. Consider, however, the sheer number of pixels or speech waves the brain has to process, and even worse, the vast space of alternative explanations for those input data (e.g., it could be that a dog was in the photo or your aunt was calling) that the brain has to take into account. From this perspective, most cognitive tasks are too computationally difficult to be able to be solved through exact Bayesian

inferences. Indeed, the computational problem faced by agents attempting to be rational (including Bayesian inference) is generally intractable (Aragones, Gilboa, Postlewaite, & Schmeidler, 2005; Sanborn & Chater, 2016; Bossaerts & Murawski, 2017; Lieder, Griffiths, & Hsu, 2018).

Thus, we are faced with an apparent paradox: how can Bayesian models of cognition be so useful, when (a) some basic elements of such models appear to be systematically biased and (b) there is a pervasive tractability problem across any application of rational models.

To reconcile the Bayesian models of cognition and the daunting intractability of these models in exact inference, the brain has to perform some approximation algorithms. In particular, I suggest that the brain may adopt sample-based approximation that removes the computations of representing a full probability distribution, instead approximating the distribution with a set of samples; we call the samples employed by the brain to conduct inference as *mental samples*. Just like the samples used to approximate π, mental samples are also stochastic and reasonably easy to generate, and in the limit, an infinite number of mental samples will produce the same answers as exact Bayesian inference. The approach that uses sampling to approximate Bayesian inference is known as a *Bayesian sampling* model (Sanborn & Chater, 2016).

Recent theoretical developments within the Bayesian sampling framework have identified many sampling algorithms, which have algorithmic limitations that can naturally lead to a number of systematic biases (Lieder, Griffiths, & Goodman, 2012; Sanborn & Chater, 2016; Vul, Goodman, Griffiths, & Tenenbaum, 2016; Dasgupta, Schulz, & Gershman, 2017; Zhu, Sanborn, & Chater, 2018). This suggests that the Bayesian sampling framework may resolve both the tractability issue of computations and the deviation from rationality. For example, if a local sampling algorithm is used by the brain, the resultant sample-based estimations are naturally biased toward the starting point of the algorithm, constituting an anchoring effect (Tversky & Kahneman, 1974; Lieder et al., 2012; Lieder, Griffiths, Huys, & Goodman, 2018). Many systematic biases of cognition can already be accommodated by Bayesian sampling models such as the

anchoring effect, availability bias, unpacking effect, conjunction fallacy, subadditivity, and superadditivity (e.g., Lieder et al., 2012; Sanborn & Chater, 2016; Lieder, Griffiths, & Hsu, 2018; Dasgupta, Schulz, Goodman, & Gershman, 2018). Furthermore, this Bayesian sampling approach has also been implemented in spiking neural networks — as in the *neural sampling hypothesis* (Moreno-Bote, Knill, & Pouget, 2011; Berkes et al., 2011; Orban, Berkes, Fiser, & Lengyel, 2016; Hennequin, Vogels, & Gerstner, 2014; Buesing, Bill, Nessler, & Maass, 2011; Savin, Dayan, & Lengyel, 2014; Savin & Deneve, 2014; Haefner, Berkes, & Fiser, 2016), suggesting a promising direction that bridges computational, algorithmic, and implementational levels of analysis of cognition (Marr, 1982).

Starting from the principle that the brain is approximating rational solutions (possibly with sampling), a rich web of theoretical insights can be derived. I first study the question: "*where to sample*?" (Chapter 2). Specifically, by assuming a mental representation of some cognitive task, where should the mind generate the next sample? Bayesian sampling accounts have to deal with the following two phenomena: both distances (Bousfield & Sedgewick, 1944; Rhodes & Turvey, 2007; Zhu et al., 2018) and autocorrelations (Gilden, Thornton, & Mallon, 1995; Farrell, Wagenmakers, & Ratcliff, 2006; Van Orden, Holden, & Turvey, 2005) of mental samples are scale-free. These spatiotemporal patterns of mental samples shed light on the algorithmic nature of the possible sample-generating processes employed by the brain. I will perform an evaluation for three candidate sampling algorithms: direct sampling, Markov chain Monte Carlo (MCMC), and Metropolis-coupled Markov chain Monte Carlo (MC^3). While the first two sampling algorithms have previously been proposed as mechanistic models of cognition (e.g., Vul et al., 2014; Sanborn & Chater, 2016; Dasgupta et al., 2017), they cannot reproduce either observed spatiotemporal pattern. The MC^3 algorithm, one of the first sampling algorithms that was designed to better explore multimodal representations, is able to capture these patterns. This result suggests that the brain may employ sampling algorithms that can search multimodal representations effectively.

When the brain has collected a set of mental samples, the next question would be "*how to make an estimate based on samples?*" (Chapter 3). Given the fact that the mental samples are inherently stochastic (to different degrees for different sampling algorithms), the brain should not trust these mental samples equally, and, if possible, should take into account the stochasticity. The optimal way to temper these intrinsic uncertainties of mental samples is, again, Bayesian inference (e.g., Bayesian Monte Carlo: Ghahramani & Rasmussen, 2003). As a proof-of-concept, I made simplifying assumptions such that the brain performs exact Bayesian inference on mental samples that are generated through direct sampling from the target distribution. This additional Bayesian inference on mental samples will alter the sample-based estimations. For example, in estimating probabilities of event *A*, the relative frequency of samples has no way of integrating the observed frequencies with prior knowledge about the behaviour of event *A*. From a Bayesian sampling perspective, agents should always bring in their prior assumption of how likely the event *A* was to occur. Indeed, the sample-based estimations improve accuracy when this additional Bayesian inference on mental samples is performed. By considering the possibility that the brain corrects for mental samples, I explore the consequence of this idea and gauge how well it explain human probability estimates.

1.3 Sample from Memory and Simulation

There are two fertile sources of data generating mental samples: memory (observed) and simulation (imagined). Throughout this thesis, I adopt restricted definitions such that samples from memory are observed experiences in the past and those from simulation are synthesised experiences for possible futures. Though both remembered old experiences and imagined experiences help us to make "discoveries" in the absence of new experience, they are now distinguishable concepts with distinct temporal tags. An informative arena to investigate the role of mental sampling from past and future is the animal learning literature where

animals' value estimates are constantly updated in response to streams of experiences. In Chapter 4 and Chapter 5, I endeavour to continue the theoretical journey regarding "*what is learning?*" with mental sampling.

Despite ambiguity in the definition of learning, experimental paradigms such as classical conditioning (e.g., Pavlov, 1927; Kamin, 1969; Lubow, 1973) and instrumental conditioning (e.g., Thorndike, 1911; Skinner, 1963; Mackintosh, 1983) offer a broad agreement with regard to an operational definition of learning: *a relatively permanent change of behaviour resulting from experiences* (Thorpe, 1956). It allows a formal quantitative measurement of learning as experience-induced behavioural changes. In this way, other antecedents of behavioural changes are explicitly excluded such as changes in motivational state (e.g., hunger or thirst) and developmental trajectory. While a robust empirical description of learning, the mechanisms constituting this learning (i.e., the processes underlying the observable behaviour changes with experience) remain outside the scope of this operational definition.

The contemporary view on classical and instrumental conditioning is dominated by the error-correction principle of learning, as manifested in the Rescorla-Wagner model (Rescorla & Wagner, 1972) and the temporal-difference model (TD model: Sutton & Barto, 1990). According to these models, learning occurs whenever there is a discrepancy between delivery of reinforcement and the animal's expectation (i.e., guess about the reinforcement). Over time, animals attempt to minimise these prediction errors. Both the Rescorla-Wagner and TD models operationalise the error-correction principle; however, the TD model further generalises the trial-level Rescorla-Wagner model to real-time, and animals are assumed to learn a guess from another guess (Sutton & Barto, 1990; Sutton & Barto, 2018). The prediction error for the TD model is thus defined as discrepancy between two temporally-separated guesses.

I will show that the classical error-correction learning rule can be dramatically improved by incorporating additional sampling mechanisms. Reusing mental samples drawn from remembered past conditioning trials can augment the standard model based on error-correction learning rules in

classical conditioning (Chapter 4). Typically, training data for a learning agent is limited to the real experiences and interactions with the environment. This is, however, a limited view on what can be treated as training data for learning agents; old experiences need not to be forgotten and may enhance the performance of learning if appropriately re-used. As I will show, this mechanism of reusing mental samples from memory can rectify a number of important failures of the basic Rescorla-Wagner model (see Miller, Barnet, & Grahame, 1995 for a review of failures and successes of the Rescorla-Wagner model). Indeed, even when past trial memories were reused at random, the *random replay* model generalises the Rescorla-Wagner model to spontaneous recovery, latent inhibition, retrospective revaluation effects, and the facilitatory effects of hippocampal lesion (Ludvig, Zhu, Mirian, Kehoe, & Sutton, under review).

A further application of mental sampling to animal learning will be discussed with a subset of experiments using instrumental conditioning — those showing how animals will sometimes engage in suboptimal choice when faced with cues for rewards (e.g., Wyckoff, 1952; Prokasy, 1956; Dinsmoor, 1983; Spetch, Belke, Barnet, Dunn, & Pierce, 1990; Stagner & Zentall, 2010; Bromberg-Martin & Hikosaka, 2011; Blanchard, Hayden, & Bromberg-Martin, 2015; Iigaya, Story, Kurth-Nelson, Dolan, & Dayan, 2016; Bennett, Bode, Brydevall, Warren, & Murawski, 2016; Zhu, Xiang, Ludvig, 2017; Rodriguez-Cabrero, Zhu, & Ludvig, in press). Like other instrumental learning procedures, animals can choose what to learn; the exploration and exploitation dilemma is present (Daw et al., 2006; Cohen, McClure, & Angela, 2007; Hills et al., 2015; Sutton & Barto, 2018). The suboptimal choice paradigm, however, stresses the informativeness of conditioned stimuli. Rewards are delivered stochastically to the animals, but there are sometimes cues that predict the arrival of these rewards. Often, in this type of experiment, animals have to choose between an informative option and a non-informative option. The sub-optimality comes from cases when animals favour the less-rewarding but more informative option, suggesting a strong desire for information even at the expense of primary rewards (e.g., food and water). The sub-optimal choice experiments have

long been used to study preference for information and, more broadly, curiosity (when information provided by the informative option are non-instrumental in the sense that it cannot alter the eventualities or their chances).

This information-induced sub-optimality challenges the standard TD model or any other value maximisation accounts based on primary rewards (Bromberg-Martin & Hikosaka, 2009; Beierholm & Dayan, 2010; Iigaya et al., 2016; Zhu, Xiang, & Ludvig, 2017). A number of significant revision and refinements has been made to the TD model make it more compatible with the observed sub-optimality in animals, including additional attention mechanisms (Beierholm & Dayan, 2010), an information bonus (Bromberg-Martin & Hikosaka, 2011; Bennett et al., 2016), or the addition of anticipatory utility (Iigaya et al., 2016). As we shall see later, these modifications of the standard TD model not sufficient for a comprehensive explanation of the information-induced sub-optimality. Alternatively, in Chapter 5, I suggest an anticipatory sampling mechanism that may unify five distinct empirical challenges from the sub-optimal choice literature: cue-outcome contingency, uncertainty resolution, delay to outcomes, reward magnitudes, and the impact of negative outcomes (e.g., Stagner & Zentall, 2010; Kendall, 1974; 1975; Green & Rachlin, 1977; Bromberg-Martin & Hikosaka, 2009; 2011; Iigaya et al., 2010; Blanchard et al., 2015; Zhu et al., 2017).

According to our model, to choose among options, animals are assumed to execute forward samplings of the future prospects of choosing any option (Zhu et al., 2017). The proposed anticipatory sampling mechanism requires animals to anticipate future episodes, and these samples from this imagined future should inform animals about the prospects of their actions. To elicit a choice, animals not only rely on the learnt value estimates of options, but also the difference between the current value estimates of options and the potential future prospects of choosing these options. This difference, named *anticipated prediction errors*, contains critical information on whether animals' well-being will be improved or deteriorated, according to their own beliefs. The model treats these

anticipated prediction errors as rewarding or punishing, just like primary rewards. In this way, many sub-optimal choices can be reinterpreted as a consequence of animals' sampling bias in their imaginative simulation to pursue certain future paths. These anticipated prediction errors quantify how much reward animals expect to receive along imagined paths and provide useful signals to guide decision making.

Chapter 2

Where to Sample?

2.1 Introduction

Suppose that I already have a mental representation of some cognitive task (we shall return to specific tasks below) from which I wish to draw samples in order to better guide my upcoming decisions, how should I explore my mental representation efficiently? In this chapter, we suggest a mechanistic model whose process of sampling matches the spatio-temporal properties of human mental sampling.

As noted in Chapter 1, in many complex cognitive domains, such as vision, motor control, language, categorisation or common-sense reasoning, human behaviour is consistent with the predictions of Bayesian models (e.g., Battaglia et al., 2013; Sanborn et al., 2013; Chater & Manning, 2006; Anderson, 1991; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Kemp & Tenenbaum, 2009; Wolpert, 2007; Yuille & Kersten, 2006). Bayes' theorem prescribes a simple normative method for combining prior beliefs with new information to make inferences about the world. Intuitively, the Bayesian approach gives a formal framework for finding the best action under uncertainty, by assigning each possible state of the world a possibility and using the laws of probability to calculate the best action. However, the sheer number of hypotheses that must be evaluated suggests that individuals are performing some kind of approximate inference, such as sampling (Vul et al., 2014; Sanborn & Chater, 2016). To illustrate, suppose a Bayesian brain must represent all possible probabilities and make exact calculations on them, the number of real numbers required to encode the joint probability distribution over n binary variables grows exponentially with 2^n, quickly surpassing the capacity of any physical system including the brain. Yet, the brain often must represent and process exactly such vast data spaces in daily

cognitive tasks, such as images or speech recognition, resolving an effectively infinite hypothesis spaces of possible scenes or sentences. It is clear that the exact representation of probabilities and explicit computation of law of probabilities (e.g., conditionalisation and marginalisation) for Bayesian computational models is impossible.

How, then, can a Bayesian model of cognition possibly work if it does not explicitly represent probabilities? Using sampling to approximate Bayesian models of complex problems makes many difficult computations easy: instead of integrating over vast hypothesis spaces, samples of hypotheses can be drawn from the posterior distribution. This makes sampling free from requiring knowledge of whole distribution. The computational cost of sample-based approximations only scales with the number of samples rather than with the size of hypothesis space, though using a small number of samples results in biased inference. Using a number of samples much smaller than the number of hypotheses makes the computation feasible, though it may introduce biases.

Interestingly, the biases in inference that are introduced by using a small number of samples match some of the biases observed in human behaviour. For example, probability matching (Vul et al., 2014; Wozny, Beierholm, & Shams, 2010), anchoring effects (Lieder et al., 2012), and many reasoning fallacies (Dasgupta et al., 2017; Sanborn & Chater, 2016) can all be explained in this way.

Yet, there is as of now no consensus on the exact nature of the algorithm used to sample from human mental representations. Previous work has posited that people either use direct sampling or Markov Chain Monte Carlo (MCMC) to sample from their posterior distribution over hypotheses (Vul et al., 2014; Lieder et al., 2012; Dasgupta et al., 2017; Sanborn & Chater, 2016). In this chapter, we demonstrate that these algorithms cannot explain two key empirical effects that have been found in a wide variety of cognitive tasks. In particular, these algorithms do not produce distances between samples that follow a Lévy flight distribution, and separately they do not produce autocorrelations in the samples that follow *1/f* scaling. A further issue is that mental representations have been shown to

be "patchy" or multimodal — there are high probability regions separated by large regions of low probability — and MCMC is ill suited for multimodal distributions. We therefore evaluate one of the first algorithms developed for sampling from multimodal probability distribution, Metropolis-coupled MCMC (MC^3), and demonstrate that it produces both key empirical phenomena. Previously Lévy flight distributions and *1/f* scaling have been separately explained as the result of efficient search and as a signal of self-organising behaviour respectively (Viswanathan, Buldyrev, Havlin, Da Luz, Raposo, & Stanley, 1999; Van Orden, Holden, & Turvey, 2003). Here we provide the first account to explain both phenomena as the result of the same purposeful mental activity.

2.2 Distances between mental samples: Lévy flights

In the real world, resources are rarely distributed uniformly in the environment. Food, water, and other critical resources often occur in spatially isolated patches with large gaps in between. As a result, humans' and other animals' foraging behaviours should be adapted to such patchy environments. In fact, foraging behaviour has been observed to produce Lévy flights, which is a class of random walk whose step lengths follow a heavy-tailed power-law distribution (Shlesinger, Zaslavsky, & Frisch, 1995). In the Lévy flight distribution, the probability of executing a jump of length l is given by:

$$P(l) \sim l^{-\mu} \tag{2.1}$$

where $1 < \mu \leq 3$, and the values $\mu \leq 1$ do not correspond to a normalisable probability distribution. Examples of mobility patterns following the Lévy flight distribution have been recorded in albatrosses (Viswanathan, Afanasyev, Buldyrev, Murphy, Prince, & Stanley, 1996), marine predators (Sims et al., 2008), monkeys (Ramos-Fernandez, Mateos, Miramontes, Cocho, Larralde, & Ayala-Orozco, 2004), and humans (Gonzalez, Hidalgo, & Barabasi, 2008).

Lévy flights are advantageous in patchy environments where resources are sparsely and randomly distributed because the probability of returning to a previously visited target site is smaller than in a standard random walk. In the same patchy environment, Lévy flights can visit more new target sites than a random walk does (Berkolaiko, Havlin, Larralde, & Weiss, 1996). More formally, it has been proven that, in foraging, the optimal exponent is $\mu = 2$ regardless of the dimensionality of the space if the following three criteria are satisfied: (a) the target sites are sparse, (b) they can be visited any number of times, and (c) the forager can only detect and remember nearby target sites (Viswanathan et al., 1999).

It has long been known that mental representations of concepts are also patchy (Bousfield & Sedgewick, 1944) and remarkably the distance between mental samples also follows a Lévy-flight distribution. For example, in a semantic fluency task (e.g., asking participants to name as many distinct animals as they can), the retrieved animals tend to form clusters (e.g., pets, water animals, African animals) (Troyer, Moscovitch, & Winocur, 1997). This same task has also been found to produce a Lévy-flight distribution of inter-response intervals (IRI) (Rhodes & Turvey, 2007).

2.2.1 New Experiment on Distances between Mental Samples

As we are interested in mental sampling, which can retrieve the same item multiple times, rather than destructive foraging, where an item once found is used up, we conducted a new memory retrieval experiment. Ten native English speakers (6 Female and 4 Male, and aged 19-25 years) were recruited from the SONA system of Warwick University (Coventry, UK). The task lasted about 60 minutes or until the participants typed 1024 words. Participants sat in a soundproof cubicle for this task and were paid 6 GBP for completing the experiment.

The following instructions appeared on the screen before the task began:

Hello and Welcome!

In this free association experiment, you are asked to type animal names as they come to mind. You will be shown the animal name you most recently reported on the screen and when you think of a different animal name, please type it into the computer.

We are interested in the free association of animal names, so we would like you to report what new animal you are thinking of whenever the animal you are thinking of changes.

It is okay to type in an animal name that you previously reported. Please let the experimenter know if you have any question before you begin.

Press any key when you are ready to continue.

Participants were also told to press ENTER when they finished typing an animal name. The inter-response interval (IRI) was the duration between last ENTER pressed and the next key response.

Participants showed power-law scaling of their IRI, replicating the main finding of Rhodes and Turvey (2007) (see Figure 3A). IRIs can be considered a rough measure of the distance between mental samples, assuming that generating a sample takes a fixed amount of time, that there are unreported samples generated between each reported sample, and that the sampler has traveled further the more unreported samples that are generated. As further support, we used a standard technique from computational linguistic to measure the distances between mental samples, again finding Lévy flight distributions for these distances (see Appendix section 2.7.3)

2.3 Autocorrelations of mental samples: *1/f* noise

Separate from investigations into the distances between mental samples, a number of studies have reported that many cognitive activities contain long-range, slowly decaying autocorrelations in time. These autocorrelations tend to follow a *1/f* scaling law (Kello, Brown, Ferrer-i-Cancho, Holden, Linkenkaer-Hansen, Rhodes, & Van Orden, 2010):

$$C(k) \sim k^{-\alpha} \tag{2.2}$$

where $C(k)$ is the autocorrelation function of temporal lag k. The same phenomenon is often expressed in the frequency domain:

$$S(f) \sim f^{-\alpha} \tag{2.3}$$

where f is frequency, $S(f)$ is spectral power resulting from a Fourier analysis, and $\alpha \in [0.5,1.5]$ is considered *1/f* scaling.

1/f noise is also known as pink or flicker noise, which varies in predictability intermediately between white noise (no serial correlation, $S(f) \sim 1/f^0$) and brown noise (no correlation between increments, $S(f) \sim 1/f^2$). Note that Lévy flights (i.e., randomly selecting a flight direction and then executing a flight distance that has power-law scaling as in Equation 2.1) are random walks and so produce $1/f^2$ noise instead of *1/f* noise.

1/f -like autocorrelations in human cognition were first reported in time-estimation and spatial-interval-estimation tasks in which participants were asked to repeatedly estimate a pre-determined time interval of 1 second or spatial interval of 1 inch (Gilden, Thornton, & Mallon, 1995). Subsequent studies have shown *1/f* scaling laws in the response times of mental rotation, lexical decision, serial visual search, and parallel visual search (Gilden, 1997), as well as the time to switch between different percepts when looking at a distal stimulus such as a Necker cube (Gao et al., 2006).

Table 1.

Empirical evidence for Lévy flights and *1/f* noise in human mental samples

Effect	Papers	Experiments	Main findings
Lévy flights	Rhodes & Turvey (2007)	Memory retrieval task	Power-law exponents IRI: $\mu \in [1.37,1.98]$
	Zhu et al. (2018)	Memory retrieval task	Power-law exponents IRI: $\mu \in [0.77,2.39]$
			Power-law exponents distance: $\mu \in [0.76,1.28]$

1/f noise	Gilden et al. (1995)	Time interval estimation	Power spectra slope: $\alpha \in [0.90,1.20]$
		Spatial interval estimation	Power spectra slope: $\alpha = 1$
	Gilden (1997)	Mental rotation	RT power spectra slope: $\alpha = 0.7$
		Lexical decision	RT power spectra slope: $\alpha = 0.9$
		Serial search	RT power spectra slope: $\alpha = 0.7$
		Parallel search	RT power spectra slope: $\alpha = 0.7$

2.4 Mental Sampling Algorithms

If we have a mental representation, which is likely to be patchy for most real-world cognitive tasks, the empirical data suggest that our brain should generate samples in a manner that follows a Lévy flight distribution in distances and *1/f* noise in time. Given that, we now investigate which sampling algorithms can capture both these aspects of human cognition.

We consider three possible sampling algorithms that might be employed in human cognition: Direct Sampling (DS), Random walk Metropolis (RwM), and Metropolis-coupled MCMC (MC^3). We define DS as independently drawing samples in accord with the posterior probability distribution. Implementing DS in the brain requires perfect representations of target distribution, be it one-dimensional or multi-dimensional. Consequently, DS is the most efficient algorithm for sampling of the three. However, DS can only be applied to relatively simple tasks. Knowing the target distribution often requires calculating intractable normalising constants that scale exponentially with the dimensionality of the hypothesis space (MacKay, 2003; Chater, Tenenbaum, & Yuille, 2006). DS has been used to explain biases in human cognition such as probability matching (Vul et al., 2014).

MCMC algorithms bypass the problem of the normalising constant by simulating a Markov chain that transitions between states according only

to the ratio of the probability of hypotheses (Metropolis et al., 1953). We define RwM as a classical Metropolis-Hastings MCMC algorithm, which can be thought of as a random walker exploring the probability landscape of hypotheses, preferentially climbing the peaks of the posterior probability distribution (Metropolis et al., 1953; Hastings, 1970). The pseudo-code for RwM can be found below. Implementing RwM in the brain is relatively easy because it only needs the local information of target distribution. However, with a limited number of samples, RwM is very unlikely to reach modes in the probability distribution that are separated by large regions of low probability. This leads to biased approximations of the posterior distribution (Swendsen & Wang, 1986; Sanborn & Chater, 2016). Random walks have been used to model clustered responses in memory retrieval (Abbott, Austerweil, & Griffiths, 2012), and RwM in particular has been used to model multistable perceptions (Gershman, Vul, & Tenenbaum, 2012), the anchoring effect (Lieder et al., 2012), and various reasoning biases (Dasgupta et al., 2017; Sanborn & Chater, 2016). RwM, however, will struggle with multimodal probability distributions regardless of dimensionality.

Algorithm Random walk Metropolis

1: Choose a starting point x_1.
2: **for** t=2 to L **do**
3: Draw a candidate sample $x' \sim N(x_{t-1}, \sigma)$
4: Sample $u \sim U[0,1]$, and compute $A = \min\{1, \frac{\pi(x')}{\pi(x_{t-1})}\}$
5: **if** $u < A$ **then** $x_t = x'$ **else** $x_t = x_{t-1}$ **end if**
6: **end for**

Our third algorithm is MC^3, also known as parallel tempering or replica-exchange MCMC, was one of the first algorithms to successfully tackle the problem of multimodality (Geyer, 1991). MC^3 involves running M Markov chains in parallel, each at a different temperature: $T_1, T_2, \ldots, T_M$. In general, $1 = T_1 < T_2 < \ldots < T_M$, and T_1 is the temperature of the interest where the target distribution is unchanged. The purpose of the heated

chains is to traverse valleys in the probability landscape and to propose moves to far-away peaks (by sampling from heated target distributions: $\pi^{1/T}$), while the colder chains make the local steps that explore the current probability peak or patch. MC^3 decides whether to swap the states between two randomly chosen chains in every iteration (Geyer, 1991). In particular, the swapping of chain i and j is accepted or rejected according to a Metropolis rule; hence the name Metropolis-coupled MCMC.

$$A^{swap} = \min\{1, \frac{\pi(x_j)^{1/T_i}\pi(x_i)^{1/T_j}}{\pi(x_i)^{1/T_i}\pi(x_j)^{1/T_j}}\} \tag{2.4}$$

The coupling induces dependence among the chains, so each chain is no longer Markovian. The stationary distribution of the entire set of chains is thus $\prod_{i=1}^{M} \pi^{1/T_i}$, but we only use samples from the cold chain ($T = 1$) to approximate the posterior distribution (Geyer, 1991). Pseudo-code for MC^3 is presented below. Note that MC^3 can reduce to RwM when the number of parallel chains $M = 1$.

Algorithm Metropolis-coupled Markov chain Monte Carlo

Choose a starting point x_1.
for t=2 to L **do**
  for m=1 to M **do**
    Draw a candidate sample $x' \sim N(x_{t-1}^m, \sigma)$
    Sample $u \sim U[0,1]$, and compute $A^m = \min\{1, [\frac{\pi(x')}{\pi(x_{t-1}^m)}]^{1/T_m}\}$
    if $u < A^m$ **then** $x_t^m = x'$ **else** $x_t^m = x_{t-1}^m$ **end if**
  end for
  repeat $\lfloor M/2 \rfloor$ **times**
    Randomly select two chains i,j without repetition
    Sample $u \sim U[0,1]$, and compute $A^{swap} = \min\{1, \frac{\pi(x_t^j)^{1/T_i}\pi(x_t^i)^{1/T_j}}{\pi(x_t^i)^{1/T_i}\pi(x_t^j)^{1/T_j}}\}$
    if $u < A^{swap}$ **then** swap(x_t^i, x_t^j) **end if**
  end repeat
end for

2.5 Algorithm Selection [1]

In this section, we evaluate whether the two key empirical effects of Lévy flights and *1/f* autocorrelations can be produced by mental sampling algorithms.

2.5.1 Producing Lévy flights with Sampling Algorithms

To simulate the sampling algorithms, we use a spatial representation of semantics (rather than graph structure used in semantic networks), and we justify this choice in the Appendix section 2.7.2. For generality, we first focus on simulating patchy environments without making detailed assumptions about any one participant's semantic space. In particular, we created a series of 2D environments using $N_{mode} = 15$ Gaussian mixtures where the means are uniformly generated from $[-r, r]$ for both dimensions, where $r = 9$ and the covariant matrix is fixed as the identity matrix for all mixtures. This procedure will produce patchy environments (for example the top panel of Figure 2). We ran DS, RwM, and MC^3 on this multimodal probability landscape, and the first 100 positions for each algorithm can be found in the top panel of Figure 2. The empirical flight distances were obtained by calculating the Euclidean distance between two consecutive positions of the sampler. For MC^3, only the positions of the cold chain ($T = 1$) were used.

[1] Relevant code for this section can be found at Open Science Framework: https://osf.io/26xb5

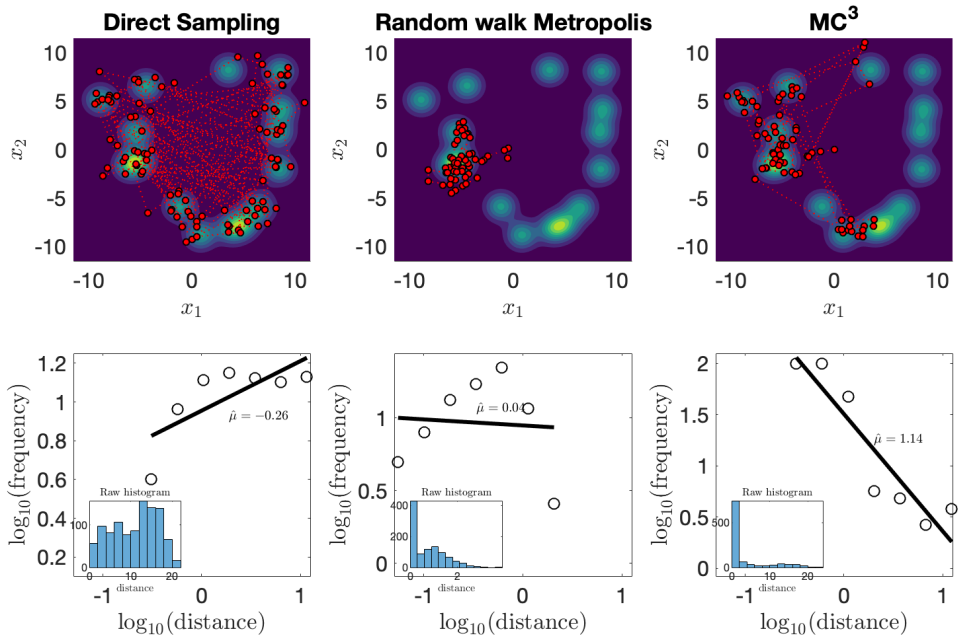


Figure 2. An example of searching behaviours in a 2D patchy environment. Each patch could represent a cluster of animal names (with two principal components x_1, x_2). Data points in lower panels represent binned histogram in log-log plot. Repeated simulation of samplers in different environments can be found in Figure 3. (**Left Panel**) Simulation result for DS. The top panel shows the trajectory of the first 100 positions (red dots). The bottom panel shows the log-log plot of flight distance distribution. The raw histogram of flight distance is also included in the bottom panel. The power-law exponent is fitted using the LBN method, which corrects for irregular spacing of points (Rhodes & Turvey, 2007). (**Middle Panel**) Same results from the RwM sampler. The Gaussian proposal distribution was an identity covariance matrix. (**Right Panel**) Same result for the MC^3 sampler with 8 parallel chains; only the positions of the cold chain are displayed here. The Gaussian proposal distributions for all 8 chains had the same identity covariance matrix. For all three samplers considered here, only the first 1024 samples were used in order to match the length of human experiments.

Power-law distributions should produce straight lines in a log-log plot. To estimate power-law exponents of flight distance, we used the normalised logarithmic binning (LBN) method as it has higher accuracy

than other methods (Rhodes & Turvey, 2007; Viswanathan et al., 1999). In the LBN, flight distances are grouped into logarithmically-increasing sized bins and the geometric midpoints are used for plotting the data. Figure 1 (bottom) shows that only MC^3 can reproduce the distributional property of flight distance as a Lévy flight with an estimated power-law exponent $\hat{\mu} = 1.14$. Both DS ($\hat{\mu} = -.26$) and RwM ($\hat{\mu} = .04$) produced values outside the range of power-law exponents found in human data. Indeed, RwM produces a highly non-linear log-log plot, differing in form as well as exponent from a Lévy flight. In the Appendix section 2.7.4, we support this result by showing how sampling from a low-dimensional semantic space representation of animal names with MC^3 can produce Lévy flight exponents similar to those of produced by participants for distances.

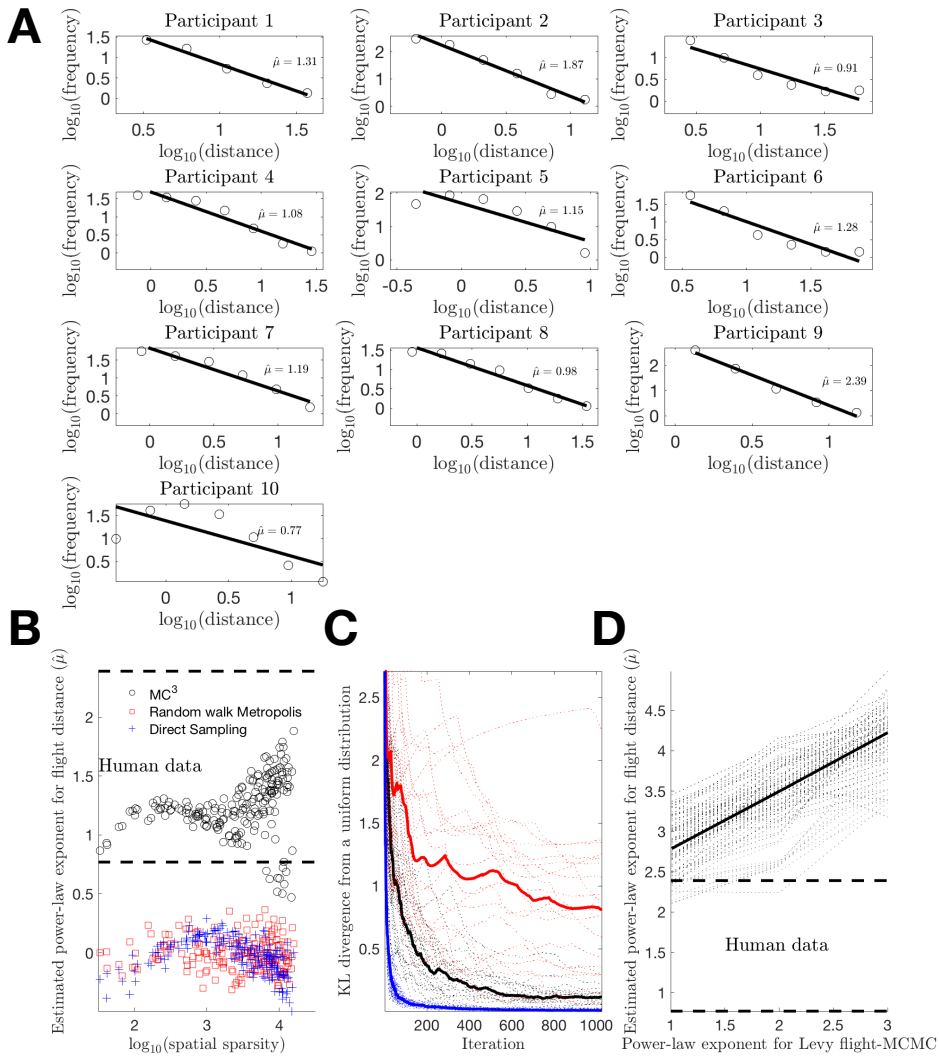


Figure 3. **(A)** Animal naming task as non-destructive mental foraging (10 participants). The estimated power-law exponents for the IRIs are $\mu \in [.77,2.39]$. **(B)** Estimated power-law exponents for flight distance distributions for the three sampling algorithms across different patchy environments, manipulating the spatial sparsity of the Gaussian mixtures. The dashed lines show the range of power-law exponents suggested by our human data. Only MC^3 falls in this range. **(C)** KL divergence of mode visitation from the true distribution for the three sampling algorithms. Red denotes RwM, black denotes MC^3, and blue denotes DS. The patchy environments are the same for all three algorithms. The quicker the sampler approaches zero KL divergence, the better the sampler is searching the patchy environment. The solid lines are medians of the dashed lines. **(D)** Simulated standard MCMC with power-law proposal distribution. The solid

line shows the median in estimated power-law exponent. The dashed lines show the range of human data.

Note that only one run of all three samplers in a patchy environment is shown in Figure 2. We also demonstrate the same samplers in different patchy environments with different spatial sparsities where the impact of spatial sparsity on the estimated power-law exponents was investigated (see Figure 3B). In these simulations, the same number of Gaussian mixtures was used but the range r was varied: with higher r, the patchy environment was more likely to be sparse. The spatial sparsity was formally defined as the mean distance between Gaussian modes. With small or moderate spatial sparsity we found a positive relationship between spatial sparsity and the estimated power-law exponents for both DS and MC^3 (Figure 3B). In this range, only MC^3 produced power-law exponents in the range reported in our mental foraging task unlike DS and RwM. For both local sampling algorithms (RwM and MC^3), once spatial sparsity was too great, only a single mode was explored and no large jumps were made.

We then varied the values of hyperparameters and tested whether this result is robust. In particular, we sampled 4 different values respectively for temperature spacing {.5, 3, 7, 10} and the number of parallel chains {2, 4, 6, 10}, resulting in 16 combinations of hyperparameters. Intuitively, larger temperature spacing, more parallel chains, and greater step size should lead to more explorative behaviour of the sampler, and vice versa. Hence, for a certain environmental structure, MC^3 could tune these hyperparameters to balance between explorative and exploitative searches. For searches in the semantic space of animal names, we ran MC^3 repeatedly 10 times, and the mean of these power-law exponents was considered. 62.5% of hyperparameters reproduced Lévy flights.

We also checked whether MC^3 really is more suitable to explore patchy mental representations than RwM. In our simulated patchy environments, which used Gaussian mixtures with identity covariance matrix, an optimal sampling algorithm should visit each mode equally often,

hence will thus produce a uniform distribution of visit frequencies over all the modes. To this end, the effectiveness of exploring such a mental representation can be examined by computing a Kullback-Leibler divergence (KL) (MacKay, 2003) between a uniform distribution over all modes and the relative frequency of how often an algorithm visited each mode:

$$D_{KL}(H_{1:t}||U) = \sum_{i=1}^{N_{mode}} H_{1:t} \log \frac{H_{1:t}}{1/N_{mode}} \tag{2.5}$$

where U is a discrete uniform distribution, N_{mode} is the number of identical Gaussian mixtures, and $H_{1:t}$ is the empirical frequency of visited modes up to time t. Samples were assigned to the closest mode when determining these empirical frequencies. The faster the KL divergence for an algorithm reaches zero, the more effective the algorithm is at exploring the underlying environment, and the DS algorithm serves as a benchmark for the other two algorithms. As shown in Figure 3C, MC^3 catches up to DS, while RwM lags far behind in exploring this patchy environment.

We checked whether the negative results for RwM were due to the choice of proposal distribution, by changing the Gaussian proposal distribution to a Lévy flight proposal distribution which has a higher probability of larger steps. Using a Lévy flight proposal distribution will straightforwardly produce power-law flight distance if the posterior distribution is uniform over the entire space (i.e., every proposal will be accepted). However, in a patchy environment, a Lévy flight proposal distribution will not typically produce a Lévy flight distribution of distances between samples that has estimated power-law exponents in the range of human data, as also can be seen in Figure 3D using different spatial sparsities. The reason for this is that the long jumps in the proposal distribution are unlikely to be successful: long jumps of ten propose new states that lie in regions of nearly zero posterior probability.

2.5.2 Producing *1/f* noise with Sampling Algorithms

To study the serial correlation of mental samples, different experimental designs were used. A typical interval estimation task requires participants to repeatedly produce an estimate of the same target interval (Gilden et al., 1995; Gilden, 1997). For instance, participants were first given an example of a target interval (e.g., 1-second time interval or 1-inch spatial interval) and then repeatedly attempted to reproduce this target without feedback for up to 1000 trials. The time series produced by participants showed *1/f* noise with an exponent close to 1. However, the log-log plot of the human data is typically flattens out for the highest frequencies (Gilden et al., 1995). This effect has been explained as the result of two processes: fractional Brownian motion combined with white noise due to motor errors at the highest frequencies (Gilden et al., 1995).

We investigated how well our three sampling algorithms can explain the autocorrelations in this temporal estimation task (Figure 4A: Gilden et al., 1995). Gaussian distributions were used as target distribution for all three sampling algorithms because the distribution of responses produced by participants was indistinguishable from a Gaussian (Gilden et al., 1995). For temporal estimation, it is known that the Gaussian distributions of responses have a scalar property that resembles Weber's law: the ratio of the mean to the standard deviation is constant (Rakitin, Gibbon, Penney, Malapani, Hinton, & Meck, 1998; Gibbon, 1977). For these simulations, we set this ratio between the mean and the standard deviation equal to 8 (Rakitin et al., 1998).

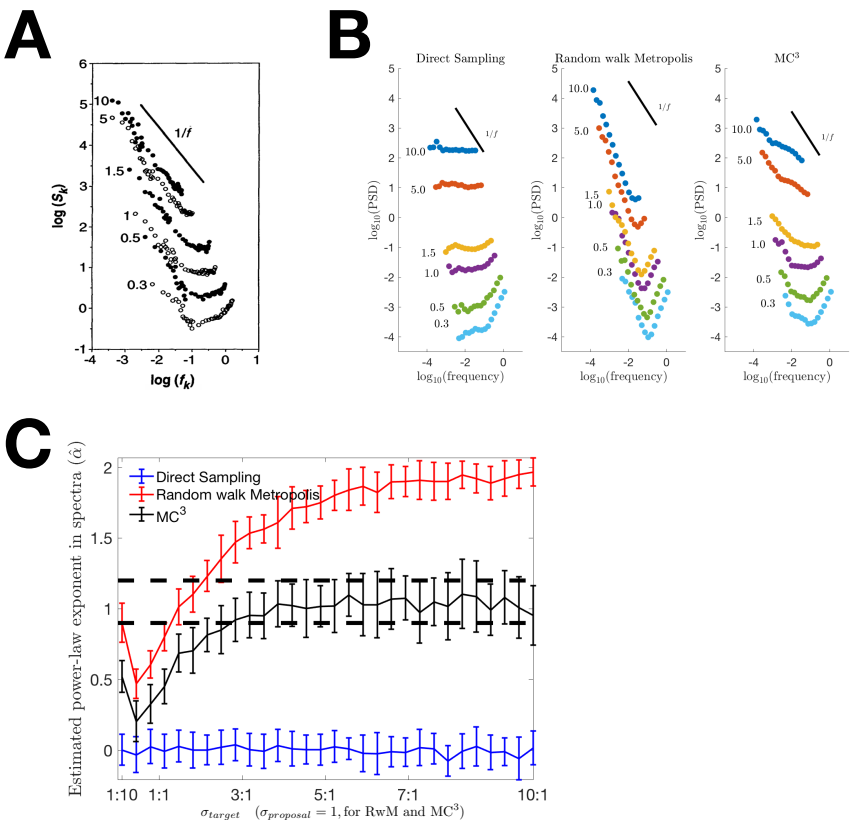


Figure 4. **(A)** Estimates of time duration show *1/f* noise. The target durations for participants to estimate are shown next to scatterplots, and the target duration ranged from 10s (top) to 0.3s (bottom). Best fit power-law exponents to the power spectra are $\alpha \in [0.90,1.20]$, and this is also the range shown in dashed lines in Figure 4C. Figure was adapted from Gilden et al. (1995). **(B)** Power spectra produced by DS (left), RwM (middle), and MC^3 (right). Only MC^3 with 8 parallel chains can generate *1/f* noise. For all the sampling algorithms, the first 1024 samples were used. **(C)** Estimated power-law exponent in power spectra are related to the ratio between Gaussian width and proposal step size. The power-law exponents for power spectra ($\hat{\alpha}$) were fitted following methods suggested by (Gilden et al., 1995; Gilden, 1997). The dashed lines show the range of $1/f^{\alpha}$ suggested by Gilden et al. (1995). Error bars indicate ±SEM. When the ratio is low the acceptance rate of proposed sample should be low; it is the opposite case for the high ratio. The asymptotic behaviours of MC^3 are *1/f* noise, of RwM are brown noise, and of DS are white noise.

We then ran the sampling algorithms on the target durations tested by Gilden et al. (1995). Unlike in the simulations of distances between samples above, the time estimates produced by participants are estimates so we can directly compare them to the samples produced by the algorithms. RwM and MC^3 were initiated at the mode of the Gaussian distribution, and there was no burn-in period in our simulations. As in Gilden et al. (1995), for all three algorithms we added Gaussian motor noise to each sample to fit the upswing in the plot at higher frequencies. As each trial in the experiment started immediately as the previous trial ended, this resulted in the recorded estimate being equal to the sample plus the motor noise, but minus the motor noise from the previous trial, producing high frequency autocorrelations. Our motor noise had a constant standard deviation of 0.1. Overall, the results show that only MC^3 produces *1/f* noise ($\hat{\alpha} \in [0.5,1.5]$), whereas DS tends to produce white noise ($\hat{\alpha} \in [0,0.5]$) and RwM is closest to Brown noise ($1/f^2$: $\hat{\alpha} \in [1.5,2]$).

RwM tends to generate Brown noise because, if every proposed sample is accepted, then the algorithm reduces to a first-order autoregressive process (i.e., AR(1)). This can be seen numerically by running the sampling algorithms using different ratios of the target distribution and proposal distribution standard deviations (Figure 4C). To see this relationship more clearly, in Figure 4C we did not add any motor noise. When the Gaussian width (σ_{target}) of the target distribution becomes much greater than the width of the Gaussian proposal distribution ($\sigma_{proposal}$), RwM produces Brown noise. In contrast, MC^3 has a tendency to produce *1/f* noise when the acceptance rate is high (Figure 4C black line). It has been shown that the sum of as few as three AR(1) processes with widely distributed autoregressive coefficients produces an approximation to *1/f* noise (Ward, 2002). As the higher-temperature chains can be thought of as very roughly similar to AR(1) processes with lower autoregressive coefficients, this may explain why the asymptotic behaviour of the MC^3 is *1/f* noise.

Note that, from an effective sample size perspective, DS is clearly the best among three sampling algorithms. The cognitive emission of *1/f* noise

is very suboptimal from a statistical standpoint as it produces a smaller effective sample size than the independent samples drawn using DS or the mild autocorrelations found in RwM. However, our sampling account provides a reason for why the mind would produce *1/f* noise: these long-range autocorrelations need to be tolerated in order to retain the possibility of generating samples from far-reaching modes.

We did a similar robustness check for hyperparameters settings using the same 16 combinations as above. For search in representation of temporal interval, only the 10s target interval was considered as it shows least influence of motor noise in the power spectra (Figure 4A). Looking across both applications, 43.75% parameters reproduced *1/f* noise. Combined, 18.75% parameters reproduced both Lévy flights in the animal name example and *1/f* noise in the target duration example.

2.6 Discussion

Lévy flight are advantageous in a patchy world and have been observed in many foraging task with humans and other animals. A random walk with Gaussian steps does not produce the occasional long-distance jump as a Lévy light does. However, the swapping scheme between parallel chains of MC^3 enables it to produce Lévy-like scaling in the flight distance distribution. Additionally, MC^3 produces the long-range slowly-decaying autocorrelations of *1/f* scaling. This long-range dependence rules out any sampling algorithm that draws independent samples from the posterior distribution, such as DS, because the sample sequence would have no serial correlation (i.e., white noise). It also rules out RwM because the current sample solely depends on the previous sample. Both of these results suggest that the algorithms people use to sample mental representations are more complex than DS or RwM, and, like MC^3, are instead adapted to sampling from multimodal distributions.

However, if people are adapted to multimodal distributions, their behaviour appears not to change even when they are actually sampling from

a unimodal distribution. In Gilden's experiments, the overall distribution of estimated intervals (i.e., ignoring serial order) was not multimodal, instead it was indistinguishable from a Gaussian distribution (Gilden et al., 1995). Assuming that the posterior distribution in the hypothesis space is also unimodal then it is somewhat inefficient to use MC^3 rather than simple MCMC. Potentially, the brain is hardwired to use particular algorithms, or it is slow to adapt to unimodal representations because it is very difficult to know that a distribution is unimodal rather than just a single mode in a patchy space. Of course, it could be that even if MC^3 is always used, that the number of chains or temperature parameters are adapted to the task at hand. In additional, it may be that a cognitive load manipulation would reduce the number of available chains and thus reduce exploration, which is an interesting prediction to test in future work.

Previous explanations of scale-free phenomenon in human cognition such as self-organised criticality argue that *1/f* noise is generated from the interactions of many simple processes that produce such hallmarks of complexity (Van Orden, Holden, & Turvey, 2003). Other explanations assume that it is due to a mixture of scaled processes like noise in attention or noise in our ability to perform cognitive tasks (Wagenmakers, Farrell, & Ratcliff, 2004). These approaches argue that *1/f* noise is a general property of cognition and do not tie it to other empirical effects. Our explanation of this scale-free process is more mechanistic, assuming that they reflect the cognitive need to gather vital information resources from multimodal probability distributions. While autocorrelations make samplers less effective when sampling from simple distributions, they may need to be tolerated in multimodal distributions in order to sample other isolated modes.

2.6.1 Neural Sampling Hypothesis

An avenue for future work is to consider how MC^3 might be implemented in the brain. Researchers have proposed a variety of mechanisms for how sampling algorithms could be implemented in the brain, and these mechanisms can account for many neural response properties including firing rate statistics in cortical neurons (e.g., Hoyer &

Hyvarinen, 2003; Orban et al., 2016; Aitchison & Lengyel, 2016). We are not aware of any implementations of MC^3 in particular, but other work has proposed how multiple chains could be implemented in neural hardware (Savin & Deneve, 2014). Adapting this existing multiple-chain scheme to implement MC^3 would require: (a) running the different chains at different temperatures, (b) tracking the cold chain for the output samples, and (c) implementing a mechanism for switching states (or equivalently switching temperatures) between chains.

2.6.2 External Search

While we have evaluated MC^3 for internal sampling, it is interesting to consider whether it might describe some aspects of external search as well. Eye movements of searching objects in natural images have also been shown to produce both Lévy flight and *1/f* noise (Rhodes, Kello, & Kerster, 2011). For example, participants in this type of task were asked to count the number of sheep in a photo and recored eye movements, which shows both Lévy flight and *1/f* noise. Certainly, the areas of interest (e.g., locations of sheep) in natural images are multimodal. How to map internal sampling mechanism to external search behaviours remain open to future research.

2.6.3 Concluding Remarks

Of course, we do not claim that MC^3 is the only sampling algorithm that is able to produce both *1/f* noise and Lévy flights. It is possible that other algorithms that deal better with multimodality than MCMC, such as running a single chain at different temperatures (Neal, 1996; Savin, Dayan, & Lengyel, 2014) or Hamiltonian Monte Carlo (Aitchison & Lengyel, 2016; Duane, Kennedy, Pendleton, & Roweth, 1987), could produce similar results. Future work will further explore which algorithms can match these key human data.

The sampling process articulated in the MC^3 algorithm suggests a more sophisticated search strategy in mental representations beyond simple local search (such as MCMC). The behavioural outputs of the MC^3

sampling process exhibit Lévy-flight distribution in distances and *1/f* noise in time. This result indicates that the mind may well-adapted to multimodal representations and actively looks out for opportunities for long-distance jumps in order to reach far-away modes. To be able to execute long jumps (and potentially not sacrifice too many benefits of a local search strategy — i.e., only relative frequency is sufficient to search locally), the MC^3 algorithm envisions a coordinated system among many samplers that have distinct search behaviours quantified by the temperature: high temperature samplers tend to be more explorative and the opposite case is true for low temperature samplers. The information is then shared among these samplers to collectively produce a picture of mental representations. In the next chapter, we will further discuss how to utilise the knowledge of how mental samples are generated to achieve better estimates of the statistics of these samples.

2.7 Appendix

2.7.1 Lévy flights do not generate *1/f* noise

In a Lévy flight, the direction of the flight is selected at random but the flight distance is distributed according to a power law (Schlesinger, Zaslavsky, & Frisch, 1995; Viswanathan et al., 1996). In a one-dimensional space, whether to move to the left or right is selected with equal probability, then the flight distance is selected according to:

$$l \sim U^{-1/(\mu-1)} \tag{2.6}$$

where U is the uniform distribution on [0,1]. This procedure guarantees that the distribution of flight distances follows a power law with exponent μ.

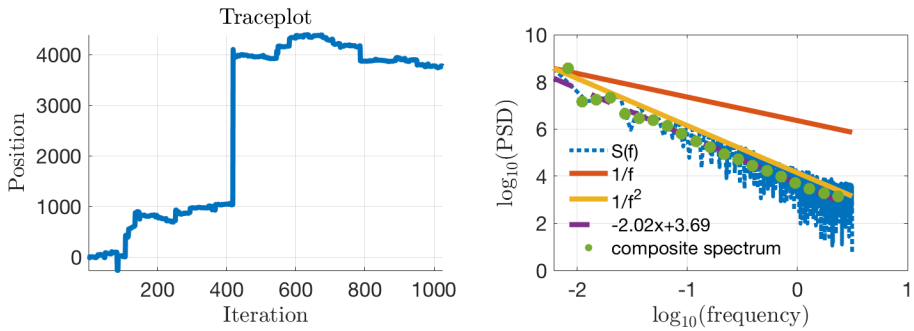


Figure S1. Autocorrelations produced by a Lévy flight. **(Left)** The trace plot of the first 1024 locations of the Lévy flight. **(Right)** The power spectra of the locations.

In Figure S1, we simulated a Lévy flight and applied the same power spectra analysis on the trace plot that we did in the main text. Lévy flights produce independent increments so the location only depends on the previous location, and indeed the simulated Lévy flight produced $1/f^2$ noise (with estimated power-law exponent $\hat{\alpha} = 2.02$).

2.7.2 Justifying a Semantic Space

Semantic representations are generally modelled with either a semantic space or a semantic network, and the algorithm that fits human data best can depend on the choice of representation (Abbott et al., 2012; Hills, Jones, & Todd, 2012). To test which representation is better to use for testing whether sampling algorithms can produce Lévy flights, we recruited two additional participants to complete a memory-retrieval task similar to the animal-naming task. However, in this task participants were allowed to report any noun as it came to mind, and not just animal names. Sampling algorithms using a semantic network should almost always predict IRI=1 in this case, since almost all the nodes in the network can be a legal response. However, a semantic space could still potentially produce power-law distributions of IRIs: under the simple assumption that the sampler reports the nearest noun, there can be many samples generated before the nearest noun changes.

Figure S2 shows that our two pilot participants produced power-law IRIs instead of constant IRIs. This relationships does not seem to hold for all of the IRIs, as the solid lines do not fit the data perfectly well, but we are most concerned with whether the longer IRIs follow a power-law distribution. When restricting our analysis to IRI>2s, the data do follow a power-law distribution as the dotted line fits the data well. This justifies our choice of a semantic space for this analysis.

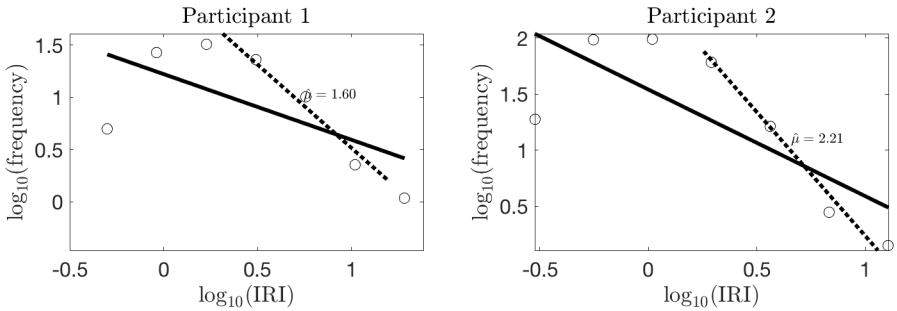


Figure S2. Histogram of IRIs (log-log plot) for two participants in noun-recall task. The estimated power-law exponents for the tail distribution are $\hat{\mu} = 1.60$ (participant 1) and $\hat{\mu} = 2.21$ (participant 2).

2.7.3. Measuring Sample Distances in the Animal Naming Task

To more directly investigate whether distances in a semantic space can be a good approximation of IRI, we mapped the animal names that our participants produced into a 300-dimensional Word2Vec semantic space (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). We first found the closest word (using the Ratcliff/Obershelp pattern-matching algorithm as implemented in the difflib,SequenceMatcher function in Python) within the Word2Vec dictionary for each participant response, as well as the animal terms identified by (Troyer et al., 1997). This resulted in 326 animal names with Word2Vec representations.

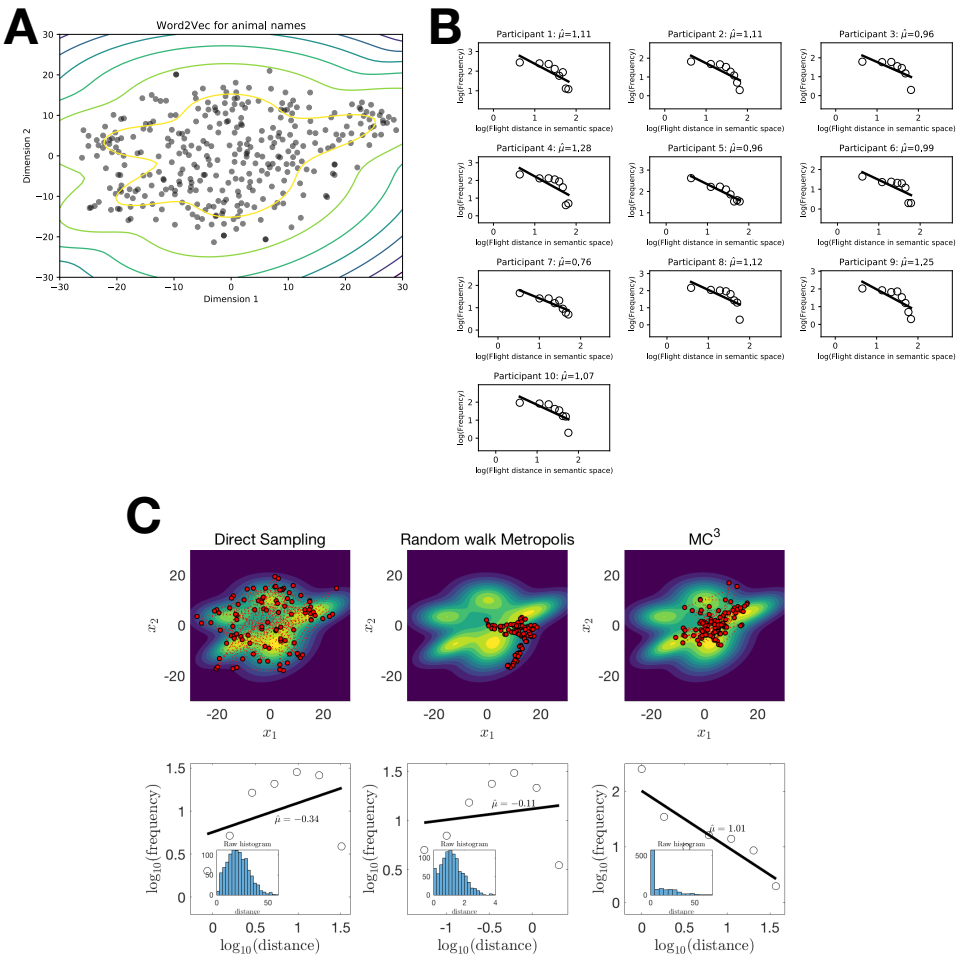


Figure S3. **(A)** 2D semantic space of all animal names. Each dot denotes one animal name. The contour represents a Gaussian mixture model on these animal names. **(B)** Histogram of flight distances for 10 participants from the animal naming task. The estimated power-law exponent $\hat{\mu} \in [0.76,1.28]$. Median correlation coefficient between the flight distance and IRIs is 0.19. **(C)** Running three sampling algorithms on the Gaussian mixture model from B. As shown above, only the MC^3 can replicate the power-law scaling of flight distance in the semantic space.

We assume that the representation of animals lies within some kind of manifold within the more general Word2Vec space, so in order to better represent the distances between animal names, we applied t-SNE to reduce the dimensionality of the space of the 326 animal names while respecting the manifold structure (Maaten & Hinton, 2008). The perplexity parameter

for t-SNE was set at 33 because this is the median category size of animal names suggested by (Troyer et al., 1997). The resultant 2D semantic space was shown in Figure S3A.

Using this 2D representation, we calculated the distances between successively reported animal names for all 10 participants, and found that the median correlations coefficient between the flight distances and IRIs was 0.19, better than the median correlation coefficients we found for the 3D (0.04) or 4D (0.04) representations. Analysing the distribution of distances between successive samples, we found they approximately have a power-law scaling. As a result, we chose to run our samplers in a 2D semantic space above.

2.7.4 Sampling in a Semantic Space

Using the low-dimensional representations we found in the previous section, we used a Dirchlet process Gaussian mixture model with the default parameters (Pedregosa et al., 2011) to infer the probability distribution of animal names. The model found eight effective Gaussian components, and the mixture distribution is plotted as contours in Figure S3C. Our three candidate sampling algorithms were run on the same mixture distribution, and the resulting power-law exponents were estimated. Only MC^3 produced exponents with the same sign as the human data.

Chapter 3

From Sample to Estimate

"The mind, like the sense of sight, has its illusions; and just as touch corrects those of the latter, so thought and calculations correct the former." (Pierre-Simon Laplace, 1825).

3.1 Introduction

The previous chapter lays out how the process of sampling could play an important role in human judgments. In this chapter, however, our focus is not on the process of sampling, but on the complementary, and neglected, question of how frequencies in a mental sample are converted into probability estimates. We will see that an analysis of this process provides a distinct mechanism through which to explain apparent biases in probability judgments.

It is well known that human probability judgments are biased, apparently suggesting that human probabilistic reasoning is not based on normative Bayesian principles, but instead on heuristic approximations of various kinds (e.g., Tversky & Kahneman, 1974; Gigerenzer & Gaissmaier, 2011). The large literature on the psychology of human probabilistic judgment has therefore emphasised human irrationality and downplayed any systematic coherence in how the brain deals with uncertainty.

Yet this research tradition, which can be traced back to Laplace's chapter "Des illusions dans l'estimation des probabilites" ("On illusions in the estimation of probabilities," in the 1995 translation by Andrew Dale from which we take all our quotes[2]) in early 1800s and later cognitive psychology and behavioural economics literature (e.g., Peterson & Beech, 1967; Kahneman, 2011; Todd & Gigerenzer, 2000), appears to stand in sharp contrast with the prevalence and usefulness of Bayesian models across

[2] Laplace's first complete presentation of the material for this chapter came in the 4th edition of his book, the *Essai Philosophique sur les Probabilites*, from 1819 (Stigler, 2005)

the cognitive and brain sciences, ranging over perception (Knill & Richards, 1996; Yuille & Kersten, 2006; Gershman, Vul, & Tenenbaum, 2009), language processing (Chater & Manning, 2006; Griffiths, Steyvers, & Tenenbaum, 2007), categorisation (Sanborn et al., 2010), naive physics (Sanborn et al., 2013; Battaglia et al., 2013), motor control (Wolpert, 2007), social reasoning (Baker, Saxe, & Tenenbaum, 2011; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017), and animal learning (Courville, Daw, & Touretzy, 2006; Gershman, Blei, & Niv, 2010; Legge, Madan, Spetch, & Ludvig, 2016). Indeed, the "new paradigm" in the psychology of reasoning (Evans & Over, 2013) even proposes that high-level explicit reasoning and argumentation is best understood in probabilistic terms (e.g., Oaksford & Chater, 1994; Hahn & Oaksford, 2007; Chater & Oaksford, 2008).

Thus, we are faced with an apparent paradox: how can Bayesian models of cognition, and indeed reasoning, be so fruitful, when what we might view as the "basic element" of such models, human probability judgment, appears to be systematically biased?

In this chapter, we confront this apparent paradox head-on: we develop a Bayesian rational model of probability judgment, which estimates probabilities from experience. This Bayesian model operates not through the explicit symbolic calculation of probabilities, but approximates probabilistic inference by drawing samples from probability distributions. As discussed in Chapter 1, one of the major discoveries of computational statistics in the last half century is that such sampling models can efficiently approximate complex probabilistic distributions (Metropolis et al., 1953; MacKay, 1998; Robert & Casella, 2013), where symbolic computation is completely intractable (Aragones et al., 2005). Indeed, such methods are routinely used to approximate probabilistic calculations in Bayesian machine learning (Neal, 2011; Rosenthal, 2011; Ghahramani, 2015), artificial intelligence (Russell & Norvig, 2016; Korb & Nicholson, 2010), and cognitive science

(Chater, Tenenbaum, & Yuille, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011)[3].

Sample-based approximation can implement Bayesian inference without explicitly representing, or manipulating, probabilities (Sanborn & Chater, 2016; Dasgupta, Schulz, & Gershman, 2017). Inevitably, however, as sampling models are an approximation to "ideal" probabilistic inference, they will systematically diverge from the norms of probability theory. We show that these departures from probability theory, given algorithmic details of sampling, generate many of the biases observed in human probability judgments. Thus, apparently paradoxically, a Bayesian rational model can automatically generate many of the systematic deviations from probability theory observed in experimental data.

3.2 A Rational Model of Probability Judgments from Sampling

How do people estimate the probability of an event? Aside from restricted domains with specially designed devices such as coins, dice, and roulette wheels, analytic calculation is typically impossible. People can, though, rely on the recall of past cases; or our ability to imagine, through a process of mental simulation, hypothetical cases. Suppose, for example, we wonder how likely we are to knock down a coconut at a coconut shy — a popular fairground game, where the goal is to throw a ball and knock over a precariously balanced coconut. We can recall past attempts at coconut shy events, by ourselves and perhaps others; and/or we can attempt to mentally simulate the process of knocking down the coconut, perhaps using some kind of naive physical model (e.g., Sanborn et al., 2013; Battaglia et al.,

[3] Another family of approximation methods, known as variational Bayes (Blei & Jordan, 2006; Blei, Kucukelbir, & McAuliffe, 2017), optimises an approximate, simplified model of the probability distribution of interest, rather than working with a sample from that distribution. This approach may also be the starting point for neuroscientific and psychological hypotheses, although we do not consider it further here (Ma, Beck, Latham, & Pouget, 2006; Gershman & Beck, 2017)

2013; Hamrick, Smith, Griffiths, & Vul, 2015). Any given "run" of such a simulation will produce a particular trajectory of the ball, collision with the coconut, and final outcome (success or failure). And different runs of the simulation will produce different results; so, by running the simulation many times, we can accumulate a sample of successes or failures, which may inform our probability judgments.

Both sources of data, memory and simulation[4], generate a set of specific instances (whether observed or imagined); and among these instances, the cognitive system can compare the number of instances in which the event of interest occurs (a coconut is successfully knocked down) and the number of instances for which it does not (the coconuts remain in place).

3.3 Empirical Evidence for the Role of Sampling in Probability Judgment

Before we develop a specific account in more detail, note that the sampling-based viewpoint gains credibility from links to existing theoretical accounts and empirical phenomena. For example, Tversky & Kahneman (1973) suggests that one important heuristic for judging probabilities is availability in memory: that is, events or types whose instances come readily to mind will be viewed as more probable than those which do not. They note, for example, that people incorrectly judge that the likelihood that words begin with a *k* to be higher than that the likelihood that a word has *k* as the third letter, because it is easier to retrieve words by their initial letter, rather than their third letter. This perspective translates naturally into a sampling framework: any factors that impact our ability to draw mental samples will influence probability judgments (as demonstrated in previous chapters).

[4] We will explore sampling from memory and simulations in more details in the following chapters.

Differences in the ease of sampling is also one source of conjunction fallacies. Tversky & Kahneman (1983) asked participants to estimate the number of words in four pages of a novel that would fit the pattern _ _ _ _ _ n _ or fit the pattern _ _ _ _ i n g. Participants both estimated the number of _ _ _ _ i n g words to be higher and found them easier to generate. That is, items which are more easily mentally sampled are rated as more probable: and the richer cue here provides a better starting point for sampling. While arising naturally from a sampling viewpoint, these results are, of course, in contradiction to the laws of probability: all words that fit the _ _ _ _ i n g pattern also fit _ _ _ _ _ n _ pattern, and hence cannot be more frequent or probable (Sanborn & Chater, 2016; Lieder, Griffiths, & Goodman, 2012).

The sampling viewpoint also provides a natural explanation for some aspects of certain types of "unpacking" effects (Dasgupta et al., 2017). People judge the probability of the "unpacked" description *being a tax, corporate, patent, or other type of lawyer* as different from an equivalent, *being a lawyer*. If the explicitly mentioned "unpacked" elements are "likely" components, this should provide a helpful cue to sampling; on the other hand, if the unpacked elements are "unlikely" then the sampling process is biased towards searching for difficult-to-find items. Thus, by biasing the starting point of a sampling process, probability judgments with unpacked description should be enhanced or reduced, by comparison with the normal description (Sanbon & Chater, 2016; Dasgupta et al., 2017). This pattern of data is observed empirically (Sloman, Rottenstreich, Wisniewski, Hadjichristidis, & Fox, 2004; Dasgupta et al., 2017).

An indirect, but suggestive, further line of evidence comes from studies of the impact of "representativeness" (Tversky & Kahneman, 1972). Consider a category with many elements, only a small subset of which can be sampled from memory, or through mental simulation. Suppose a person is asked how likely it is that a particular item is generated from that category. Almost certainly that target item will not have been generated in the sample; the participant may naturally rely on the similarity of the target item with items in the sample: is the target 'like' the sampled items or not? So, for example, when considering the probability that a sequence HHHHHH has

been generated from flips of a fair coin, this sequence will be viewed as less likely than more irregular sequences, because it is more uniform than and thereby dissimilar to, typical sequences (Tversky & Kahneman, 1973).

3.4 From Sample Frequencies to Probability Judgments

The next question, although less studied, is how sample frequencies should be converted into probability judgments. Perhaps, the problem may seem almost trivial: surely, we can simply take the relative frequencies (e.g., of successful throws at the coconut in comparison with all throws), and identify these as the probabilities.

Such approach is potentially justifiable because, under certain conditions (e.g., the samples are independently drawn from a fixed distribution) as the sample size tends to infinity, these relative frequencies will, with high probability, be close to the true probability. Indeed, this is the rationale for the frequentist interpretation of probability: that probabilities are limiting frequencies (e.g., von Mises, 1957). Taken as a psychological proposal concerning how people form probability judgments, we call this the *relative frequency* approach to probability judgment.

This approach, while simple, leads to some unappealing consequences. Suppose, for example, that we have just one sample: perhaps we have flipped a coin once, and it landed Heads. According to the relative frequency approach, we will judge the probability of the coin falling Heads to be:

$$P_{\mathrm{RF}}(\mathrm{Heads}) = \frac{\text{No. of Heads}}{\text{No. of total flips}} = \frac{1}{1} = 1 \tag{3.1}$$

and Tails to be:

$$P_{\mathrm{RF}}(\mathrm{Tails}) = \frac{\text{No. of Tails}}{\text{No. of total flips}} = \frac{0}{1} = 0. \tag{3.2}$$

Indeed, according to this viewpoint, it is difficult to avoid the prediction that anything that has never happened before will be judged to have a probability of *0*. For example, if I play the lottery with the same number each week, it is overwhelmingly likely that I will encounter an unbroken succession of losses, but I do not conclude that the particular number therefore cannot possibly win.

From a Bayesian viewpoint, which we will develop below, what is missing in a relative frequency model is any way of integrating the observed frequencies with my prior assumption about the behaviour of coins or lotteries (e.g., that coins are mostly, but not always, fair; that the prior probability of winning a lottery is very low, and so on).

3.5 A Bayesian Sampling Model of Conservatism in Probability Judgment

How, then, might we develop a purely Bayesian approach? First, we suppose that people begin with a prior concerning the possible bias of the coin, lottery, or real-world event. Following standard Bayesian statistical practice, the natural prior distribution for the bias parameter is the so-called conjugate prior of the probabilistic process of interest — here, flipping a coin; as discussed further below, this is the Beta distribution. Moreover, symmetry considerations (e.g., the fact that there is no a priori reason to expect a bias toward Heads rather than Tails), requires that the conjugate prior be a symmetric Beta distribution: Beta(β, β), which has a single free parameter, β. This prior is then continuously updated in the light of the data: that is, on instances we are able to sample, whether retrieved from memory or generated by simulation. So, for example, as the number of successive Heads we observe increases, the more we suspect that the coin has a bias towards Heads: the posterior probability distribution of possible biases shifts in favour of biases that favour Heads. How do we then convert this posterior distribution over biases (note that this is a so-called second-order probability: a probability distribution over probabilities)? The natural

approach is to take the expected value of this distribution[5]: roughly, the average of the biases, weighted by their posterior probabilities.

We have outlined a simple Bayesian approach to probability judgment; to make this model complete requires specifying one parameter, β. We will call this the "standard Bayesian model". But how credible is this Bayesian approach as an account of human probability judgments? On the face of it, this approach seems to have two powerful arguments in its favour; but also two apparently equally powerful arguments against.

On the positive side, the first argument for the appeal of the model is that it is normatively justified. Thus, the question of *why* people should follow this model in forming probability judgments then has a clear answer: people should do this because it is the right solution to the problem. The second argument is by analogy with the apparent success of Bayesian models in other cognitive domains. Probabilistic methods have been successful at modelling human data in domains as varied as perception, categorisation, motor control, reading, language processing, naive physics, folk psychology, and reasoning (e.g., Chater, Tenenbaum, & Yuille, 2006; Oaksford & Chater, 2007; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Sanborn & Chater, 2016; Gershman & Beck, 2017). Indeed, the problem of probability judgment seems a particularly simple and direct application of the Bayesian approach. The failure of a Bayesian model in this simple case might even cast doubt on the credibility of far more complex Bayesian models in these other domains.

On the negative side, though, there are two apparently serious challenges for this approach. The first is that, as noted above, the Bayesian calculations (e.g., inferring and averaging over the posterior distribution) appear computationally daunting. We approach this challenge by borrowing standard methods from computational statistics mentioned above: the Bayesian calculations can be approximated by *sampling* from the relevant posterior probability distributions, rather than being computed directly. We

[5] The mean of posterior minimises squared errors (L2-norm); alternatively, one may use the median of posterior, which minimises absolute deviations (L1-norm).

have argued elsewhere that this may be the most appropriate interpretation of many Bayesian psychological models: that the brain is a Bayesian sampler, but does not represent or calculate with probabilities (Sanborn & Chater, 2016).

The second challenge appears more difficult. A Bayesian account, with the firm normative basis outlined above, seems ill-suited to explain the systematic biases observed when people generate probability judgments. As we have indicated, the key contribution of this thesis is to argue that many of these biases in judgment and choice arise naturally, and indeed, inevitably from the sampling approximation. Indeed, from this viewpoint, many observed probabilistic biases can be viewed as "traces" of the sampling process that underpins human probabilistic judgments. But how far is this perspective really justified?

Perhaps the most fundamental and important systematic bias in probability judgment, which has been observed repeatedly, is conservatism: people tend to avoid the extremes (i.e., values close to 0 or 1) in their probability estimates (Peterson & Beech, 1967; Edwards, 1968; Kaufman, Lord, Reese, & Volkmann, 1949; Fiedler, 2000; Hilbert, 2012). Conservatism is widespread: it has both been demonstrated in the aggregation of evidence (Peterson & Beech, 1967) and in simple probability estimates (Erev, Wallsten, & Budescu, 1994). Many have argued that there is a cognitive mechanism that regresses people's estimates toward .5 (Erev et al., 1994; Dougherty, Gettys, & Ogden, 1999; Hilbert, 2012; Costello & Watts, 2014). Specifically, the closer the true probability of an event A, $P(A)$, is to 0, the more likely it is that the estimated probability, $\hat{P}(A)$, is greater than $P(A)$, whereas the closer $P(A)$ to 1, the more likely it is that $\hat{P}(A)$ is less than $P(A)$.

Interesting, though, the systematic "bias" of conservatism is generated directly by the standard Bayesian model we have outlined (see Section 3.2). As described above, the Beta distribution prior over probabilities will moderate extreme relative frequencies. From this point of view, labelling conservatism as a "bias" is misleading. From the point of view of frequentist statistics, it is the case that, where the true probability is extreme, for example, 0, then the standard Bayesian approach will

overestimate that probability given a sample. In frequentist statistics, any difference between the expected value of an estimate, and the true value, counts as a bias. But from a Bayesian perspective, this phenomenon follows from *adhering* to the laws of probability. After all, when the true probability to be estimated is 0, a rational updating model will inevitably overestimate this probability from any finite sample (on average) — after all, a rational Bayesian model cannot rule out the possibility that the event has a positive probability, but simply has yet to occur by chance. Therefore, from the present Bayesian standpoint, some degree of conservatism is *normatively required* and hence is not necessarily properly labelled as a bias at all.

How conservative should people be? In our model, this depends on their prior distribution, characterised by the value of the β parameter in the symmetrical Beta distribution. Another potentially relevant factor, though, is the degree of correlation between samples (as we have discussed in Chapter 2). While identical independent draws are suggested by drawing from an urn with replacement (as in Edward's famous urn experiment), natural sources of data typically have interdependencies at many scales (Gilden et al., 1995; Gilden, 2001). And indeed, when people are sampling, not from observation, but from memory or mental simulation, such interdependencies will be large and unavoidable (Bousfield & Sedgewick, 1944; Zhu et al., 2018). To the extent that a person does not assume independence, further conservatism is justified — if, for example, people assume that events run in "streaks" then observing an event occurring successively many times should be weaker evidence that it is unlikely: after all, an opposite streak might be about to start at any time. For now, we assume independence, but we will return to the question of correlations in later sections.

Instead of conservatism being the result of noise as suggested by Costello & Watts (2014), we propose that it is a rational adjustment for small samples drawn from a person's belief distribution. While we assume that samples drawn will generally reflect the underlying probabilities accurately, the second stage corrects for the intrinsic uncertainty in the probabilities as a result of having a limited number of samples. This correction produces a

biased estimate that is on average more accurate than the uncorrected, unbiased estimate.

It is worth noting that our approach falls into the class of *rational process models*, which explain biases as the result of the approximation algorithm used to perform inference (Griffiths, Vul, & Sanborn, 2012; Sanborn et al., 2010; Sanborn & Chater, 2016). Recently, this approach has been extended to derive biases from a rational use of time or limited cognitive resources (Griffiths, Lieder, & Goodman, 2015; Lieder, Griffiths, & Hsu, 2017). A Bayesian sampling model is in the same spirit of the resource-rational framework as it aims to produce the best possible adjustments given a limited number of samples. In addition, its two-stage nature echoes work in neuroscience that has posited that brain regions and even neurons perform inference on the input that they receive (Pfister, Dayan, & Lengyel, 2010; Deneve, 2008).

3.6 Probability Theory plus Noise (PTN) Model

There is, though, an alternative, and arguably simpler, model of the mapping from frequencies to probability judgments to consider — that probability regression does not arise from potentially elaborate Bayesian calculations, but simply from noise in the process of storing and retrieving memories of past events. This "Probability Theory plus Noise" (PTN) approach has been pursued by Costello and Watts, in an important recent series of papers (Costello & Watts, 2014; 2017; Costello et al., 2018). The PTN model suggests that, for example, when recalling past throws at the coconut shy, our memory is noisy: some failures will be mis-remembered as successes; and some successes will be misremembered as failures. Their initial model (Costello & Watts, 2014) makes the simplest possible assumption: that the probability of misclassification is a fixed constant, which is the same for both positive and negative instances. If probability judgments were determined purely by noise of this type, then each event A, and its complement not-A, would be assigned a probability close to .5

(varying depending on the particular sample drawn); and hence a mix of veridical and noisy memories will "regress" observed relative frequencies towards .5, in proportion to the level of noise.

According to the PTN model, many "rational" patterns in the data on human probability judgment should remain intact. Misclassifications can "flip" the classification of items in the mental sample, but probabilities are still "read off" of the relative frequencies of items in this "modified"/"corrupted" sample. These relative frequencies, all derived from the same, albeit corrupted, mental sample should therefore obey the laws of probability. Using this line of reasoning, Costello and Watts identify a number of probabilistic identities that should be respected, even with "regressed" probability judgments. For example, to choose a somewhat simpler case for illustration, $P(A) + P(\neg A) = 1$ still applies in the PTN model: if A is a low probability event, then there will be more switches from not-A to A than the reverse. But each event is, nonetheless, either A or not-A, so the sum of the relative frequencies should still be 1. Costello & Watts also derive a number of more complex identities that should not be preserved in the PTN account. They have verified the predictions of the PTN account in a series of experiments (Costello & Watts, 2018; Costello, Watts, & Fisher, 2018).

The PTN model, at first glance, looks like a rival to a rational Bayesian account — which departs from rationality in the light of putative mechanistic factors, concerning the noisiness of memory. As we shall see, though, it turns out that a natural Bayesian sampling model generates predictions that are, in expectation, identical to those of the PTN model. Moreover, the two approaches diverge regarding the variability of probability judgments, yielding empirical predictions that can be tested experimentally.

3.7 Conservatism: Capturing the Key Identities

The identities in Costello, Watts, and colleagues (2014; 2016; 2018) all involve participants' estimates of either single events or combinations of two events. In these experiments, for any pair of event A and B, participants were asked to estimate $P(A)$, $P(B)$, $P(A \cap B)$, and $P(A \cup B)$. Based on each participant's probability estimates, we can examine a number of identities that add or subtract these estimates, introduced by Costello, Watts, and colleagues. What all of these identities have in common is that if participants are making estimates consistent with probability theory, the result of these identities should be indistinguishable from 0. For example, Costello & Watts (2014) introduce:

$$X(A, B) = P(A) + P(B) - P(A \cap B) - P(A \cup B) \tag{3.3}$$

and

$$Y(A, B) = P(A) + P(\neg A \cap B) - P(B) - P(A \cap \neg B). \tag{3.4}$$

As required by the probability theory, both $X(A, B)$ and $Y(A, B)$ should equal to 0 for all events A and B. Table 2 summarises a set of identities that have tested in experiments (Costello & Watts, 2014; Costello et al., 2018). Crucially, all of these identities were found to be reliably different than 0, and in such a direction as to implicate conservatism as the cause.

Table 2.

Summary of combined probability expressions tested by Costello & Watts (2014) and Costello et al. (2018)

Identities	Calculations
X	$P(A) + P(B) - P(A \cap B) - P(A \cup B)$
Y	$P(A) + P(\neg A \cap B) - P(B) - P(A \cap \neg B)$
Z_1	$P(A) + P(B \cap \neg A) - P(A \cup B)$
Z_2	$P(B) + P(A \cap \neg B) - P(A \cup B)$
Z_3	$P(A \cap \neg B) + P(A \cap B) - P(A)$
Z_4	$P(B \cap \neg A) + P(A \cap B) - P(B)$

Z_5	$P(A \cap \neg B) + P(B \cap \neg A) + P(A \cap B) - P(A \cup B)$
Z_6	$P(A \cap \neg B) + P(B \cap \neg A) + 2P(A \cap B) - P(A) - P(B)$

3.8 Computational Models of Human Probability Judgments

In this section, we introduce and compare 3 computational models of probability judgments.

3.8.1 Relative Frequency Sampling Model

The first model we consider is simply a sample-based approximation to the true probability. The probability estimate of an event A occurring can be obtained by sampling a set of episodes and counting the number in which event A occurred. From N total samples, $S(A)$ are marked as successes, or occurrence of event A. Given the true probability of event A, $P(A)$ and event not-A with probability $1 - P(A)$, this process can be characterised as N independent Bernoulli trials with success probability $P(A)$. Therefore, the eventual number of samples that successfully counts as event A follows a binomial distribution:

$$S(A) \sim \text{Bin}(N, P(A)) \tag{3.5}$$

and the number of samples that failed to count as event A is the rest:

$$F(A) \sim N - \text{Bin}(N, P(A)) \tag{3.6}$$

Given that $S(A)$ out of N samples registered as event A, people can simply respond with the relative frequency of occurrence of event A as their probability estimate of event A,

$$\hat{P}_S(A) = \frac{S(A)}{N} \sim \frac{\text{Bin}(N, P(A))}{N}. \tag{3.7}$$

By the law of large numbers, the relative frequency of the samples as probability estimates will be very close to the true probability when the total number of samples is large enough. Bernoulli estimated that more than 25,000 samples are needed for "moral certainty" about the true probability

of an event, where moral certainty means that, at least 1000:1 odds, the true probability falls within .02 of the estimated probability.

3.8.2 Bayesian Sampling Model

As noted above, while a pure sampling model will produce the correct probabilities from relative frequencies in the limit, it can produce extreme conclusions where the number of samples is small. Consider the case where you draw a single sample from the posterior of the event; it seems like a poor idea to report that you are 100% certain that an event occurred. The Bayesian approach moderates such extreme conclusions, leading to conservatism.

In the simulation below, we use a symmetric Beta distribution Beta(β, β) as the prior on all possible probability estimates. The Beta distribution is a conjugate prior probability distribution for the Bernoulli and binomial distributions. It is defined on the interval [0,1], which is of course also the interval for probability estimates. This prior reflects the degree of belief placed on every possible probability estimate that range from 0 to 1.

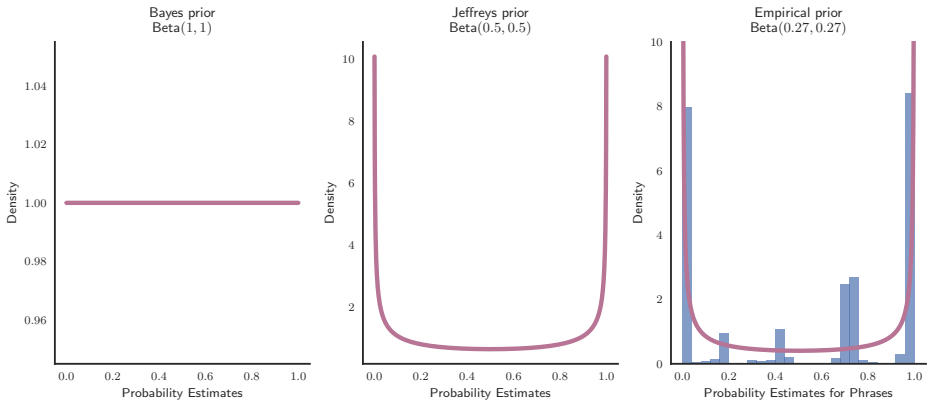


Figure 5. Illustrations of the Bayes prior (left), the Jeffrey prior (middle), and an empirical prior (right). The empirical prior was obtained by fitting the histogram of the normalised frequencies (adjusted by the proportion of uses) of the probability-describing phrases in natural language against the mean probability estimates of the same phrases (British National Corpus; adapted

from Stewart et al., 2006, Table 2). The purple curve shows the best-fitted symmetric Beta distribution:Beta(.27,.27).

What generic prior beliefs should people have about the posterior probabilities that they sample from? What makes a good generic prior for the Beta distribution is a contentious topic. As shown in Figure 5 (left), a uniform distribution, Beta(1,1), was suggested by Bayes and later adopted by Laplace in his female birth rate analysis (Bayes & Price, 1763; Laplace, 2012). Their justification was one of "ignorance" or "lack of information" as there was no reason to consider the case $p = p_1$ was more likely than the case $p = p_2$ for all possible values of $p \in [0,1]$. A uniform probability density function (pdf) is consistent with the no-preference principle on p. However, this no-preference principle does not generalise to, even, monotone transformations of p. For example, a uniform pdf on p clearly assigns preference over some log-odds ratio $\eta(p) = \log \frac{p}{1-p}$ than another; $\eta(p = .8) \approx .6$ and $\eta(p = .9) \approx .95$. An alternative prior is Jeffreys' prior (Figure 5 middle), Beta(.5,.5), which fixed the lack of monotone transformation problem for binomial distributions.

On the extreme end, Haldane's prior, Beta(0,0), is sometimes considered to represent complete uncertainty (or total ignorance) about prior information (Jaynes, 2003). Haldane's prior asserts that people are not even sure whether it is possible that samples generated in the first stage will yield either event A or event not-A. However, in reality, people should know a priori that if a sample is not marked as event A then it has to be marked as event not-A.

Though Bayes' prior, Haldane's prior, and Jeffreys' prior each have their own justifications and theoretical benefits for Bayesian inference under different contexts, for this chapter we choose to empirically estimate the generic prior people use. In particular we use the data from Stewart, Chater, & Brown (2006) who asked participants to report their probability judgments for a range of probability-describing phrases in natural language (e.g., "doubtful", "uncertain", "fair chance", "likely"). They also collected the

adjusted frequencies (raw frequency times proportions of probability uses) of each phrase based on the data from British National Corpus world edition (http://www.natcorp.ox.ac.uk/index.html). We fit these data using a symmetric Beta distribution and found that the best-fitting distribution was Beta(.27,.27)[6] when maximising the likelihood of the histogram for the Beta parameter. We will use this empirical prior throughout the simulation sections, assuming that all people share this empirical prior and use it across all situations, though this assumption is only a starting point and is likely too strong.

With the prior fixed at $\beta^* = .27$, we now consider how people would respond to the incoming samples in the first stage assuming they are drawn from the true probability $P(A)$. Given N samples collected, the Beta distribution in the second stage should get updated according to Bayes' rule. Formally, for $S(A)$ samples successfully marked as event A and $F(A)$ failed to be marked as event A, people will have a posterior probability for probability estimates that is distributed according to $\text{Beta}(\beta + S(A), \beta + F(A))$. We assume that participants then report the mean of their posterior distribution as their probability estimate. For any $x \sim \text{Beta}(a, b)$, we have the mean of x: $\mathbb{E}[x] = \frac{a}{a+b}$. Therefore, the probability estimate is a simple linear transformation of the number of successes:

$$\hat{P}_{BS}(A) = \frac{S(A) + \beta}{N + 2\beta} \sim \frac{\text{Bin}(N, P(A)) + \beta}{N + 2\beta} \tag{3.8}$$

The purpose of the correction is of course to improve the accuracy of the probability estimates. Because we derived this correction using a Beta(.27,.27) prior and used the mean of the posterior distribution as the estimate, then our estimate will certainly show improved accuracy, in terms of mean squared deviation from the true posterior probability, assuming that the distribution of posterior probabilities matches our prior.

---

[6] Fennel and Baddeley (2012) analysed blog posts and found that the distribution of the probabilities of good and bad events occurring followed Beta distributions with parameters much greater than one. We chose to use these self-generated probabilities as they seemed to be more likely to reflect subjective probabilities, though we note that very similar predictions would result using their priors.

3.8.3 Applying the Probability Theory plus Noise model

The probability theory plus noise model assumes that people estimate probabilities in a way that is fundamentally rational, but is perturbed by random noise (Costello & Watts, 2014). They suggest the bias in probabilistic estimates is the result of a memory retrieval process (or potentially an inference process), with a tallying of number of retrievals marked as event A and event not-A. Therefore, they assume that the total number of event A will be: $T_A = M \times P(A)$ where M is the number of samples drawn in total. Note that both M here and N from the sampling models above represent the total number of samples or instances drawn from memory; here we use different symbols in the interest of clarity in the simulations.

The critical mechanism proposed by the probability theory plus noise model is that recalling samples from memory is perturbed by random noise, in which each flag can be misread with a probability of d (Costello & Watts, 2014). Therefore, the eventual count of event A, $C(A)$, will be different from T_A. If we assume that random noise affects individual sample independently, then $C(A)$ should be the sum of all of the true flags minus the portion misread as false, plus all false flags that were misread as true. These two flag-flipping processes can be characterised as the sum of two binomial processes:

$$C(A) = T_A - \text{Bin}(T_A, d) + \text{Bin}(M - T_A, d). \tag{3.9}$$

The estimated probability of the event A is thus:

$$\hat{P}_{PTN}(A) = \frac{C(A)}{M} = \frac{T_A - \text{Bin}(T_A, d) + \text{Bin}(M - T_A, d)}{M} \tag{3.10}$$

3.9 Predicted Average Probability Estimates: A Mimicry Theorem

Remarkably, it turns out that the second two models, while having very different origins, precisely mimic each other's behaviour, regarding

expected probability estimates. For the probability theory plus noise model, the average probability estimates predicted by the model depends on the flag-flipping processes governed by the random noise. Every single flag, regardless of whether it is true or false for the event A, has a probability d of being read incorrectly. Thus, the flags follow a binomial distribution, $\text{Bin}(M, p)$, with mean $M \times p$ and variance $M \times p(1-p)$. $\hat{P}_{PTN}(A)$ thus has mean value of

$$\mathbb{E}\left[\hat{P}_{PTN}(A)\right] = (1-2d)P(A) + d \tag{3.11}$$

As seen in Figure 6 (left), the $\mathbb{E}\left[\hat{P}_{PTN}(A)\right]$ predicted by the probability theory plus noise model is a linear transformation of the true probability $P(A)$.

In a significant elaboration of their approach, the extended probability theory plus noise model, Costello, Watts, and Fisher (2018) suggest that increased random error is needed for conjunctive or disjunctive events. The rate of random error is enhanced from d (for single events) to $d + \Delta d$ (for conjunctions and disjunctions). In this model, the expected value of probability estimates for a conjunctive event $A \cap B$ is:

$$\mathbb{E}\left[\hat{P}_{PTN}(A \cap B)\right] = (1-2[d+\Delta d])P(A \cap B) + [d+\Delta d]. \tag{3.12}$$

Similarly, the expected value of probability estimates for a disjunctive event $A \cup B$ is:

$$\mathbb{E}\left[\hat{P}_{PTN}(A \cup B)\right] = (1-2[d+\Delta d])P(A \cup B) + [d+\Delta d]. \tag{3.13}$$

For the sampling model, the number of samples that shows an event occurred again follows a binomial distribution, and the expected estimate is unbiased:

$$\mathbb{E}\left[\hat{P}_{S}(A)\right] = P(A). \tag{3.14}$$

However, for the Bayesian sampling model, people respond with the mean of the posterior distribution as their probability estimate. Given that the mean of a Beta distribution is a fixed linear transformation of its parameters, the variability of probability estimates solely depends on the variability of the samples (i.e., $\text{Bin}(N, P(A))$). Hence, the expected value of the probability estimates predicted by the Bayesian sampling model is

$$\mathbb{E}\left[\hat{P}_{BS}(A)\right] = \frac{N}{N+2\beta}P(A) + \frac{\beta}{N+2\beta}. \quad (3.15)$$

where N is the number of sample drawn to approximate the true probability and $\beta = .27$ is the shape parameter of the generic prior over probabilities. Comparing to Equation (3.8) which gives one probability estimate, Equation (3.15) considers the expectations as the average values of repeated probability estimates made by the same model. Note that if

$$d = \frac{\beta}{N+2\beta} \text{ (bridge condition)}, \quad (3.16)$$

then $1 - 2d = \frac{N}{N+2\beta}$ and both the Probability Theory plus Noise and the Bayesian sampling models predict the exact same mean probability estimates (Figure 6).

For a conjunctive or disjunctive event, the bridge condition that connects the extended probability theory plus noise model (Costello et al., 2018) and the Bayesian sampling model is simply $d + \Delta d = \frac{\beta}{N'+2\beta}$ and then $1 - 2[d + \Delta d] = \frac{N'}{N'+2\beta}$, where N' is a new mental sample size for calculating the probabilities of combined variables.

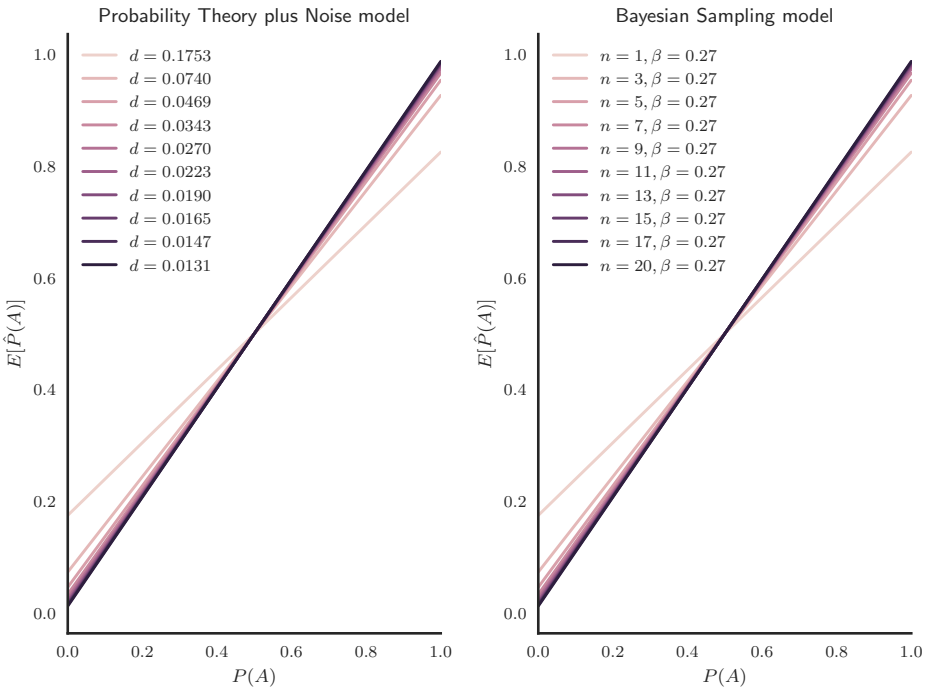


Figure 6. An illustration of model behaviours. The relationship between the true probability (x-axis) and the expected probability estimates (y-axis) predicted by the probability theory plus noise model (left) and the sampling plus correction model (right).

The predictions for the combined probability identities in Table 2 are based purely on the average probability estimates (Costello & Watts, 2014; Costello et al., 2018), so we see that the Bayesian sampling model exactly matches the predictions of the probability theory plus noise model as long as the bridge condition holds. For example, Costello and Watts (2017) found that the best-fitting values were $d = .05, \Delta d = .04$ for the probability estimation data reported in Zhao, Shah, and Osherson (2009). For this fitting result, the Bayesian sampling model can predict quantitatively the same mean values by simply applying the bridge condition: $d = \frac{\beta}{N + 2\beta} = .05$ and $d + \Delta d = \frac{\beta}{N' + 2\beta} = .09$. When the empirical generic prior is used in both cases (i.e., $\beta^* = .27$) and setting $N = 4.86, N' \approx 2.46$, the Bayesian sampling model should fit the mean values of probability estimations just as well as the probability theory plus noise model. In this sense, the increased noise

suggested by the probability theory plus noise model for conjunctive events is simply the reduced total number of samples retrieved for the same conjunctive events (a decrease from N to N').

Table 3.

Summary of model predictions (left to right: probability theory, probability theory plus noise model, sampling model, Bayesian sampling model) on the average values of the combined probability expressions from Table 1.

Identities	Prob Theory	PTN	Sampling	BS*
$\mathbb{E}[\hat{X}]$	0	$2\Delta d(P(A)+P(B))-2\Delta d$	0	$2\Delta d(P(A)+P(B))-2\Delta d$
$\mathbb{E}[\hat{Y}]$	0	$2\Delta d(P(A)-P(B))$	0	$2\Delta d(P(A)-P(B))$
$\mathbb{E}[\hat{Z}_1]$	0	$2\Delta d\,P(A)+d$	0	$2\Delta d\,P(A)+d$
$\mathbb{E}[\hat{Z}_2]$	0	$2\Delta d\,P(B)+d$	0	$2\Delta d\,P(B)+d$
$\mathbb{E}[\hat{Z}_3]$	0	$2\Delta d(1-P(A))+d$	0	$2\Delta d(1-P(A))+d$
$\mathbb{E}[\hat{Z}_4]$	0	$2\Delta d(1-P(B))+d$	0	$2\Delta d(1-P(B))+d$
$\mathbb{E}[\hat{Z}_5]$	0	$2d+2\Delta d$	0	$2d+2\Delta d$
$\mathbb{E}[\hat{Z}_6]$	0	$2d+2\Delta d$	0	$2d+2\Delta d$

*The bridge conditions were applied to the Bayesian sampling model: $d=\frac{\beta}{N+2\beta}$ and $\Delta d=\frac{(N-N')\beta}{(N+2\beta)(N'+2\beta)}$.

3.10 Where do Bayesian Sampling and PTN Differ?

While the Probability Theory plus Noise and Bayesian sampling models make indistinguishable predictions for the average estimates, they do make distinct predictions about the variability of estimates. The probability theory plus noise model predicts that the variance is independent on the true probability $P(A)$ and instead depends only on the degree of random noise (d) and the number of samples recalled (M) (Costello & Watts, 2014). In contrast, the variance of probability estimates predicted by the Bayesian

sampling model depends on the true probability (see Table 4). Holding the number of samples drawn from the posterior distribution (N) and symmetric Beta prior parameter (β) constant, the predicted variance of the probability estimates will have a quadratic relationship with $P(A)$, peaking at $P(A) = .5$.

Table 4.

Predicted variance of human probability estimates from the Probability Theory plus Noise model and the Bayesian sampling model.

	Probability Theory plus Noise	Bayesian Sampling*
Variance of Probability Estimates	$\frac{d(1-d)}{M}$	$\frac{(1-2d)^2 P(A)(1-P(A))}{N}$

*The bridge conditions were applied to the Bayesian sampling model: $d = \frac{\beta}{N + 2\beta}$

To determine which account better matches human data, we looked at the variability of the estimates of the two data sets we examined earlier: Costello et al. (2018) and Stewart, Chater, and Brown (2006). First, Costello et al. (2018) reports a series of experiments (Experiment 1, 2, and 3 in their paper) that asks participants to estimate probabilities of a range of weather events (e.g., cold, windy, or sunny) or to estimate probabilities of a range of future events (e.g., "Germany is in the finals of the next World Cup (July 2018)"). People were allowed to judge the probability in both frequency (response with number of occurrences over 100 days) and probability (response with probability of occurrence on a randomly selected day) form.

Second, we also consider the variability of the probability estimates of phrases reported by Stewart et al. (2006). As seen above, the same dataset has been used to calculate the empirical prior. However, we excluded data points with zero variance, such as the phrase "fifty-fifty chance", which was estimated to be 50% probability by all participants. We assume that for these

questions participants do not perform sampling but instead directly convert the phrase into a probability (cf. Kemp & Eddy, 2017).

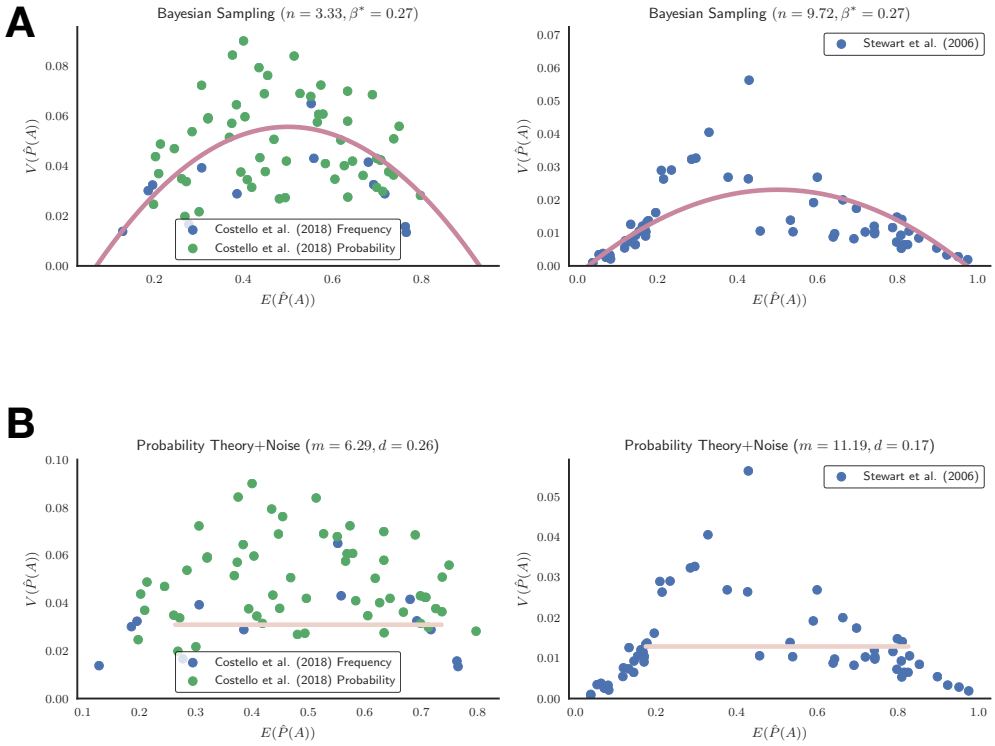


Figure 7. The relationship between the mean and the variance of people's probability estimates (left panel: Costello et al., 2018; right panel: Stewart et al., 2006). **(A)** The sampling plus correction model predicts a quadratic relationship (purple lines). **(B)** The probability theory plus noise model predicts a constant relationship (purple lines). Best-fitting model parameters are displayed in the titles, and the MSE of each model in predicting the empirical tasks can also be found in Table 5.

In Figure 7, an inverted-U shaped relationship can be found between the means and variances of the probability estimates. This is qualitatively consistent with the relationship predicted by the Bayesian sampling model. We fit both the Bayesian sampling model and the probability theory plus noise model to the data. In addition, the Bayesian sampling model uses the empirical prior as above. For both Costello et al. (2018) and Stewart et al.

(2006) tasks, we fixed $\beta^* = .27$ ex ante and thus only the total number of samples N was permitted to vary in the fitting for the Bayesian sampling model. Table 5 shows that, for both tasks, the Bayesian sampling model can achieve lower MSE and thus a better fit to the relationships between the expectations and variances of probability estimates. Because the Bayesian sampling model used fewer parameters than the probability theory plus noise model, the lower MSE values convincingly demonstrate that the Bayesian sampling model provides a better account of the variability.

Table 5.

Fitting results for both the probability theory plus noise model and the sampling plus correction model.

	Probability Theory plus Noise		Bayesian Sampling	
Experiment	Best-fitted parameters	MSE	Best-fitted parameters	MSE
Costello et al. (2018)	$M = 6.29, d = 0.26$	4.99×10^{-4}	$N = 3.33 (\beta^* = 0.27)$	2.46×10^{-4}
Stewart et al. (2006)	$M = 11.19, d = 0.17$	1.11×10^{-4}	$N = 9.72 (\beta^* = 0.27)$	0.65×10^{-4}

3.11 Discussion

In this chapter, we have argued that sampling can play a crucial role in forming probability judgments, and indeed is key to explaining aspects of well-known biases in human judgments including availability and representativeness (Lieder et al., 2012; Sanborn & Chater, 2016; Lieder et al., 2018). But this approach raises a neglected problem: how should sample frequencies be converted into probability ratings? Researchers have often implicitly assumed that probabilities can be computed taking relative frequencies, but we have seen that this gives inappropriately extreme results for small samples.

Here we provided a rational Bayesian account of how this problem can be addressed. It turns out, unexpectedly, that the approach perfectly mimics the predictions, in expectation, for a major recent theoretical account: Costello and Watts's (2014; 2018) probability theory plus noise model. We have noted, though, that the two approaches differ regarding the variability of probability estimates, and the empirical data favour the Bayesian sampling account. Here we consider where our approach fits into other work on bias in probability estimates, how our approach could be extended to more realistic sampling algorithms, and how it can explain other biases such as the conjunction fallacy and subadditivity.

The general approach outlined here (whether using the Bayesian sampling or PTN mechanisms) also captures a variety of interesting further phenomena. We considered unpacking in the introduction, where single 'unpacked' or 'packed' descriptions are compared, and observed that whether the unpacked description is judged as more or less probable depends on the likelihood of the specific unpacked descriptions that are chosen. But there is also a much more stable unpacking effect, which arises where a single probability judgment (e.g., probability of an air crash) is compared with the sum of 'unpacked' judgements (probability of an air crash due to engine failure; probability due to terrorist attack; and so on). Here, conservatism will raise each of these small probabilities, and the summing of raised probabilities will amplify the effect further.

3.11.1 Other Accounts of Bias in Probability Estimates

There are other empirically successful models of probability reasoning in the literature such as heuristic accounts (e.g., *inductive confirmation model*: Crupi & Tentori, 2016; *configural weighted average model*: Nilsson, Juslin, & Winman, 2016). However, the inductive confirmation model predicts a negative correlation between the predicted and empirical average conjunction fallacy rate (Costello & Watts, 2016), and the configural weighted average model does not provide a satisfying account on the rate of the conjunction fallacy (Costello & Watts, 2016). Furthermore, both models did not specify the detailed computational mechanism for the probability

estimate of a single event. This basically prevents these models from explaining a number of empirical results, particularly, the combined probability expressions in Table 1.

While the very existence of cognitive biases like conservatism has been seen as a sign that people are fundamentally irrational, others have argued that these irrationalities appear because people have been reasoning against the wrong normative standards (Oaksford & Chater, 2007). There have been compelling accounts of how deviations from the "correct" response can be the result of rational inference. However, these accounts cannot explain the types of inconsistencies seen in the identities in Table 2 – no matter how the problem is interpreted, these identities should still correspond with the predictions of probability theory.

3.11.2 Integrating with More Realistic Sampling Algorithms

The Bayesian sampling model assumes that people draw independent samples from their posterior distribution. But, as we have implicitly touched on in the introduction, this does not match the empirical data on how people generate hypotheses. Instead, people generate correlated hypotheses in which the identity of the next hypothesis depends on what was produced earlier. For example, in an animal-naming task, participants were asked to free recall animal names whenever it comes to mind (Bousfield & Sedgewick, 1944). The result indicated that animal names, which were reported temporally close, are also semantically similar. Similar results on the autocorrelation of mental samples are also reported in repeated time or spatial estimation tasks (Gilden et al., 1995).

These results mean that people must instead be using an algorithm that generates autocorrelated samples such as Markov Chain Monte Carlo (MCMC; Metropolis et al., 1953) or more complex alternatives (Courville & Daw, 2008; Lieder et al., 2012; Gershman et al., 2009; Zhu et al., 2018). Autocorrelated samples contain less information than independent samples, so autocorrelated samples must be weighted differently than independent

samples. Fortunately, there is an easy way to do so if the amount of autocorrelation is known. The effective sample size can be calculated

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho(k)} \tag{3.17}$$

where N is the total number of samples and $\rho(k)$ is the autocorrelation at lag k. The autocorrelated samples can then be reweighted by $\frac{ESS}{N}$ to be equivalent to independent samples. This reweighting means that the number of independent samples estimated from human data does not need to be equal to a whole number. Of course, the autocorrelation will not be known perfectly if only a short sequence of samples is generated, but here a generic value estimated over a lifetime of experience could be used here.

3.11.3 Conclusions

We have introduced a Bayesian sampling model, in which people are assumed to first draw samples from their posterior distribution and then make the best estimate possible given those samples. An exact Bayesian model does not have this uncertainty about probabilities – while there is uncertainty about the true state of the world, there is never any uncertainty about the probability of any state or collection of states. Our approach thus better matches the phenomenology of making probability estimates, as any hemming and hawing about what estimate to make clearly points to some uncertainty about the probabilities. To explain a variety of classic deviations from probability theory, we then simply assume that people are aware (at some level) of this uncertainty and adjust for it appropriately.

3.12 Appendix

3.12.1 Detailed Derivation of Model Predictions

In this section, we provide detailed mathematical operations that lead to our results in the main text. First, we consider the Probability Theory plus Noise

model (Costello & Watts, 2014). As described above, the PTN model predicts human probability estimates:

$$\hat{P}_{PTN}(A) = \frac{C(A)}{M} \tag{3.S1}$$

where the probability estimates of an event A are taken as the relative frequency of samples eventually marked as event A (after being corrupted by the read-out noise), $C(A)$, and total samples retrieved, M. The noise randomly flips the identities of samples, in this case, from event A to event not-A and vice versa. This effectively results in two independent binomial process: (a) T_A samples originally marked as event A now have $\text{Bin}(T_A, d)$ of them that changed to event not-A; (b) symmetrically, $M - T_A$ samples originally marked as event not-A now have $\text{Bin}(M - T_A, d)$ of them turned to event A. In the end, the PTN model predicts that total samples being read out as event A post noise process should be:

$$C(A) = T_A - \text{Bin}(T_A, d) + \text{Bin}(M - T_A, d) \tag{3.S2}$$

where d is the degree of random noise.

Based on Equation 3.S1 and 3.S2, we can derive the expected value of the probability estimates predicted by the PTN model

$$\begin{aligned} \mathbb{E}\left[\hat{P}_{PTN}(A)\right] &= \frac{T_A - dT_A + d(M - T_a)}{M} \\ &= (1 - 2d)\frac{T_A}{M} + d \\ &= (1 - 2d)P(A) + d \end{aligned} \tag{3.S3}$$

Similarly, we obtain the variance of the probability estimates:

$$\begin{aligned} \mathbb{V}\left[\hat{P}_{PTN}(A)\right] &= \frac{\mathbb{V}[C(A)]}{M^2} \\ &= \frac{d(1 - d)(T_A + m - T_A)}{M^2} \\ &= \frac{d(1 - d)}{M} \end{aligned} \tag{3.S4}$$

Then we consider the properties of the relative frequency sampling and Bayesian sampling models. People generate veridical samples from the true probability $P(A)$. If we assume such a process produces independent

and identically distributed samples, then out of a total of N samples, $S(A) \sim \text{Bin}(N, P(A))$ should be successfully marked as event A. In this end, the rest samples will fail to be marked as event A: $F(A) \sim N - Bin(N, P(A))$. For the relative frequency sampling model, a probability estimate can be made by simply taking the relative frequency of $S(A)$ and the total samples:

$$\hat{P}_S(A) = \frac{S(A)}{N} \sim \frac{\text{Bin}(N, P(A))}{N} \tag{3.S5}$$

However, people may also not fully trust these samples, maybe due to small sample size, stochasticity in the sample-generation process, or an existing stronger prior belief on what the true probability ought be. They can then incorporate such prior beliefs with these samples. Here, we assume people have a symmetrical prior belief that follows a Beta distribution: $\text{Beta}(\beta, \beta)$. With additional $S(A)$ samples indicating event A and $F(A)$ indicating event not-A, a Bayesian agent then updates its belief to the posterior accordingly: $\text{Beta}(\beta + S(A), \beta + F(A))$. Given that, a Bayesian agent could choose to produce its probability estimates as the mean of the posterior:

$$\hat{P}_{BS}(A) = \frac{S(A) + \beta}{N + 2\beta} \sim \frac{\text{Bin}(N, P(A)) + \beta}{N + 2\beta} \tag{3.S6}$$

We also note that the choice of reporting mean values of the posterior as probability estimates is based on Bayesian decision theory as the mean minimises squared errors. Alternatively, people could choose the median of the posterior as their probability estimate, which minimises absolute deviations. Following the formulisation of (S6), we can compute the expected value and variance of probability estimates predicted by the Bayesian sampling model:

$$\mathbb{E}\left[\hat{P}_{BS}(A)\right] = \frac{N}{N + 2\beta} P(A) + \frac{\beta}{N + 2\beta} \tag{3.S7}$$

$$\mathbb{V}\left[\hat{P}_{BS}(A)\right] = \frac{NP(A)(1 - P(A))}{(N + 2\beta)^2} = \frac{(1 - 2d)^2 P(A)(1 - P(A))}{N} \tag{3.S8}$$

3.12.2 Bayesian sampling improves accuracy from relative frequency sampling

As discussed above, we know the relative frequency sampling model will produce unbiased probability estimates, whereas Bayesian sampling predicts biased estimates. So how could a biased estimate achieve higher accuracy to the true probability than an unbiased estimate? To demonstrate this, we conducted a simulation in which we compared probability estimates predicted by the sampling and Bayesian sampling models, using several possible distributions of the true probabilities (Figure S4 x-axis). For this simulation, we repeatedly drew probabilities from the true distribution, $p_{true} = \text{Beta}(\beta_{true}, \beta_{true})$ and let both models make estimates about these true probabilities. Next, we computed the mean squared errors (MSE) between the true probabilities and estimates produced by both models. The smaller this MSE is, the more accurate the predicted probability estimates are.

In Figure S4, we subtracted the MSE of the sampling model from the MSE of the Bayesian sampling model to quantify the degree of improvement from incorporating a symmetric prior. For small number of samples (e.g., $N = 1,2,3,4$), the Bayesian sampling method significantly improved the accuracy of the estimates. For larger numbers of sample (e.g., $N > 10$), both models produce a similar level of accuracy. The estimates from the Bayesian sampling model are more advantageous as the value of β_{true} increases. This is because the sampling model predicts estimates that are equivalent to an estimate from Bayesian sampling using Haldane's Beta(0,0) prior, so the further β_{true} is from zero the better the Bayesian sampling will be.

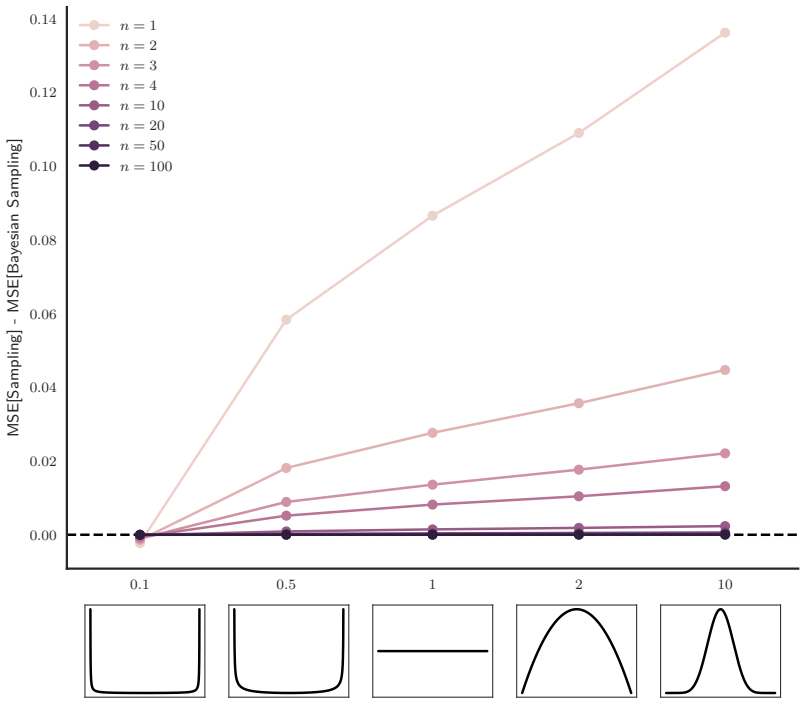


Figure S4. The degree of improvement in the probability estimate (y-axis) due to the inclusion of the correction step in the sampling plus correction model. X-axis depicts the true probability distributions from Beta(0.1,0.1) (most left) to Beta(10,10) (most right). An empirical prior, Beta(0.27,0.27), was used in the correction step as explained in the text.

Chapter 4

Sample from Memory

"Cognition is recognition." (Douglas Hofstadter, 1995)

4.1 Introduction

We now consider sampling from past computations and, more specifically, how the ability to do so may improve the efficiency of learning[7]. Here, we further restrict the examined scope of learning phenomena to classical conditioning (also known as Pavlovian conditioning). During classical conditioning, human and animal subjects alter the magnitude and timing of their conditioned responses (CR), as a consequence of the contingency between the conditioned stimulus (CS) and the unconditioned stimulus (US). The most famous example of classical conditioning was described by Pavlov (1927) where dogs were repeatedly given food powder (US) following a presentation of a sound (CS). The dogs salivated initially only at food delivery, but gradually the sound started to elicit salivation after repeated sound-food (CS-US) pairings.

The classical conditioning paradigm provides valuable empirical data for computational models of animal learning because, unlike instrumental conditioning (which we will return to in the next chapter), the experimenter has significantly more control over when each learning episode should occur. The ability to learn about the CS-US relationship plays a fundamental role in this learning process that allows animals to adapt to imminent biologically significant events (Hollis, 1982; 1997). This mechanism could be further used to, presuming a common set of theoretical principles between humans and other animals, explain the development of many abnormal behaviours

[7] Selective reuse from past inferences (amortised inference) have also been shown to improve the accuracy of inference (Stuhlmuller, Taylor, & Goodman, 2013; Gershman & Goodman, 2014).

in humans such as drug abuse (Siegel, 1989) and anxiety disorders (Bouton, 2002).

Most contemporary theories of classical conditioning employ a variation on the error-correction principle that was most notably proposed by Rescorla & Wagner (1972). The critical feature of this error-correction theory is that learning (or the change in associative strength of a stimulus) happens whenever there is a discrepancy between what animals expected from CS and what magnitude of US animals actually received. This principle alone can explain a rather wide range of experimental findings such as acquisition, extinction, and blocking. A great deal of empirical data, however, contradicts with Rescorla & Wagner's model such as spontaneous recovery and latent inhibition (see Miller et al., 1995 for a review of failures of the model). In this chapter, we will review a range of classical empirical findings in the animal learning literature, and suggest a simple extension of the Rescorla & Wagner's model — adding an additional memory component which allows animals to replay past experiences — that can avoid many of the limitations the original model. The proposed model, the *random replay* model, emphasises the role of sampling from past experiences as a complementary mechanism for associative learning and builds on an earlier replay model (Ludvig, Zhu, Mirian, Kehoe, & Sutton, under review). The random replay model assumes a more realistic memory model than the assumption of an infinite memory size in the earlier replay model. In the random replay model, memory stores a fixed number of past conditioning trials with guaranteed storage of new trial and random drop-out of an existing trial from memory. This newer memory model is now able to exhibit a certain degree of forgetting and recency.

The idea of replay has been used in neuroscience to describe the reactivation of place cells in the hippocampus during periods of rest or sleep (e.g., Davidson, Kloosterman, & Wilson, 2009; Gupta, van der Meer, Touretzky, & Redish, 2010; Euston, Tatsuno, & McNaughton 2007; Wilson & McNaughton, 1994). Reusing past experiences to assist learning also enables many applications in artificial intelligence that achieve impressive performances on video games (Mnih et al., 2015) and the ancient Chinese

board game Go (Silver et al., 2016; Silver et al., 2017). The proposed random replay mechanism is not the only way to reuse past experience. Another approach to reuse past experiences in reinforcement learning is to use these data to build a model (i.e., model learning) and then to simulate and generate new data from this learnt model (Sutton, 1990; Sutton, Szepesvari, Geramifard, & Bowling, 2008)[8]. While recent theoretical works have demonstrated an analytically optimal design of memory size and replay mechanisms in simple games (Liu & Zou, 2017; Zhang & Sutton, 2017), here we focus on the explanatory power of the simplest replay mechanism—replay at random—in animal learning. We will also show that introducing additional memory of past trials and the random replay mechanism can explain the "counterintuitive" behavioural facilitation in classical conditioning for hippocampal lesioned animals (see Schwarting & Busse, 2017 for a review).

4.2 Computational Models of Associative Learning

A computational model of associative learning should provide a formal framework for animal learning phenomenon both mechanistically and normatively. The mechanistic questions concern when and how the associative strength of a stimulus should change. The normative question concerns why the associative strength should change in the manner suggested by the mechanism.

4.2.1 Rescorla-Wagner Model

As mentioned above, the leading theory of animal learning was suggested by Rescorla & Wagner (1972). In their model, unlike earlier models, learning occurs not directly because CS-US pairings, but because such pairing is unanticipated on the basis of the current associative strength, which functions effectively as a prediction of US occurrence. This idea is

[8] We will return to this idea in the subsequent chapter

formally defined as an error-correction learning rule, whereby changes in associative strength between CS and US occur whenever differences between what was expected and what actually happened:

$$\delta_t = R_t - V_t \tag{4.1}$$

where δ_t is the difference between animal's expectations and reality (also known as the prediction error) for trial t; R_t is the actual US (often rewards such as food and liquid or aversive outcomes, like electronic shocks); V_t is the associative strength for the same trial.

To handle multiple stimuli (e.g., both tone and light) within one trial, the Rescorla-Wagner learning rule works by computing the overall associative strength for the trial as the sum of the associative strengths for all stimuli available on that trial:

$$V_t = \sum_s V_t(s)\mathbb{I}_{A_t}(s) \tag{4.2}$$

where $V_t(s)$ represents the individual associative strength for stimuli s at trial t; A_t is the set of stimuli present on that trial; $\mathbb{I}_A(s)$ is an indicator function, and it takes value of 1 when stimulus s is in the set A_t (i.e., stimulus s was present on trial t) and 0 otherwise[9].

Whenever the prediction error δ_t is non-zero, the Rescorla-Wagner learning rule prescribes that associative strengths should get updated in proportion to the prediction error:

$$V_{t+1}(s) = V_t(s) + \alpha\delta_t\mathbb{I}_{A_t}(s) \tag{4.3}$$

where $\alpha \in [0,1]$ is the learning rate that controls the size of the update.

With this simple Rescorla-Wagner learning rule, the model successfully accommodates a number of learning phenomenon in classical conditioning, including acquisition, blocking, and conditioned inhibition (see Pearce & Bouton, 2001 for a review).

[9] The indicator function can be viewed as a special case of the eligibility trace in reinforcement learning where trace length is one (Singh & Sutton, 1996; Maei & Sutton, 2010).

4.2.2 The Random Replay Model

We consider a simple extension to the Rescorla-Wagner model where animals can store past trials of conditioning and actively re-use these trial memories to assist learning (Ludvig et al., under review). As illustrated in Figure 8, there are two parallel streams of information processing in the random replay model. First, there is the usual process of associative learning as formalised in the original Rescorla-Wagner model. The animals encounter a CS-US pairing on a trial and update the associative strength of the CS in the same manner as in the classic Rescorla-Wagner model. Second, past trials (i.e., CS, US, and the timing of the trial) are also remembered in a trial memory. The animal can thus draw samples from this trial memory and replay these sampled trials like normal trials. That is, during the replay process, new prediction errors are computed based on the current associative strength and the content of the trial sampled from memory.

Clearly, how the trial memory stores past trials and how to sample from the trial memory determines the model behaviour. Here, as a proof-of-concept for the replay idea, we adopt the simplest memory storage mechanism that allows a certain degree of forgetting and recency effects[10]. The storage of trials works as if it were a leaky bucket: (a) the memory has a limited capacity (i.e., only a fixed number of trials are remembered), and (b) once the number of trials exceeds the memory capacity, a random trial is dropped out, and (c) the most recent trial is always successfully remembered. The first two rules ensure that forgetting of past trials is present, and the last two rules further regulates the forgetting such that it should be more likely for older trials than for newer trials (i.e., recency). In addition, we assume the simplest trial-retrieval mechanism — random sampling from the memory bucket with replacement.

This replay process will provide a sample of past trials that contains similar information to a normal trial: the CS, US, and the relative timing.

[10] Analytical solutions of optimal memory size and optimal replay mechanism are only available in simple settings (Liu & Zou, 2017).

The animal, then, can compute prediction errors for these samples using the Rescorla-Wagner rule:

$$\delta_t^{replay} = R_t^{replay} - V_t. \tag{4.4}$$

The difference from the standard Rescorla-Wagner learning rule is that R_t^{replay} is the remembered US for the replayed trial. V_t is, however, still the present associative strength. Similar to the Rescorla-Wagner model, the associative strength needs to be updated whenever the prediction errors from replayed trials, δ_t^{replay}, are also non-zero:

$$V_{t+1}(s) = V_t(s) + \alpha^{replay}\delta_t^{replay}\mathbb{I}_{A_t^{replay}}(s) \tag{4.5}$$

where a different and smaller learning rate $\alpha^{replay} < \alpha$ is used to update the current associative strength; $\mathbb{I}_{A_t^{replay}}(s)$ is the indicator function now placed on the replayed trial, which takes a value of 1 when the stimulus (e.g., CS or US) was presented in the replayed trial, and 0 otherwise.

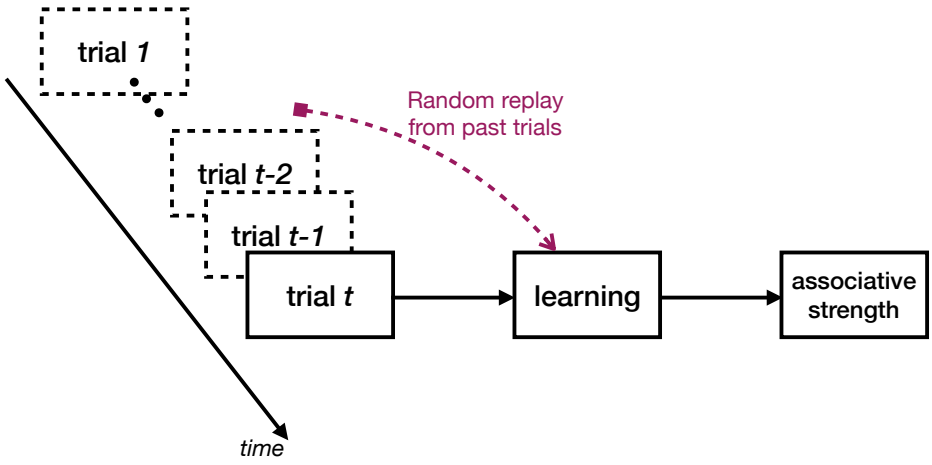


Figure 8. Schematic of the random replay model. The standard error-correction learning rule is depicted in solid arrows. In addition, the model assumes a memory of past trials, which are then randomly sampled and then replayed (dashed arrows). The replayed trials are treated like any other trial and are used to update associative strength through the standard error-correction learning rule.

4.3 Towards a Unifying Account of Classical Conditioning

Empirical studies of classical conditioning have provided a vast dataset to constrain theoretical models. To demonstrate the explanatory power of the random replay model, we simulate six well-established empirical findings that have been replicated across species or preparations (e.g., Rescorla, 2004; Schmajuk & Alonso, 2012; Schwarting & Busse, 2017): acquisition, spontaneous recovery, latent inhibition, retrospective revaluations, acquisition-extinction interval, and facilitatory lesion effects. These empirical findings suggest desirable models predictions based on the standard classical-conditioning paradigm (4.3.1, 4.3.2, 4.3.4, and 4.3.5), the timing of the trial distribution (4.3.3), and the effects of hippocampal lesions (4.3.6).

The qualitative model predictions are presented for the core phenomena and detailed quantitative fitting to the exact parameters of particular dataset is left to future work. Each of these six learning phenomena has had tailored explanations within associative learning framework, but no unifying theoretical model has been able to explain all of them. In the simulation below, we test the model predictions with one set of parameters: learning rate for normal trials $\alpha = .05$, learning rate for replayed trials $\alpha = .005$, number of replays $N_{replay} = 5$ (working memory size), memory capacity $N_{total} = 200$ (long-term memory size). To account for the behaviour of hippocampal-lesioned animals, we suggest two parameter changes: a lower overall memory capacity $N_{total}^{(lesion)} = 50$ but a higher learning rate for replayed trials $\alpha_{replay}^{(lesion)} = .05$.

4.3.1 Acquisition

During simple acquisition, a CS is repeatedly paired with a US. For instance, a tone might be repeatedly paired with food or a light might be repeatedly paired with a puff of air to the eye. Animals can learn to predict the US (e.g., food or puff of air) by making an appropriate anticipatory response to the CS (e.g., salivate or blink). This is the simplest learning

phenomena in classical conditioning and often acquisition trials are denoted as X+ where "X" indicates the presentation of the CS and "+" indicates the presentation of the US (see Table 6 for a complete list of trial types considered in the simulations).

Table 6.

All types of classical conditioning trials considered in this Chapter. CS1 and CS2 denotes two different conditioned stimuli. US denotes the unconditioned stimulus.

Trial Type	CS1	CS2	US
Null	absent	absent	absent
X+	present	absent	present
Y+	absent	present	present
XY+	present	present	present
X-	present	absent	absent

On each trial, animals observe X+ and incrementally update the associative strength of X based on the prediction errors (Equation 4.3). Given enough of these updates, animals, following Rescorla-Wagner's model, can learn to elicit an appropriate conditioned response to the CS. For the random replay model, however, additional replays take place between trials; in those replays, past trials (also X+ in acquisition) are sampled at random and replayed. The replay process effectively leads to further increments in the associative strength of X, and faster learning of the association between the CS and US is predicted. Though making distinct predictions about the speed of acquisition, it might be impossible to separate the random replay model from Rescorla-Wagner's model based solely on this behavioural data. As we shall see in the next section, however, the necessity of additional replay processing becomes much more apparent when more experimental variables are manipulated.

4.3.2 Spontaneous Recovery

Spontaneous recovery has long been seen at odds with the basic Rescorla-Wagner model and the widespread observations of this phenomenon suggests that an additional mechanism is needed beyond the error-correction principle (Rescorla, 2004; Sissons & Miller, 2009). In classical conditioning, the arranging of a positive relation between a CS and a US typically establishes a conditioned response to that CS (as in the standard acquisition trial). After animals have acquired the CS-US relation, during extinction training where the CS-US relation is removed (typically through presenting CS-alone trials), animals progressively cease responding. According to the Rescorla-Wagner model, the associative strength and thus conditioned response should be completely eliminated by the end of extinction. By simply introducing a delay after extinction, however, the extinguished response reappears when the same CS is represented to the animals, sometimes at nearly full strength — hence so-called spontaneous recovery (Napier, Macrae, & Kehoe, 1992; Kehoe, 2006). Moreover, the degree of recovery increases as the delay between the end of extinction and recovery test is increased (e.g., Haberlandt, Hamsher, & Kennedy, 1978).

Table 7.

Experimental procedures in classical conditioning

Phenomena	Phase 1	Phase 2	Phase 3
Spontaneous recovery	X+	X-	test X
Latent inhibition	X-	X+	test X
Backward blocking	XY+	X+	test Y
Recovery from overshadowing	XY+	X-	test Y
Recovery after blocking	X+	XY+	X-

The basic Rescorla-Wagner model predicts that the associative strength approaches 0 by the end of extinction. There is no chance that the model would predict a recovery while no further training is provided to the

animals during the delay between the end of extinction and the recovery test. The associative strength should remain at 0 because effectively animals observe "Null" trials during the delay. Within an associative framework, a number of explanations for spontaneous recovery have been proposed and most share a common intuition: the two conflicting experiences (i.e., X+ in acquisition and X- in extinction) should be temporally weighted differently with more recent experiences having a greater effect on learning. With the passage of time, however, the two sets of experiences become increasingly similar in recency (i.e., both are temporally more distant from the present), resulting in relatively more influence of X+ than before. The conditioned response thus partially returns, generating spontaneous recovery. This idea has been formally defined as (a) different decay rates for excitatory and inhibitory processes (Bouton, 1993), (b) a temporal weighting rule (Devenport, 1998; Devenport, Hill, Wilson, & Ogden, 1997), (c) diffusion of stimulus characteristics over time (Estes, 1955), and (d) inference of different latent causes for acquisition and extinction (Courville, Daw, & Touretzky, 2006; Courville & Daw, 2008; Gershman, Blei, & Niv, 2010; Gershman & Niv, 2012).

The random replay model can be thought of as a mechanistic account of these previously suggested models for spontaneous recovery. By the end of extinction, both X+ and X- trials are stored in the trial memory, but with more X- than X+ given the forgetting feature of the memory storage rule (random drop-out of existing trial memories). The delay period after extinction essentially creates a series of Null trials and will also be incorporated into the trial memory. As a result, the trial memory satisfies the intuition that (a) more recent experiences have greater influence on performance and (b) with an increase in delay, more Null trials will be added to the memory and hence the relative influence of distant trials (in this case X+) against recent rials (in this case X-) should also increase. In Figure 9 (blue lines), the random replay model is simulated to explain spontaneous recovery with parameter values mentioned above. The model predicts that a longer extinction-test delay should generate greater degree of recovery as in

the empirical study on the conditioned eye blink response in rabbits (Haberlandt et al., 1978).

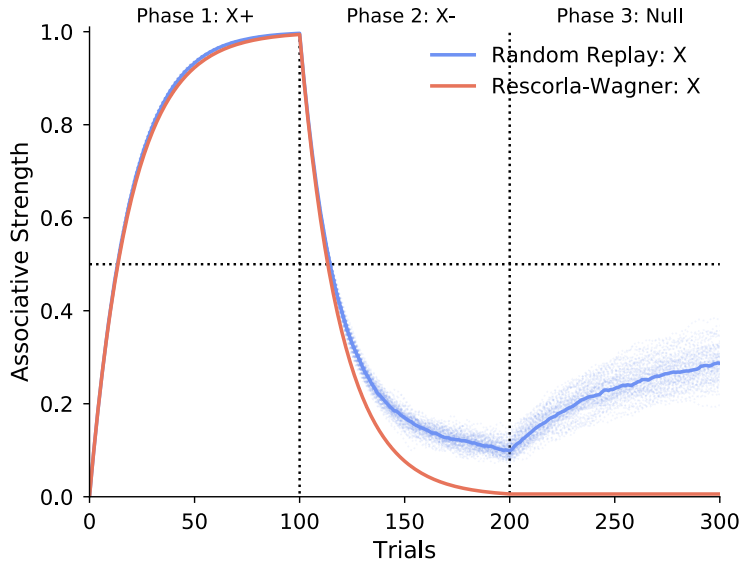


Figure 9. Spontaneous recovery predicted by the random replay model (blue), whereas the classic Rescorla-Wagner model (red) predicts no recovery in the third phase. Both models are repeatedly simulated for 100 times with exact same set of parameters. The median value of these simulated runs are depicted as solid lines.

4.3.3 Impact of Length and Timing of Training Phases

The random replay model makes further testable predictions about spontaneous recovery. For example, suppose a significantly longer extinction training phase where all X+ trials drop out of the trial memory, recovery should be muted in this situation because there are no X+ trials left in memory to be replayed. Similarly, when the acquisition trials are few, the number of X+ trials that are stored in memory before extinction is small, so even a shorter extinction phase would be expected to fully wash out the X+ from the trial memory. Therefore, the degree of spontaneous recovery should depend on the relative length of the acquisition and extinction trials.

This prediction matches what has been observed in fear conditioning (e.g., Rescorla, 2006; Laborda & Miller, 2013): massive extinction treatment reduces the return of fear.

Indeed, the magnitude of spontaneous recovery varies inversely with the acquisition-extinction interval: there is a greater recovery for a shorter interval between acquisition and extinction (Figure 10A: Rescorla, 2004). The experimental procedure can be divided into five phases. First, animals were presented with R_1+ (e.g., CS: white noise, US: food) until their conditioned responses reached a threshold. Second, animals received the same treatment with an alternative R_2 CS (e.g., light). If the animals were trained with the light first, then they received the white noise next (and vice versa). Third, animals exposed to an extinction contingency for both R_1 and R_2 (i.e., trials were either R_1- or R_2-). The presentation of R_1- and R_2- trials was counterbalanced. Fourth, a resting period was inserted (i.e., extinction-test interval as in the standard spontaneous recovery experiment). Finally, conditioned responses were recorded for both R_1 and R_2 after a fixed resting period for both CSs. The main finding was that recovery was greater for a stimulus trained closer in time to extinction that for one trained in the more distant past (Rescorla, 2004).

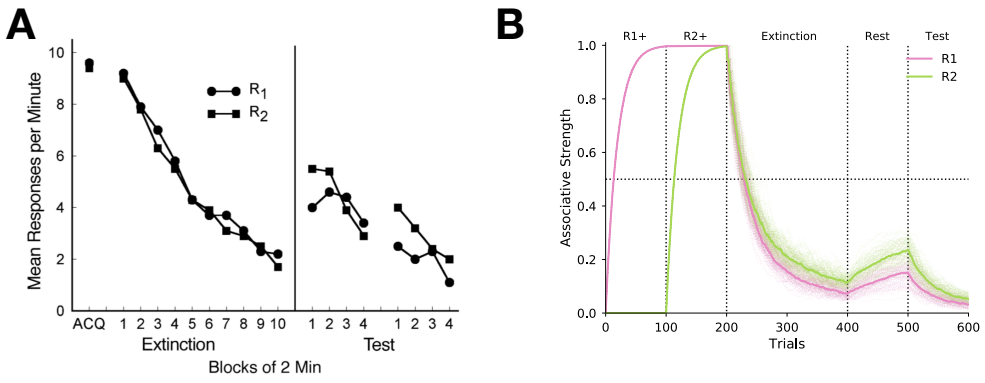


Figure 10. Shorter acquisition-extinction intervals produce a greater degree of spontaneous recovery. **(A)** Experimental data with pigeon subjects. Responses that had been trained distantly from (R_1: longer acquisition-

extinction interval) or proximally to (R_2: shorter acquisition-extinction interval) extinction. R_2 exhibits greater recovery than R_1 in a recovery test. The figure was adapted from Rescorla (2004). **(B)** Simulation of the random replay model. The Rescorla (2004) experiment has five phases (left to right: R_1+, R_2+, random mixture of R_1- and R_2-, rest, and test). The random replay model predicts that R_2 emerges to recover more than R_1.

Figure 10 shows how this finding follows naturally from the random replay model. Because the trial memory favours storing temporally more recent trials, animals are expected to have more remembered trials of R_2 than R_1. Random sampling from such a memory ensures that learning with replayed trials also prioritises recent trials. Following from the replay model's explanation for standard spontaneous recovery whereby the degree of recovery is positively related to the proportions of previous trials in the memory, R_2 (with more reinforced trials still in the trial memory) is thus predicted to have greater recovery than R_1. Stronger versions of trial spacing effects, in which duration of a trial and the duration of the inter-trial interval were manipulated (e.g., Balsam et al., 2010; Gallistel & Gibbon, 2000, 2002), can also be accounted by this simple replay mechanism (Ludvig et al., under review).

4.3.4 Latent Inhibition

In latent inhibition, initial exposures to unreinforced presentations of a stimulus (i.e., X-) reduces the speed of conditioning when that stimulus is later paired with a US (i.e., X+) (Lubow, 1973; Lubow & Moore, 1959). This finding poses yet another challenge to the Rescorla-Wagner model: the associative strength for X should remain at 0 during the initial X- trials when the animals expect no US from X and also receive no US (i.e., no prediction error). Later associative learning models postulated attentional mechanisms to explain latent inhibition, whereby animals learn not to pay attention to the stimulus when it has been repeatedly presented alone before (Mackintosh, 1974; Pearce & Hall, 1980).

From the random replay model's perspective, initial conditioning with X- trials should also be remembered by the animals. When in the second phase, X+ trials are presented, animals should also replay some of earlier X- trials, thereby slowing down conditioning to X. In Figure 11B, we depict one simulation of the replay model (blue lines for normal animals) with the latent-inhibition procedure. In the first phase, 100 unreinforced trials were presented to animals. As with the experimental data, the random replay model shows no change in associative strength in this stimulus-alone presentation; however, the trial memory gradually fills up with a number of X- trials. In the second phase, the animal experiences 100 reinforced presentation of the same stimulus (X+). Because the memory was dominated by X- before the start of the second phase, sampling X- trials from memory means that there was a negative prediction error (see Eq 4.4), and thus associative strength goes down. This results in a slower learning than if there had been no previous exposure to X- trials.

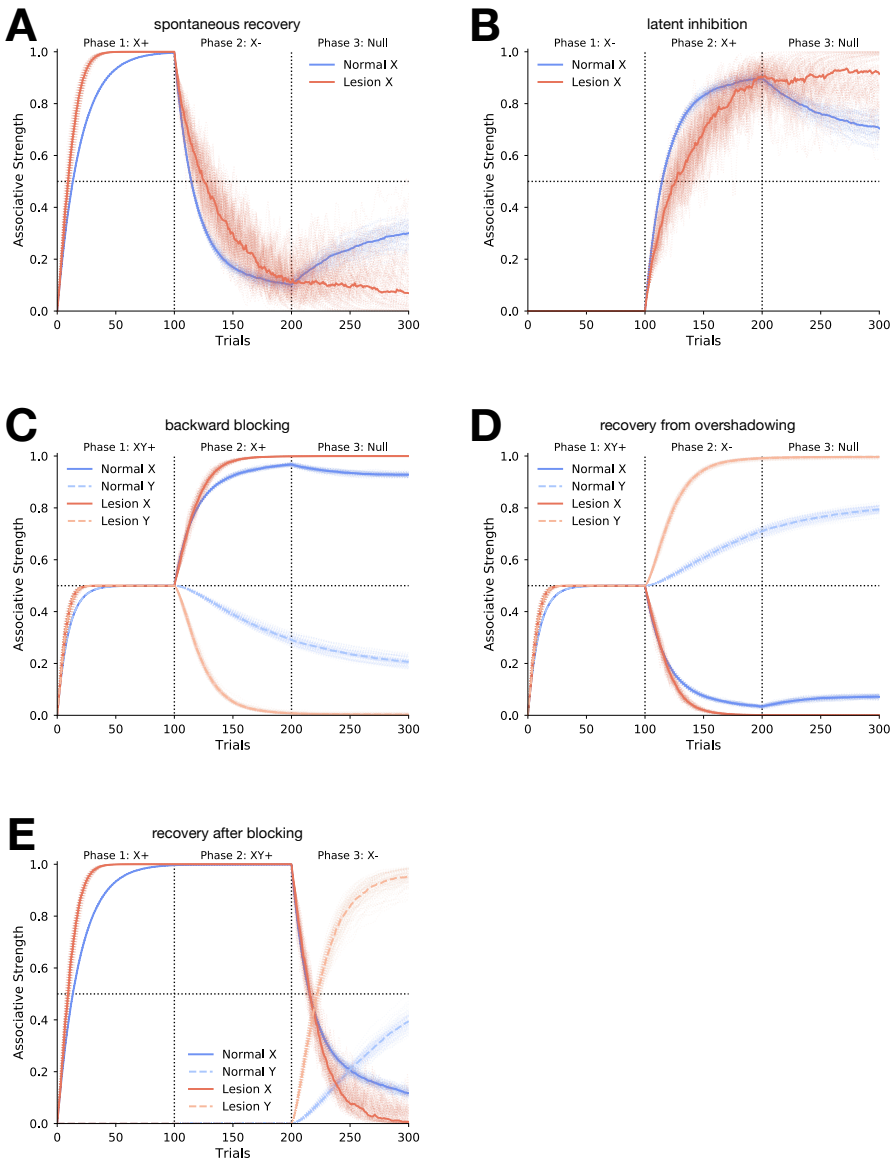


Figure 11. The random replay model of classical conditioning for both normal and hippocampal lesion animals. The learning phenomena were in figure title and the simulated experimental procedures were separated by dashed lines. **(A)** Spontaneous recovery. **(B)** Latent inhibition. **(C)** Backward blocking. **(D)** Recovery from overshadowing. **(E)** Recovery after blocking. All simulations presented here used the exact same set of model parameters in the main text and also for Figure 9 and 10 above.

4.3.5 Retrospective Revaluation

A third class of associative learning phenomena that can be reproduced by the random replay model is retrospective revaluation. These learning phenomena typically include two or more CSs (e.g., light and tone can be simultaneously presented within the same trial). In retrospective revaluation, animals change their responding to one CS on the basis of further trainings with a different CS (e.g., Dickinson, 1996; Van Hamme & Wasserman, 1994). We will investigate three instances of retrospective revaluations below: backward blocking, recovery from overshadowing, and recovery after blocking.

4.3.5.1 Backward Blocking

In backward blocking, animals are first presented with reinforced compound XY+ trials; for example, both tone and light are paired with food delivery. Then in the second phase, they are trained with one of the stimuli in isolation (e.g., X+) — i.e., only tone is paired with same food delivery. When tested after this second phase, the conditioned response for the other stimulus (e.g., Y) that was not presented in the second phase appears to become lower (Shanks, 1985; Miller & Matute, 1996). Though animals never re-encountered the Y stimulus, the training of X+ trials functions in the second phase as if it blocks the previously learnt associative strength of Y in the first phase; hence the name, backward blocking.

The key intuition behind many existing explanations of the phenomenon is that, during the first phase of XY+ training, animals not only learn that both X and Y predicts the US, but they also implicitly learn that X and Y are somehow related. This idea is crucial to understand many retrospective revaluation phenomena and has been instantiated in both latent-cause models (Courville & Daw, 2008; Gershman et al., 2010) and through within-compound associations in standard associative models (Dickinson, 1996; Van Hamme & Wasserman, 1994; Markman, 1989; Tassoni, 1995). According to these models, a representation of the absent stimulus (Y) becomes activated by presenting the other stimulus (X), because

the XY+ training results in (a) a generative model where two latent causes (e.g., X, Y) are necessary for a US or (b) within-compound associations are formed. Just as presenting Y- trials reduces the associative strength of Y, mentally activating a presentation of Y (through its associations with X) in the second stage of the XY+ X+ experiment should also reduce the associative strength of Y.

The Rescorla-Wagner model has no implicit or explicit associations between X and Y, and therefore it predicts no blocking for Y in the second phase. The random replay model, however, offers a simple mechanistic account for how the associations between X and Y were formed. After the first phase of training with XY+ trials, the trial memory will consist of many XY+ trials. Memories of these XY+ trials will be replayed during the second phase of training with X+ trials. Such replays result in the reduction of the associative strength of Y. To illustrate why this happens, we suppose an idealised training case where US has unit value and, upon the completion of XY+ training, animals learn that the associative strengths for X and Y are equal to 0.5. In the second phase, X is now paired with the same unit US, so the associative strength of X will gradually increase towards 1, through the standard error-correction learning rule. Meanwhile, when replaying XY+ trials, the sum of the associative strengths of X and Y are now greater than 1 (because Y is at 0.5 and X is greater than 0.5). This will produce negative prediction errors for both X and Y and decrease their associative strengths through the error-correction learning rule for replayed trials. However, given the fact that replayed learning rate is smaller than the normal learning rate, the overall trend for associative strength of X is to increase towards 1 and of Y is decrease towards 0, explaining the backward blocking phenomenon. In Figure 11C (blue lines), we simulate the random replay model and a decrease in associative strength of Y can be found while the training in the second phase only contains X+.

4.3.5.2 Recovery from Overshadowing

Recovery from overshadowing is very similar to backward blocking, except the second phase of training is altered from X+ (acquisition) to X- (extinction). That is, animals are first trained with XY+, which is then followed by extinction of one stimulus X-. During the extinction of X, however, the level of responding to the absent stimulus Y increases (Matzel, Schachtman, & Miller, 1985; Wasserman & Berglan, 1998), in an almost inverse case to backward blocking.

Figure 11D (blue lines) shows what the random replay model does in a recovery from overshadowing experiment. The exact same mechanism of trial memory and random replay is used in the simulation. The replay of XY+ trials during X- extinction generates a positive prediction error for both X and Y because, following the same illustrative situation above, $V(X) < 0.5$ and $V(X) + V(Y) < 1$ where the US was assumed to have a unit value. This effectively boosts the associative strengths of both X and Y, but the boost for X is offset by the ongoing extinction of X, especially with the higher learning rate for real versus replayed experiences.

4.3.5.3 Recovery after Blocking

A slightly more complicated experiment procedure is needed to elicit recovery after blocking. Initially, animals go through a standard blocking protocol, which includes two phases of training (X+ then XY+). This results in limited responding to the added stimulus Y; that is, the learning of Y is blocked by X because X alone is sufficient to predict the US. Then a third phase of X extinction training (X-) is introduced to animals. The 'surprise' empirical finding is that the level of responding to stimulus Y increases during this extinction training of X (Blaisdell, Gunther, & Miller,1999), hence the name, recovery after blocking.

Recovery after blocking is again quite readily accommodated by the random replay model (see Figure 11E). During the first phase of X+ training, the associative strength of stimulus X increases to asymptote, with no change in the strength of the yet-to-be-experienced stimulus Y. In the

second phase of XY+ training, no prediction errors are present because X alone can predict the US fully. Thus, there is no learning for both X and Y. So far, the Rescorla-Wagner model and the random replay model predict the exact same qualitative changes in the associative strengths of X and Y, though the random replay would expect the acquisition of X to be faster than the Rescorla-Wagner model would (as discussed above). The only difference is that the trial memory of the random replay model now contains a mixture of both X+ and XY+ trials. These additional memories and replays of X+ or XY+ do not change the associative strength and sustain the blocking effect.

The trial memory and replay process start to make distinct predictions from the Rescorla-Wagner model in the third phase, which consists of X- extinction training (see Table 7). The replay of XY+ during X- extinction training generates positive prediction errors for both X and Y, and therefore increase in the associative strength of the previously blocked stimulus (Y). In Figure 11E (blue lines), we depict a simulation of the random replay model using the same set of parameters across as elsewhere in this chapter.

4.3.6 Facilitatory Lesion Effects

The random replay model proposes a simple extension to the basic Rescorla-Wagner model by highlighting the importance of trial memory and random replay from such memory. Prevailing neuroscientific theories argue that hippocampus is critical to many functions of memory, and behavioural changes related to memory are often observed in animals with hippocampal destruction (e.g., Douglas, 1967; Cohen & Eichenbaum, 1993; Hirsh, 1974; Sutherland & Rudy, 1989; Ludvig, Sutton, Verbeek, & Kehoe, 2009).

In this section, we focus on the behavioural effects of hippocampal lesions in classical conditioning (see Schwarting & Busse, 2017 for a review). Schmaltz and Theios (1972) show that rabbits with their hippocampus removed show both faster acquisition of a conditioned nictitating membrane response and slower extinction than unoperated rabbits. Later studies

replicated the facilitatory effect of these hippocampal lesions in acquisition on different species (e.g., rats in Schmajuk & Isaacson, 1984) and on both appetitive and aversive conditioning (see Weiss & Disterhoft, 2015 for a review).

To account for the faster acquisition but slower extinction in hippocampal lesion animals, the random replay model has to assume two computational consequences of hippocampal lesion: (a) the capacity of the trial memory should be reduced, and (b) the learning rate for replayed trials should be enhanced. It is not obvious why the learning rate should be increased, but this adjustment may be justifiable as the main learning now take place in striatum for hippocampal-lesion animals (Bornstein & Daw, 2012). With these two adjustments on model parameter values, the random replay model can reproduce the facilitatory effect in acquisition and inhibitory effect in extinction for hippocampal-lesioned animals (Figure 11A, red lines). Simultaneously, the random replay model suggests a number of new predictions for possible behavioural changes for hippocampal lesion animals in latent inhibition and other retrospective revaluation experiments (Figure 11B to 11E, red lines). In latent inhibition, hippocampus lesioned animals are predicted to have slower acquisition in second phase and no recovery in the third phase. Facilitatory effects are expected in all retrospective revaluations for hippocampus lesioned animals. To the best of our knowledge, the literature lacks sophisticated empirical investigations of the performance of hippocampal lesion animals on most of these tasks.

4.4 Discussion

The random replay model is a formal model of associative learning. The model suggests simple extension to the Rescorla-Wagner model, by including an additional trial memory component and the randomly resampling trials from that memory to replay and learn from with the same error-correction learning rule. As demonstrated above, these extensions to

the error-correction learning models can capture a number of classical conditioning phenomena.

The idea of replay is inspired by lengthy neuroscience evidence of the behaviour of place cells in the hippocampus (e.g., Davidson et al., 2009; Redish, 2016; Gupta et al., 2010). These neurons seem to exhibit experience replay, mostly during the resting state of animals, in spatial navigation tasks (Gupta et al., 2010; Dave & Margolish, 2000; Gardner & Moser, 2017; Wu, Haggerty, Kemere, & Ji, 2017; Foster, 2017). Reusing past experiences to assist learning also has also been demonstrated to be useful in artificial intelligence. Deep learning agents with experience replay can achieve human-level performance on Atari video games (Mnih et al., 2015) and the ancient Chinese board game Go (Silver et al., 2016; Silver et al., 2017).

For simplicity, here, we only consider a very simple curation and retrieval mechanisms for the trial memory: A fixed size memory bucket that adds in the most recent trial and randomly drops out old trials when the bucket is full. Sampling past trials from the memory is also random. While simple, the trial memory and random replay mimic a mental model of the world presumed in model-based reinforcement learning algorithms (Daw, Niv, & Dayan, 2005; Balleine & O'Doherty, 2010; Daw et al., 2011; Vanseijen & Sutton, 2015). The mental model of the world typically has a full description of how the world works. In conditioning tasks, the world model summarises the transition function between states and reward function. The computational advantage of using model-based reinforcement learning is that a single model transition lumps together many real-world transitions and effectively reduces what needs to be kept in memory. The trial memory proposed here resembles a mental model of world because, as more actual trials are observed and curated, the more accurate the trial memory approaches to the true statistics of the world.

Replaying past trials at random, however, is clearly non-optimal because not all previous trial are relevant for the current learning objective. Other replay schemes have been previously suggested to further enhance the efficiency of learning such as prioritising certain trials to replay based on their absolute reward prediction errors (Lin, 1992; Liu & Zou, 2017; Zhang

& Sutton, 2017) or based on the expected value of improvements (Mattar & Daw, 2018). How different replay schemes affect learning and which scheme is adopted by animals are interesting questions to explore in the future.

The random replay model does not challenge the normative error-correction principle of learning. The model, however, advocates an augmented view of what are the training data used for learning: sampled memory from past is also a critical, but previously neglected, part of data that need to be considered through the error-correction rules. We showed that many phenomena observed in classical conditioning reveal traces of a mental process that draws samples from memory and actively reuses these samples for learning.

Chapter 5

Sample from Simulation

"There is a reflex which is still insufficiently appreciated and which can be termed the investigatory reflex. I sometimes call it the 'What-is-it?' reflex." (Ivan Pavlov, 1924)

5.1 Introduction

Through interactions with the environment, animals not only acquire value estimations for various states in the environment, but also get to know about how the environment works in general (e.g., state transitions and reward functions). The learnt transition and reward function comprise a *world model* which summarises the environment's dynamics. In this chapter, we will further demonstrate that humans and other animals behave as though they can draw samples from simulating such a world model. More importantly, this ability to sample from simulations (i.e., imagined mental samples), as we shall see below, can be used to potentially explain why we are curious creatures.

Humans and other animals have strong preferences for informative options. We are all highly curious creatures and will explore unknown options and even sometimes sacrifice rewards to resolve uncertainty earlier. When faced with delayed, uncertain rewards, humans and other animals usually prefer to know the eventual outcomes in advance. This search for information will even occur independent of any potential profit, when there is no possible effect on the delivery of primary rewards, as if consuming information itself was rewarding (e.g., Wyckoff, 1952; Prokasy, 1956; Spetch, Belke, Barnet, Dunn, & Pierce, 1990; Stagner & Zentall, 2010; Bromberg-Martin & Hikosaka, 2011; Blanchard, Hayden, & Bromberg-Martin, 2015; Iigaya, Story, Kurth-Nelson, Dolan, & Dayan, 2016). On occasion, this information seeking leads to seemingly suboptimal choices with animals preferring options with lower average reward rates (Spetch et al., 1990;

Stagner & Zentall, 2010). In this chapter, we develop a new computational model of this information-induced sub optimality based on the idea that animals' choices reflect the anticipated prediction errors from any upcoming cues in addition to the expected rewards.

5.2 Empirical Evidence for Information-Induced Sub-Optimality

The strong preferences for advanced information about rewards has been widely observed across species, including rats (Prokasy, 1956; Chow, Smith, Wilson, Zentall, & Beckmann, 2017), pigeons (Spetch et al., 1990; Zentall, Laude, Stagner, & Smith, 2015), starlings (Vasconcelos, Monteiro, & Kacelnik, 2015), monkeys (Bromberg-Martin & Hikosaka, 2009; 2011; Blanchard et al., 2015), and humans (Iigaya et al., 2016; Zhu, Xiang, & Ludvig, 2017). In some cases, animals even give up food or water for advanced information about impending rewards, even when these advanced signals do not change the rate of eventual reward delivery. For example, pigeons reliably choose an alternative that provides delayed access to food only 50% of the time over one that always provides the same amount of food with the same delay, but only when an immediate cue is provided, which tells the pigeons whether or not the food will eventually be available on that trial (Kendall, 1974; 1975; Spetch et al., 1990; Gipson et al., 2009). The choice of the 50% alternative over a 100% alternative is clearly suboptimal in terms of reward maximisation. Similarly, when choosing between delayed, probabilistic rewards, monkeys and humans will strongly prefer an option that informs them about the eventual outcome of that trial over one that leaves the resolution of uncertainty to the time of reward delivery (Bromberg-Martin & Hikosaka, 2009; 2011; Iigaya et al., 2016; Zhu et al., 2017).

Figure 12 presents a schematic that encapsulates many of the experiments that have been used to study this curiosity-like behaviour (e.g., Wyckoff, 1952; Green & Rachlin, 1977; Spetch et al., 1990; Roper &

Zentall, 1999; Stagner & Zentall, 2010; Bromberg-Martin & Hikosaka, 2009; 2011; Iigaya et al., 2016; Fortes, Vasconcelos, & Machado, 2016; Zhu et al., 2017). In these experiments, animals pick between two options with uncertain, delayed outcomes, where the cued option (red, top row in Figure 11) provides immediate cues as to the eventual outcome and the other, uncued option (blue, bottom) does not. For cued options, subjects receive a predictive cue after the initial-link (IL), which either perfectly predicts reward *R* after the terminal-link (TL) delay (the green S+ cue) or which perfectly predicts no reward with the same delay (the yellow S0 cue). The probability of the reward-predictive cues varies across experiments and is indicated here by *q*. For the uncued options, non-predictive stimuli (represented as the black S* cue) always appear after choosing the uncued option, leaving the animal in a state of uncertainty. A reward *R* then follows with a probability of *p* after the same delay of TL.

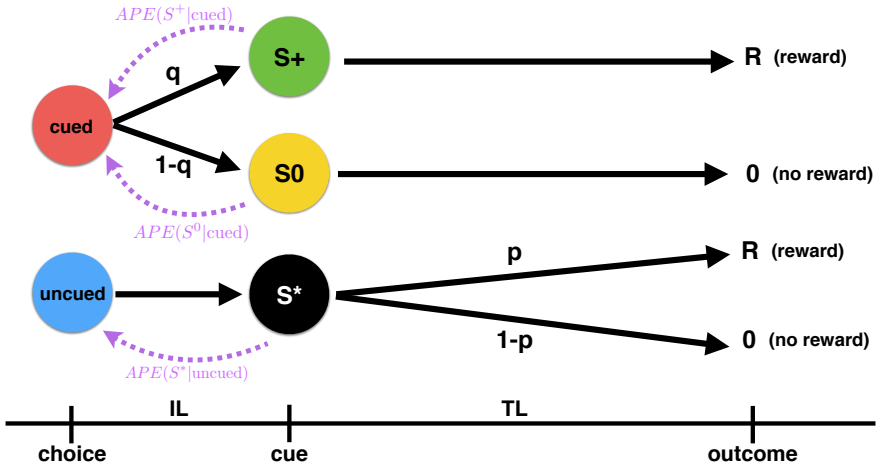


Figure 12. Formal representation of the information-choice task as a Markov Decision Process (MDP). Two offers (red and blue circles) are presented, and the animal must choose one of them. A cue then appears after this initial choice (Initial Link: IL), which is either informative (green S+ indicates a rewarding outcome; yellow S0 indicates a neutral outcome with probability *q* and *1-q* respectively) or uninformative (black S* leaving the animal in a state of uncertainty). Following a delay (Terminal Link: TL), the animal obtains

the outcome (reward or no reward). The anticipatory signals proposed by the APE model are illustrated as the purple dashed lines.

To apply computational models to these experiments, this information-choice task was represented as a Markov Decision Process (MDP). The agent starts in the choice state with two actions available (*cued* or *uncued*). The cued action leads to one of the cued states (either S+ or S0) stochastically, whereas the uncued action leads to the uncued state (S*). The uncued state only resolves its uncertainty at time IL+TL (i.e., the end of a trial), when it leads to reward with probability of p and no reward with probability of $1\text{-}p$.

5.3 Computational Models of Suboptimal Choices

To explain the patterns of information-induced suboptimal choice discussed above, we introduce a new model — the Anticipated Prediction Error (APE) model. We explore and evaluate the model in contrast with a baseline reinforcement-learning model: The Temporal-Difference (TD) learning model (Sutton & Barto, 1998). We start by outlining the standard TD model and then show how the APE model builds on and extends this basic framework.

5.3.1 The Temporal-Difference (TD) Model

As a baseline model, we consider the TD model of animal learning (Sutton & Barto, 1987; Moore, Desmond, Berthier, Blazis, Sutton, & Barto, 1986; Moore & Blazis, 1989; Moore, Choi, & Brunzell, 1998; Ludvig, Sutton, & Kehoe, 2012). The model suggests a normative account of animal learning, where animals are trying to form accurate *long-term predictions of rewards*. Animals following the TD learning rule are assumed to estimate a value function (V) for each state in their environment. Similar to the associative strength in the Rescorla-Wagner model discussed in Chapter 4,

the value of a state is an estimate of future rewards and is learned through an incremental update mechanism. At trial t, animal should update estimations for state values as follows:

$$V(s_t) = V(s_t) + \alpha \delta_t \tag{5.1}$$

where α is the learning rate, and δ_t is the temporal difference error:

$$\delta_t = R_t + \gamma V(s_{t+1}) - V(s_t) \tag{5.2}$$

where R_t is the immediate reward received upon transition from state s_t to s_{t+1}, and γ is a discount factor between 0 and 1.

The information-choice paradigm in Figure 12 is basically an *instrumental-conditioning* experiment where learning depends on the consequences of an animal's choices: the delivery of a reinforcing stimulus is contingent on what the animal does. These experiments can trace their history all the way back to Thorndike's experiments that motivated his famous Law of Effect, where cats progressively decreased their time to escape puzzle boxes with each successive attempt (Thorndike, 1898). In Thorndike's box, what to choose next based on previously learned knowledge was crucial to find a better policy to escape. Thus, the main difference in the instrumental case from our random replay model of classical conditioning (see the previous chapter) is that learning and control each have effects on the other: the learning system should guide choices, and choices will subsequently affect the observed data for the learning system.

For the MDP considered here (Figure 12), choice is only possible at one state C (the beginning of the trial) between progressing to either state C_{cued} or state C_{uncued}. In the modelling work below, for simplicity— selecting actions based on the state value, we assume a softmax action selection rule for all the models.

$$P(C_{cued}) = \frac{\exp[\beta \hat{V}(C_{cued})]}{\exp[\beta \hat{V}(C_{cued})] + \exp[\beta \hat{V}(C_{uncued})]} \tag{5.3}$$

where the probability of choosing the cued option, $P(C_{cued})$, is dependent on the difference in decision value, $\hat{V}(C_{cued}) - \hat{V}(C_{uncued})$; and β is the inverse

temperature parameter that controls the degree of randomness of action selection; with smaller β, animals select actions more randomly.

Note that we deliberately use two different notations for state values: $V(s)$ and $\hat{V}(s')$. Both state values are learned through experiences, whereas the former is used to form accurate predictions of long-term rewards, and the latter is used to generate actions. The benefit of this distinction will be clearer when we introduce the APE model. For the TD model, $V(s)$ is always equal to $\hat{V}(s)$.

5.3.2 The Anticipated Prediction Error (APE) Model

The key intuition behind the APE model is that animals both learn from their experiences as per the TD rule above, but also enhance that experience with samples of potential future outcomes at the time of choice (i.e., sample from simulations of imagined future episodes). From those anticipated samples, they calculate a prediction error, which is then used to adjust the value learned through the TD rule. Importantly, this sampling process can lead to biased learning values that do not necessarily accord with the encountered experiences. To explain the information-induced suboptimal choices, it is necessary to assume that there is an optimism bias for rewards (Sharot, Riccardi, Raio, & Phelps, 2007; Sharot, 2011), or more generally, that animals are more likely to sample from both big wins and big losses (Ludvig, Madan, & Spetch, 2014; Madan, Ludvig, & Spetch, 2014; Lieder et al., 2018).

More formally, the new model proposed here supposes that information-seeking behaviours are due to a forward-sampling process, whereby animals draw mental samples of possible future states, and then calculate any prediction errors for that sample. The mean value of these prediction errors is defined as the APE:

$$APE(s' \mid s) = T_{ss'} \times [R_{ss'} + \gamma^{D_{ss'}} V(s') - V(s)] \tag{5.4}$$

where $T_{ss'}$ is the transition probability from state s to subsequent s', $R_{ss'}$ and $D_{ss'}$ are the immediate reward received upon and the time needed for the transition respectively, and γ denotes the discount factor. These APEs are

hypothesised to reflect the degree of anticipation for a future imagined state s' given the current state s.

The key assumption of the model is that these anticipated prediction errors do not just drive learning, but are rewarding in and of themselves (see McDevitt et al., 2016). Positive APEs are thus reinforcing and negative APEs punishing. Therefore, when animals make choices, they consider both the value function as learned through the TD model as well as the APEs derived through forward sampling. Accordingly, we can define the decision value $\hat{V}(s)$ as the weighted sum of the APEs for all possible future states plus the value function of state s:

$$\hat{V}(s) = V(s) + \sum_{s_i \in F(s)} w_i APE(s_i | s) \tag{5.5}$$

where $F(s)$ is the set of all possible future states that are reachable from state s, w_i is the weight associated with the future state s_i, and reflects the sampling bias toward that future trajectory. For simplicity, we only consider $F(s)$ as the immediate next state from state s; that is, animals were assumed to only engage in one-step look ahead into future.

For the APE model, the decision value is different from the value function because the inclusion of the anticipatory sampling process. Besides that, the same softmax choice rule was applied to generate actions. It is also worth noting that the APE model can be viewed as an extension of TD model. The APE model argues for separation of learning (as in the value function) and control (as in the decision value) through an additional anticipatory sampling process (represented by the sampling weights). However, when there is no anticipatory sampling (i.e., sampling weights are equal to 0), the APE model reduces to TD model; this also means that the values driving learning and control are no longer separated. In the simulation below, we treat the TD model as a special case of the APE model where sampling weights are 0.

5.4 Explaining Suboptimal Choices with the APE model

In this section, we characterise the behaviours of both the TD and APE models based on a series of information-choice tasks. We broadly classified the task, based on the primary experimental variables in the test, into five applications: (a) cue-outcome contingency, (b) uncertainty resolution, (c) delay to outcomes, (d) reward magnitudes, and (e) negative outcomes (see Table 8). Alongside each application, we present results from both the TD and APE models.

Table 8.

Key experimental variables that have been found to determine the degree of suboptimal choice in information-choice tasks.

Experimental Variables	Papers	Species
Cue-outcome contingency	Stagner & Zentall (2010)	Pigeon
	Kendall (1974; 1985)	Pigeon
	Roper & Zentall (1999)	Pigeon
	Gipson et al. (2009)	Pigeon
	Fortes et al. (2016)	Pigeon
Uncertainty resolution	Green & Rachlin (1977)	Pigeon
	Dunn & Spetch (1990)	Pigeon
	Bromberg-Martin & Hikosaka (2009; 2011)	Monkey
	Vasconcelos et al. (2015)	Starling
Delay to outcomes	Spetch et al. (1990)	Piegon
	Iigaya et al. (2016)	Human
Reward magnitudes	Blanchard et al. (2015)	Monkey
	Bennett et al. (2016)	Human
Negative outcomes	Zhu et al. (2017)	Human
	Charpentier, Bromberg-Martin, & Sharot (2018)	Human

5.4.1 Cue-outcome Contingency

Advance information is generally valuable. When such information is earned by trading against primary rewards (e.g., food and water), an important question is how much of this primary reward, if any, would creatures be willing to sacrifice for advance information. Using a series of information-choice tasks (see Figure 14A), Stagner and Zentall (2010) found that pigeons learn to trade, at least, 30% of their potential reinforcement rate for "useless" advance information. In their task, pigeons repeatedly chose between a 20% reinforcement gamble (that provided advance information about the eventual outcome) and a 50% reinforcement gamble (no advance information) that gave reward after a delay. The advance information was provided by coloured cues. In the informative case, the cue colour indicated whether or not reward would be available at the end of the trial, whereas in the informative case the coloured cues could lead to either reward or no reward. Figure 13 shows how the experiment had 4 phases: training, reversal, discrimination, and non-discrimination. Throughout this experiment, the cue-outcome contingencies were manipulated but a consistent preference for informative cues was observed (Figure 14A). Indeed, such information that cannot change the subsequent outcomes is often called non-instrumental information.

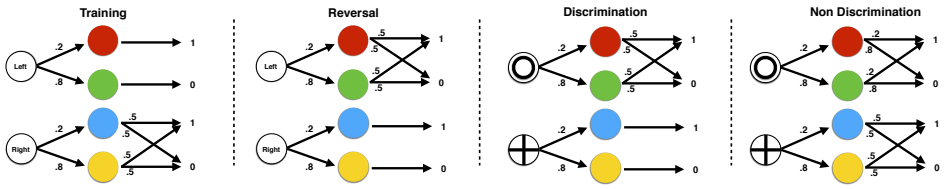


Figure 13. Schematic illustration of the four-phase information-seeking task in Stagner & Zentall (2010). In the *Training* phase, pigeons learned to prefer the cued option (depicted as "Left"). Then the cue-outcome contingency was reversed in the *Reversal* phase, and pigeons still learned to prefer the cued option (now "Right"). In the third *Discrimination* phase, novel choice stimuli ("Circle" and "Plus" shapes) were introduced while keeping the same cue-outcome contingency. The shapes were counterbalanced across pigeons, and they again learned to prefer the cued option ("Plus" in this case). In the final

Non-Discrimination phase, choosing either option should not provide valuable advance information, and pigeons learned to prefer the one with the higher expected value ("Plus" in this case).

In this case, the pigeons gave up the potential for a 2.5x higher reinforcement rate for a piece of "useless" information that just revealed the eventual outcomes 10 seconds early, but did not alter them. This strategy has been interpreted as potentially optimal for birds in nature in that these extra few seconds can be spent on other tasks or even to just hide safely from a predator (e.g., Vasconcelos, Monteiro, & Kacelnik, 2015).

We first test whether the TD learning rule is sufficient to explain the observed preference for suboptimal choices. The TD model assumes no additional memory or computation from the animals, except value functions for the individual states. Figure 14B (top panel) shows how the TD model selects the option with the higher reinforcement rate with enough training, which is the optimal behaviour, but the exact opposite of what pigeons do. This insensitivity to advance information of the TD model is also reported in Bromberg-Martin and Hikosaka (2009; 2011).

Clearly, pigeons are sensitive to the advance information and learn to pick the option with advance information (as in the first three phases of the experiment). Simultaneously, however, when advance information is removed from consideration, the same pigeons do learn to select the higher reinforcement rate similar to the TD model (as in the final phase). How could TD model explain the pigeons' behaviour so well without advance information and so badly with that information?

We now examine a simple extension of the TD model, the APE model, that further assumes an anticipatory sampling process for animals which can simulate samples of future episodes from their model of world. For example, in the *Training* phase of Stagner and Zentall (2010)'s experiment (Figure 14), when pigeons evaluate the value of "Left" option, they should generate samples for what will happen next after choosing the "Left". Choosing "Left" could lead to the "red" or "green" states with 20%

and 80% probability respectively. To illustrate, suppose that animals do not discount future rewards (i.e., $\gamma = 1$) and have trained long enough (i.e., animals have the correct value functions of states, transition and reward functions of tasks), thus the value for "red" and "green" should be:

$$V(\text{red}) = 1$$
$$V(\text{green}) = 0 \tag{5.6}$$

On average, the anticipated prediction errors per sample should be:

$$APE(\text{red}|\text{Left}) = T(\text{red}|\text{Left}) \times [V(\text{red}) - V(\text{Left})]$$
$$APE(\text{green}|\text{Left}) = T(\text{green}|\text{Left}) \times [V(\text{green}) - V(\text{Left})] \tag{5.7}$$

where $T(s'|s)$ is the transition probability from state s to s'. Because animals have the right model of world, the transition function is correctly represented in the brain. For simplicity, we do not consider details of how animals learn a world model from experience. In the simulations below, we simply assume that, after K trials, a correct world model including transition and reward functions should be learnt by animals where $K \sim N(\mu = 50, \sigma = 10)$.

On average, the anticipatory sampling process generates samples in proportion to w^{red} and w^{green} respectively for the "red" and "green" states. Hence, the total amount of APEs should be the sum of all the samples' APEs:

$$w^{\text{red}} APE(\text{red}|\text{Left}) + w^{\text{green}} APE(\text{green}|\text{Left}). \tag{5.8}$$

To make the difference clear, $T(s'|s)$ represents the transition function, which is assumed to be determined by the task, and w denotes the sampling weights, which are determined by the internal processes. Pigeons then combine knowledge from both the anticipatory sampling process and the standard value function to construct the decision value for choosing "Left":

$$\hat{V}(\text{Left}) = V(\text{Left}) + w^{\text{red}} APE(\text{red}|\text{Left}) + w^{\text{green}} APE(\text{green}|\text{Left}) \tag{5.9}$$

Similarly, for the other option (i.e., choosing "Right"), the decision value based on the APE model should be:

$$\hat{V}(\text{Right}) = V(\text{Right}) + w^{\text{blue}} APE(\text{blue}\,|\,\text{Right}) + w^{\text{yellow}} APE(\text{yellow}\,|\,\text{Right}) \tag{5.10}$$

Interestingly, in this idealised scenario (i.e., pigeons have the correct model of the MDP), both $APE(\text{blue}\,|\,\text{Right})$ and $APE(\text{yellow}\,|\,\text{Right})$ are equal to 0 because $V(\text{Right}) = V(\text{blue}) = V(\text{yellow}) = .5$. This is a unique feature of the information-choice task that the absence of advance information also means there is no rapid change in the value function. Therefore, the decision value of choosing "Right" is effectively identical to its value function:

$$\hat{V}(\text{Right}) = V(\text{Right}) \tag{5.11}$$

Given that, the choice probability of the option with advance information (in the Training phase it is "Left") is determined by the differences in decision values between "Left" and "Right". Here, the preference for choosing the informative option ("Left") over choosing the non-informative option ("Right") is solely driven by the net sampling weights of good news (w^{red}) and bad news (w^{green}). The APE model predicts that sampling more positive future states leads to information-seeking behaviour; conversely, sampling more negative future states leads to information avoidance behaviour. In the information-choice task of Stagner & Zentall (2010), the net sampling weights between good news and bad news determines the choice preference:

$$\Delta w = w^{\text{red}} - w^{\text{green}} \tag{5.12}$$

In Figure 14C, we show that APE model can reproduce the changes in choice probability in response to changes in the cue-outcome contingency.

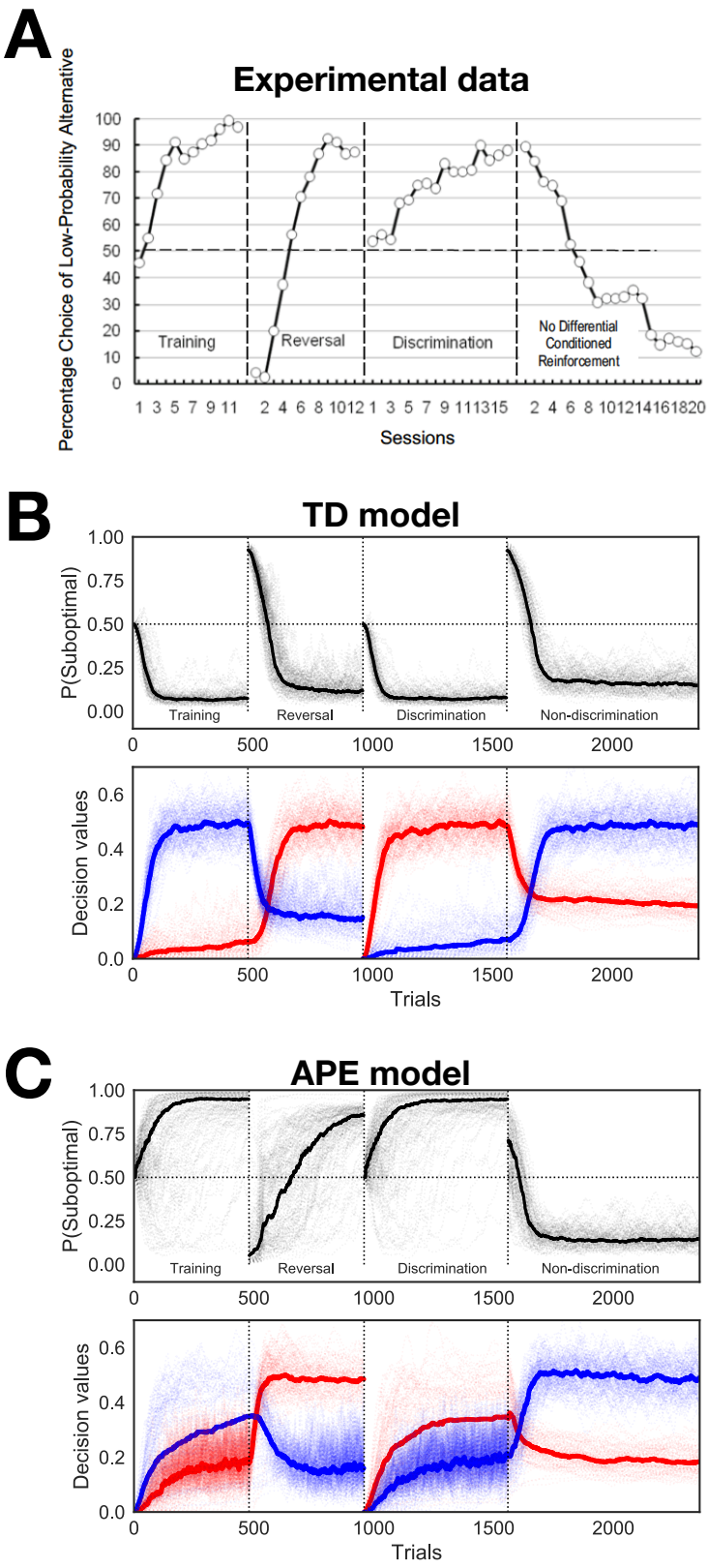


Figure 14. **(A)** Behavioural data from the four-phase task of Stagner & Zentall (2010). Strong preferences for advance information and a lower reinforcement rate option emerged through experience in the Training,

Reversal, and Discrimination phases. When the advance information is absent (Non-discrimination phase), pigeons learnt to choose the option with the higher reinforcement rate. The figure was adapted from Stagner & Zentall (2010). **(B)** The TD model predicts no preference for advance information. Value functions were initialised at 0. On the first trial in the Reversal phase, cue values were reset to 0 to account for changes in cue-outcome contingency. On the first trial in the Discrimination phase, value functions were agin reset to 0 for the new choice context. **(C)** The APE model can capture the dynamics of the choice probability for the suboptimal option with advance information. The same simulated procedure was used as for the TD model. We set the learning rate, inverse temperature in the softmax choice rule, and the discount factor at $\alpha = .06$, $\beta = 6$, $\gamma = .98$ for both models. The APE model has an additional parameter: the sampling bias for good news, which was set at $\Delta w = 4$. Both the TD and APE models were repeatedly simulated with the same set of parameters 100 times, and the solid lines denote the median of individual simulated runs (dashed lines).

5.4.2 Uncertainty Resolution

In Stagner & Zentall (2010)'s version of the information-choice task, the expected reward rate was different between informative (20% reward) and the non-informative options (50% reward). Though strong information-induced suboptimality was observed, the experimental design inevitably confounded the expected reward and the amount of resolved uncertainty associated with the advance information. To isolate the effect of uncertainty resolution, an earlier series of experiments considered a type of information-choice task where the expected rewards for both the informative and non-informative options were identical (e.g., Green & Rachlin, 1977; Bromberg-Martin & Hikosaka, 2011).

We focus on the Green & Rachlin (1977) task (Figure 15A) because it provides a broad manipulation of the degree of uncertainty being resolved by the advance information. In their task, pigeons were trained to choose

between two options that led to different degrees of uncertainty resolution (i.e., probabilities of reward), but the expected rewards for both options remained same throughout the experiment. The benefit of choosing the informative option (i.e., the cued option) was that pigeons were informed about the eventual outcome, through signalled light colours, 30 seconds earlier, thereby eliminating any uncertainty. On the other hand, choosing the non-informative option provided one non-discriminative colour, and pigeons remained uncertain about the outcome for the same 30 s until food delivery.

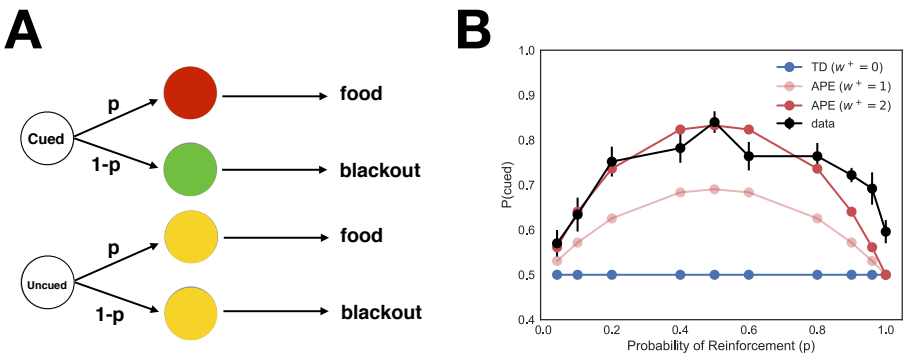


Figure 15. The effect of uncertainty reduction on choice of the informative option. **(A)** Illustration of the experimental procedure used in Green & Rachlin (1977) to study pigeons' preference for uncertainty reduction. Both options had the same probability of reinforcement (p), but after choosing the cued option, pigeons were informed of the eventual outcome immediately from the colour signals. Choosing the uncued option, however, left pigeons in a state of uncertainty until the end of trial (30 s later). The tested values of p varied across the range of 4%, 10%, 20%, 40%, 50%, 60%, 80%, 90%, 96%, and 100%. **(B)** Experimental data from Green & Rachlin (1977) are in black, and error bars indicate ±SEM. The simulations of the TD model (blue) and APE model (shades of red for different values of the w^+ parameter). The softmax inverse temperature and discount factor were set at $\beta = 6$, $\gamma = .98$. The APE model can reproduce the quadratic relationships observed in data.

Figure 15B (black line) shows how the experimental results suggest a quadratic relationship between choice probability of the cued option and the probability of reinforcement (p). From an information-theoretical perspective, the amount of uncertainty can be quantified as the Shannon entropy:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log P(x_i) \tag{5.13}$$

where X denotes all n possible future outcomes of choosing the option. The concept of entropy quantifies our intuition in guessing Heads or Tails of a coin toss. We are most uncertain about the outcome when it is a fair coin and become gradually more certain when the coin is biased toward either Heads or Tails. In this task, the entropy also exhibits a quadratic relationship with the probability of reinforcement (by analogy, one can imagine that Reward can be coded as Heads, and No Reward can be coded as Tails), which peaks when $p = .5$ and decreases when p is away from .5. Using the amount of Shannon entropy reduction as an information bonus has even been formally implemented in a previous extension of the TD model (e.g., Bennett, Bode, Brydevail, Warren, & Murawski, 2016). Though the Shannon entropy fits this part of data, we will see later that this idea lacks explanatory power in information-choice tasks concerning delay to outcome, reward magnitude, and negative outcomes. The reasons are that Shannon entropy does not vary with time or reward and is always non-negative.

Figure 15B (red lines) also shows how the APE model can predict the quadratic relationship because the anticipated prediction errors are proportional to $p(1-p)$. To illustrate, we consider the idealised condition again (where pigeons have the correct model of the task). Pigeons should learn that the value for the cued option and the uncued option are the same: $V(\text{cued}) = V(\text{uncued}) = p$, assuming that food has unit value. In addition, for the task depicted in Figure 15A, $V(\text{red}) = 1$, $V(\text{green}) = 0$, and $V(\text{yellow}) = p$. For example, according to Equation 5.4, the size of the anticipatory prediction error for the good news (red) would be calculated as:

$$\begin{aligned} APE(\text{red}\,|\,\text{cued}) &= p \times [V(\text{red}) - V(\text{cued})] \\ &= p(1-p) \end{aligned} \tag{5.14}$$

Similarly, we calculate the amount of APE for bad news (green) as

$$\begin{aligned} APE(\text{green}\,|\,\text{cued}) &= (1-p)\times[V(\text{green})-V(\text{cued})] \\ &= -(1-p)p \end{aligned} \tag{5.15}$$

This example calculation indicates that APE for future states in Green & Rachlin (1977)'s task should be quadratic with p, resulting in choice preference for the APE model that are also quadratic with the probability of reinforcement.

5.4.3 Delay to Outcome

A third well-documented empirical finding in the information-choice task is that the degree of preference for the informative option increases with longer delays to the eventual outcome (i.e., longer TL in Figure 16). Longer delays to the outcomes mean that any advance information provides an even earlier resolution of the uncertainty. Intuitively, if the amount of uncertainty resolved by the information is constant, earlier reception of the information should be more valuable than later. As shown in Figure 16 (Left), both pigeons (Spetch et al., 1990) and humans (Iigaya et al., 2016) exhibit significant modulation of their preference for informative options with different delays to the reward.

Note that the unique property of the anticipatory sampling process is that it requires animals to conduct mental time travel to future states (Clayton & Dickinson, 1998; Roberts, 2002; Clayton, Bussey, & Dickinson, 2003; Zentall, 2005; Schacter, Addis, & Buckner, 2007; Roberts, 2014)[11]. In the information-choice task, such an anticipatory sampling process is more likely to draw a future cue state (e.g., good news, bad news, or neutral news) when the duration of that cue presentation is longer. For example, the probability of drawing a future sample, which happens to be good news, should be proportional to the duration of the presentation of the good news

[11] There is a debate on whether non-human animals are capable of mental time travel. The Bischof–Köhler hypothesis argues for human uniqueness in the ability to travel mentally in time; other animals cannot anticipate future needs or drive states and are bound to a present that is defined by their current motivational state (Köhler, 1917; Bischof, 1978; Suddendorf & Corballis, 1997).

signal — in this case, the duration is equal to the terminal link (TL) or delay from choice to reward. To formally describe this property, we further constrain the sampling weights to be a linear function of TL:

$$w = g \times TL \tag{5.16}$$

where g is a gain factor that represents the degree of influence by the TL duration on sampling weights. Similarly, the behaviour of the APE model in this task should be driven by the difference in this gain factors for the good news and bad news. Here, we set the gain factor for good news and bad news to $g^+ = .1,\ \ g^- = 0$. The TD model becomes a special case of the APE model when all gain factors are equal to 0.

In Figure 16, we present model simulations of both the TD and APE model. The human study provides more diagnostic data (Iigaya et al., 2016) where the expected reward for both the informative and non-informative option are the same. The TD model always predicts indifference in choices across any length of the TL, whereas the APE model predicts an increase in preference for the informative option with longer TLs.

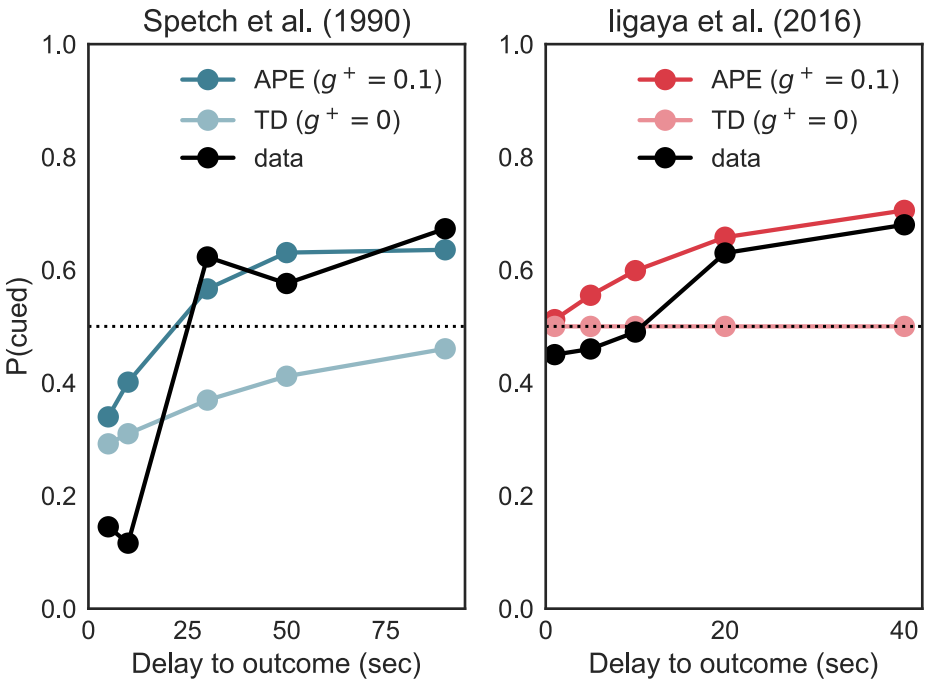


Figure 16. Delay to outcome manipulation in the information-choice task. **(Left)** The pigeon study found that longer delays induce greater preference for the cued option (Spetch et al., 1990). The experiment contained a cued option with a 50% reinforcement rate and an uncued option with a 100% reinforcement rate. The duration of the TL was varied across 5, 10, 30, 50, and 90 seconds. **(Right)** The human study found a similar pattern (Iigaya et al., 2016). Both the cued and uncued option had a 50% reinforcement rate. The TD model fails to reproduce the increase in preference for cued options with an increase in TL, whereas the APE model successfully captures this relationship. We set the inverse temperature in softmax and the discount factor $\beta = 2$, $\gamma = .98$ for both models. The APE model has an additional parameter $g^+ = .1$.

5.4.4 Reward Magnitude

Suppose we purchased two lotteries, which both have a 50% chance of winning something and a 50% chance of winning nothing. One lottery provides a potential reward of $100, and the other provides a potential rewards of $1, and the outcome can only be known in a month from now. Fortunately, we are given the opportunity to find out right now the outcome of one of the lotteries. Which lottery would you prefer to know about? Intuitively, the advance information that would resolve the uncertainty of the lottery with the higher payoff would seem to be more valuable. However, this intuition is at odds with information theory as both lotteries have the same amount of entropy — hence one should really be indifferent if strictly following information theory.

Using a similar information-choice task paradigm, Blanchard et al. (2015) formally tested this intuition and directly probed how animals choose between information with the same amount of uncertainty resolved but with different upcoming payoffs. Figure 17A shows how the overall task structure was similar to the standard information-choice task in Figure 12 with only the reward magnitudes manipulated. Monkeys were asked to choose between a cued option and an uncued option. Choosing the cued option resulted in either good news or bad news with a 50/50 chance. The good news not only indicated an upcoming water reward, but also the amount of water to come (represented by the height of white bar in the experiment). Choosing the uncued option resulted in random states which contained no further information, and the monkey only learned the outcome at the end of trial. The amount of juice delivered to monkeys ranged from 75 to 375μL in 15μL increments.

Figure 17B shows the behavioural results of two monkeys as a psychometric curve of the choice probability of the cued option against the differences in reward magnitudes (i.e., the amount of liquid) between the cued and uncued options. If advanced information was worth nothing to the animals (i.e., no value), then the animals should be indifferent between two options when the difference in water rewards was 0. Indeed, this is the prediction of the TD model as shown in Figure 17C (black line). The

empirical psychometric curves, however, were shifted leftwards for both monkeys, suggesting that monkeys were willing to sacrifice some amount of water in exchange for learning the eventual outcome sooner. This phenomena can be reproduced by the APE model (Figure 17C blue and red lines) because the anticipated prediction errors naturally grow with the reward magnitudes (see Equation 5.4).

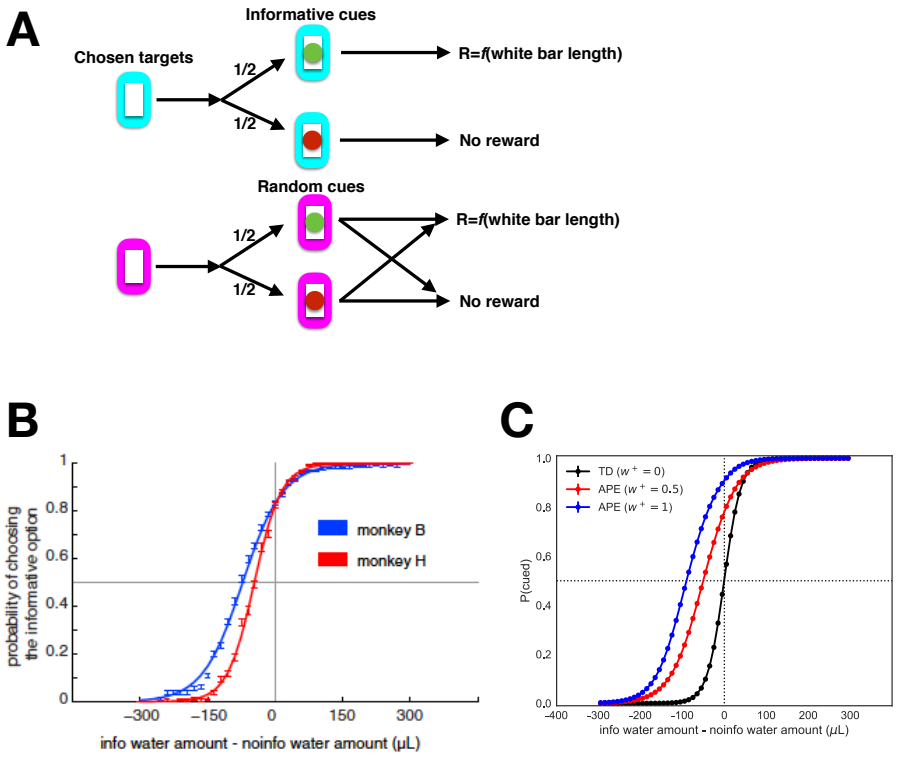


Figure 17. Reward magnitude manipulation. **(A)** An illustration of the experimental procedure of the monkey study reported in Blanchard et al. (2015). On each trial, monkeys were presented with two offers in sequence, each followed by a dark screen period (order is counterbalanced). Then they had to choose between a cued offer (cyan bar) and an uncued offer (magenta bar). The height of the central white bar indicated the amount of liquid potentially available on that trial, and the green and red dots revealed whether the risky option won or lost respectively. The probability of reinforcement for both options was 50% throughout experiment. After a 2.25s cue presentation, the monkeys received outcome delivery. **(B)**

Behavioural results. Preference for the cued option as a function of the liquid amount difference between the cued and uncued options. Error bars indicate ±SEM. The figure is adapted from Blanchard et al. (2015). **(C)** Predictions of the TD and APE model. We set the inverse temperature and discount factor as $\beta = 6$, $\gamma = .98$ for both models. The APE model has an additional sample weightings parameter as shown in the figure legends.

5.4.5 Negative Outcomes

Though the standard economic analysis suggests that information is valuable only to the extent that it can lead to better decisions, humans and other animals should just ignore the non-instrumental information which has no prospect of improving their decision making. As reviewed above, the suboptimal-choice literature provides a wealth of empirical data that humans and other animals still seek out non-instrumental information, even at the cost of primary rewards. However, animals, in many other situations, actually avoid the non-instrumental information, once again without any strategic rationale (Jenkins & Boakes, 1973; Golman, Hagmann, & Loewenstein, 2017).

To further examine the learning curve of information-preference behaviour, we conducted an information-choice task with humans (Zhu et al., 2017). Figure 18A illustrates the basic design. Participants were instructed to choose between two options. Just as in the standard information-choice task, after choosing the "Find Out Now" option, informative cues were presented as animal symbols, and participants could infer the upcoming outcomes by which symbol appeared. By choosing the other "Keep It Secret" option, participants always encountered the same animal symbol, leaving them in a state of uncertainty. To mimic the information-choice task used with other animals, we deliberately set the outcomes to be only consumable immediately. Hence, unlike monetary rewards used to incentive human participants, they were rewarded or punished by images (Crockett et al., 2013; Iigaya et al., 2016). Positive

images were erotic, negative images were aversive (e.g., gruesome pictures), and the neutral image was a circle.

We consider three conditions for the task with a similar procedure, but different outcomes. The *Good* condition involved positive and neutral images, the *Bad* condition involved negative and neutral images, and the *Mix* condition involved positive and negative images. These outcomes were always delivered with 50/50 odds on each trial.

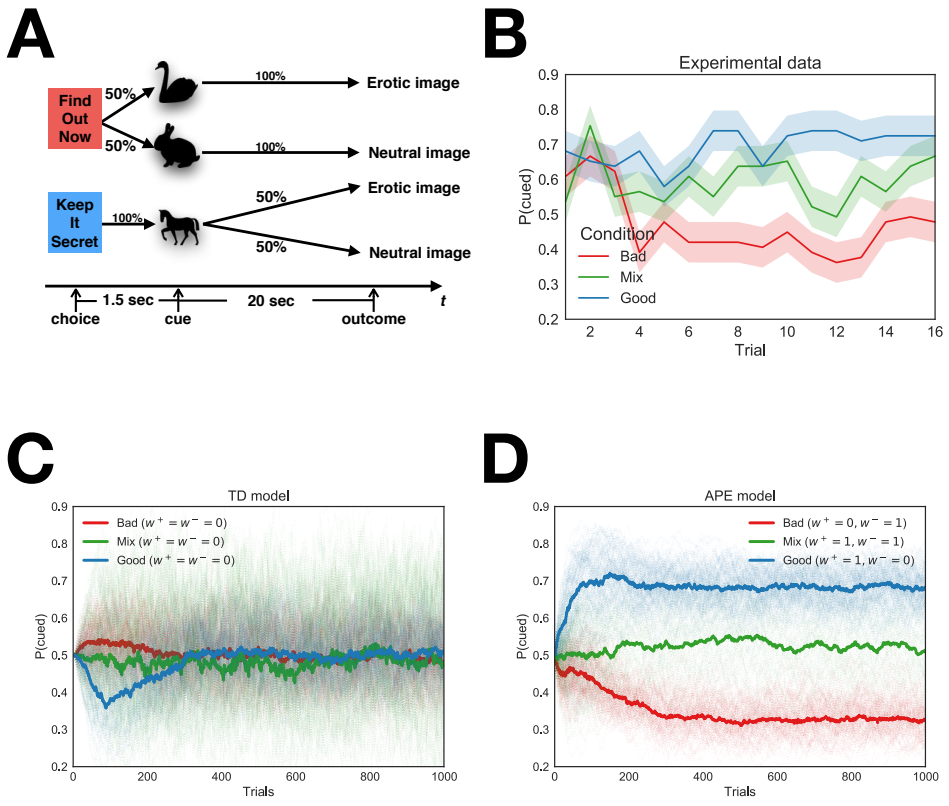


Figure 18. People preferred advance information, but less so when aversive outcomes were included in the gamble. **(A)** Experimental procedure of the human information-choice task reported in Zhu et al. (2017). Participants chose between an informative "Find Out Now" option and a non-informative "Keep It Secret" option. By choosing the informative option, participants could know immediately the nature (appetitive, aversive, or neutral) of upcoming images by inference from the animal symbols. By choosing the non-informative option, however, the same animal symbol

always appeared, and the final outcome was only revealed at the end of trial. The diagram only depicts the Good condition, which contains 50% erotic and 50% neutral images. We also tested a Bad condition (50% aversive and 50% neutral images) and a Mix condition (50% erotic and 50% aversive image). **(B)** The time series of choice probability for the informative option (i.e., "Find Out Now"). Shaded area indicates ±SEM. **(C)** Predicted time series of choice probability from the TD model. The model was repeatedly simulated 100 times with the same set of parameters (dashed lines). The solid lines are the median of the dashed lines. At asymptote, the TD model chooses indifferently between the two options. **(D)** Similar simulations from the APE model. The asymptotic behaviour of the APE model agrees with the human data. Both models share the same learning rate, inverse temperature, and discount factor $\alpha = .06$, $\beta = 3$, $\gamma = .98$. For the APE model, additional sampling weights parameters were used, as displayed in the figure legend.

As shown in Figure 18B, people preferred advance information about potential positive outcomes (i.e., erotic images), as in the other information-choice tasks (e.g., Iigaya et al., 2016). When the advance information can potentially reveal the upcoming images to be aversive, in both the Bad and Mix conditions, there was less information-seeking than in the Good condition where there was no worry of receiving such advanced information. In addition, the learning curve indicates that people learn to seek out good news through preferring options with advance information, and to actively avoid bad news through preferring to remain uncertain.

We simulated both the TD and APE models on the information-choice task presented in Zhu et al. (2017). The appetitive images have a positive unit value, the neutral images have zero value, and the aversive images have a negative unit value. As shown in Figure 18C, the asymptotic behaviour of the TD model is to choose indifferently between the informative and non-informative options across all three conditions. Note

that the TD model predicts a short period of information avoidance in the Good condition (see blue solid line in Figure 18C). The aversion to information of the TD model was previously observed (Bromberg-Martin & Hikosaka, 2009; Niv, Joel, Meilijson, & Ruppin, 2002; Denrell, 2007). This is a direct result of integrating learning (i.e., the estimation process based on data: TD learning rule) and control (i.e., the data generation process: softmax choice selection). The estimation of the mean value of an option is based on data that generates itself based on current estimations. Indeed, the integration of learning and control as in the TD model inevitably produces a bias against uncertain options; this point was recently formally proved and generalised to any system that combines learning and control (Nie, Tian, Taylor, & Zou, 2017).

The emergence of information seeking or avoidance through experience can be reproduced by the APE model for all three conditions. In the simulation, we follow previous empirical and theoretical works on how people should sample when extreme events are present (i.e., appetitive and aversive images) (e.g., Lieder et al., 2018; Tversky & Kahneman, 1973; Ludvig, Madan, & Spetch, 2014): the generated samples from imagined future states should be biased toward the most extreme events. This implies higher sampling weights for advanced information that reveals the highly appetitive and aversive images. The sampling weights associated with appetitive and aversive images were thus set to 1. This extremity bias drives the sampling process more toward the advance information that predicts appetitive and aversive images, and hence learns to seek or avoid information, pending the valence. Therefore, the APE model predicts information seeking in the Good condition, information avoidance in the Bad condition, and close to indifference to information in Mix condition.

5.5 Discussion

We presented a novel model of learning to prefer information. The APE model can be seen as an extension of the basic TD model with an

additional sampling process from simulations (i.e., imagined future episodes) assumed. This sampling process generates anticipated prediction errors based on the imagined future prospects and current state valuations. These APEs are treated like primary rewards, which combined with a bias toward sampling trajectories toward positive outcomes, leads to information seeking in situations with potential positive outcomes and to information avoidance in situations with potential aversive outcomes. The positive APEs encourage a preference for that future state whereas the negative APEs discourage such a preference.

We tested the APE model against the classical TD model on the empirical findings using the information-choice paradigm. The empirical findings were categorised into five main groupings with each manipulating one key experimental variable in the task: cue-outcome contingency, uncertainty resolution, delay to outcomes, reward magnitudes, and negative outcomes (see Table 8 for summary). The TD model fails to reproduce any of these applications because of the model's insensitivity to information. The APE model successfully explained all five categories of empirical data using a single, common mechanism of anticipatory sampling from simulations. A number of other modifications of TD model have built on the common assumption that receiving information cues can be rewarding. Information-bonus models treat the act of obtaining information valuable as inherently valuable (Bromberg-Martin & Hikosaka, 2009; 2011; Bennett et al., 2016). The anticipatory utility model postulates a positive utility of anticipating an upcoming certain reward (Loewenstein, 1987; Iigaya et al., 2016). The APE model can be viewed as a mechanistic account for those models (only for the information-bonus model in appetitive conditioning) and possibly could also explain the origin of anticipatory utility.

The information-choice task is a special form of instrumental conditioning. The key difference from the classical conditioning phenomena discussed in Chapter 4 is that animals can "voluntarily" choose among stimuli. Because this additional degree of freedom in experimental design where the animals can choose what to learn, in instrumental conditioning, it has been thought to be difficult to identify exactly when each learning

episode occurs (Skinner, 1963). The APE model presented here as a combination of the standard TD learning and an anticipatory sampling process can be a candidate model to understand instrumental conditioning.

The existing experimental paradigm can be rendered as a shallow decision tree (see Figure 12), and the APE model presented here is assumed to have only one-step anticipation on the model of world. With larger branching and/or deeper decision trees, it becomes quickly impractical to mentally imagine all possible future trajectories. A recent smartphone-based study adopted a four-stage information-seeking game, and observed systematic deviations from the optimal strategy suggested by dynamic programming (Hunt et al., 2016). This type of task poses yet another computational challenge for the models discussed here and urges the models to adapt to more complicated decision tasks. As the APE model embedded with look-ahead experiences, we consider this model more convenient to incorporate with other planning algorithms. Recent developments of planning algorithms suggest that complicated tree search can be roughly divided into two components. For example, in Monte-Carlo tree search (Coulom, 2006), agents can act optimally according to an optimal policy on the part of tree that has been well-explored and act randomly on the less-known part of tree. The applicability of APE model on this type of more extensive tree search can be an interesting future research question.

Chapter 6

Conclusions

6.1 Towards A Theory of Sampling Brain

This thesis has focussed on how mental sampling can enhance computational models of cognition. I pursued a theory of Bayesian sampling (Chapter 2-3) and augmented classical learning models with mental sampling (Chapter 4-5).

The key idea behind the Bayesian sampling is that Bayesian models of cognition need not require the brain to represent and calculate all probabilities, but these can be approximated instead through samples (Sanborn & Chater, 2016; Gershman et al., 2012; Griffiths et al., 2012; Vul et al., 2014; Wozny et al., 2010; Lieder et al., 2012; Zhu et al., 2018). I suggested that a specific sampling algorithm, MC^3, may explain how and why mental samples are autocorrelated. However, from a statistical perspective, independent samples can be best justified because they contain more information than autocorrelated samples and a waste of cognitive resources is to be expected from autocorrelated samples. The MC^3-type of mental sampling algorithm paves a rationale for autocorrelated samples that the brain has to tolerate some degrees of autocorrelation in order to retain the possibility of generating samples from far-away modes.

By considering the stochasticity in the sample-generation process and the limited number of samples generated, the brain can proactively correct for the intrinsic uncertainties of these mental samples. I found evidence in human probability estimates where sample-based estimations for probability judgments are tempered by an additional Bayesian inference on the mental samples. While simplifying assumptions were added to our analysis (i.e., direct sampling and exact Bayesian inference), this result suggests a promising direction for understanding the origin of judgments. A set of

mental samples might be generated at first, but the estimates are not only based on the statistics of these samples. There may exist an additional correction process applied upon these samples before a judgment is formally reported.

I also studied the impact of two sources of mental samples on learning where value estimates are repeatedly updated with regard to new experiences. Learning problems served as a valuable empirical benchmark for the use of mental samples from memory and simulation. Evidence for the reuse of past experiences was found in classical conditioning. According to the model developed here, replayed experiences go through the exact same error-correction learning rule as a new experience does. Even if the sampling from memory is as simple as random, the random replay model can accommodate many classical conditioning phenomena, including spontaneous recovery, latent inhibition, retrospective revaluation, and hippocampal-lesion effects. The model also prescribes a number of novel predictions that could be tested in future behavioural experiments and with hippocampal-lesioned animals. More importantly, the random replay model advocates an extended view on training data for learning systems: past experiences from an agent's interactions with the environment may not be forgotten and could be deliberately reused to shape future behaviour.

Whenever there is a choice among options, animals may activate an anticipatory sampling process based on their current model of world. The prospects from choosing any option can be mentally quantified through the anticipated prediction errors, which reflects the difference in value estimates between the option and imagined samples from choosing that option. Many sub-optimal choices can be explained in this way, through known sampling biases such as an optimism bias and a bias towards extreme events. The Anticipated Prediction Error model described a mental process that, through the combination of a standard error-correction learning rule and anticipatory sampling, reaches a risky choice. Beyond its explanation of preference for information and curiosity-like behaviour, the insights from sampling has further potential in explaining gambling behaviour, drug abuse, anxiety, and many other mental disorders.

6.2 Neural Mechanisms of Sampling

I have demonstrated many appealing properties of sampling on the computational and algorithmic level of analysis of cognition (Marr, 1982). Sampling has also gained momentum to become a unifying framework due to its generalisation to biologically-plausible implementations in spiking neural networks. The neural sampling hypothesis proposes that probability distributions are encoded in samples of neural populations (Orban et al., 2016; Berkes et al., 2011; Hoyer & Hyvarinen, 2003). According to this hypothesis, neural variability is not a nuisance, but rather a vital part of how the brain encodes probability distributions and performs computations with them. The first application of a sampling scheme in spiking neural networks is in visual competition (Hoyer & Hyvarinen, 2003): when viewing ambiguous stimuli with two stable percepts (e.g., Necker's cube and Rubin's vase/face), neurons oscillate between two distinct states (Blake & Logothetis, 2002). That is, the same stimulus can yield two distinct firing rates in the same neuron, resembling a sampling process between two likely percepts of the stimulus.

Matching patterns between neuronal firing rate variability statistics and sampling stochasticity have been further demonstrated in a series of subsequent studies (e.g., Berkes et al., 2011; Savin & Deneve, 2014; Aitchison & Lengyel, 2016). As neurons are not likely to have global access to the activity of all other neurons in the population, all the biologically-plausible sampling algorithms in the literature have features of local sampling (e.g., Savin & Deneve, 2014; Aitchison & Lengyel, 2016). The MC^3 algorithm suggested in Chapter 2 also shares this local search property, but implies adjustments to the existing neural sampling schemes for the algorithm to be implemented in neural hardware: (a) multiple chains of sampling running in parallel but at different temperatures, (b) tracking the cold chain for output, and (c) a switching mechanism between chains. The first two adjustments, as they are similar to a distributed MCMC sampling algorithm, have been implemented in a spiking neural network (Savin & Deneve, 2014). The neural architecture used for distributed MCMC basically combines spatial (information integrated across neurons) and temporal coding (information

integrated across time) of probability distributions. This is a promising architecture to deploy MC^3-like sampling algorithms. However, the switching scheme between chains is more challenging to be implemented in neural populations, and perhaps, there could be a higher-level meta-controller that alternates among low-level neurons which are dedicated to drawing samples.

6.3 Limitations and Alternatives of Sampling

I have discussed the sampling brain as a general form for cognition and have selected classical cognitive phenomena that comply with this view. There are other prominent alternatives to sampling. I will focus on two of them: variational Bayesian inference and reasoning principles. Though many differences in explaining cognition (as we will discuss below), both sampling and variational methods can be broadly classified as approximation algorithms that attempt to get closer to the true probability distribution and hope for a vanishing approximation error with longer computations. In contrast, reasoning principles cut out any approximation process and can provide accurate answers straightaway.

6.3.1 Variational Bayes

The variational methods presume that the true probability distribution belongs to some parametric family of probability distributions (e.g., Gaussian distribution; Jordan, Ghahramani, Jaakkola, & Saul, 1999; Wainwright & Jordan, 2008). Unlike the sample-based approximation, if the target probability distribution is not considered in the parametric family, the approximation error will never approach to zero. In essence, variational methods tend to attain efficiency at the expense of flexibility and bias. Often in practice, variational methods use a simpler probability distribution to approximate a complex probability distribution (Blei, Kucukelbir, & McAuliffe, 2017). Then the inference problem is substituted with an

optimisation problem where some "distances" between the target distribution and parametric family are minimised.

Both sampling and variational algorithms have been proposed as cognitive mechanisms for mental activities (Sanborn, 2017). As I have demonstrated across this thesis, successful applications of sampling in Bayesian models of cognition rely on its ability to accommodate two key aspects of mental activities: (a) stochasticity and (b) limited cognitive resources. Variational Bayes algorithms are also able to replicate these two features of mental activities (Sanborn, 2017; Gershman & Beck, 2017). The difference between sampling and variational methods could be most profound in their neural architectures. Neural implementations of variational methods regard neural noise as a nuisance that should be averaged out across a large population of neurons (e.g., Rao, 2004; Ma et al., 2006; Beck, Pouget, & Heller, 2012). However, the neural sampling hypothesis views the neural noise as a necessary part of stochasticity that supports sampling processes.

It is also possible that the brain could use both sample-based and variational approximation approaches. Hybrids of sampling and variational methods, which combine the strengths of both methods, have been developed to tackle complex probability distributions in machine learning (e.g., variational particle approximations: Saeedi, Kulkarni, Mansinghka, & Gershman, 2017). In addition, the correction for mental samples discussed in Chapter 3 can be potentially enhanced through a "better" correction prior. A better correction prior requires smaller distances between the correction prior and the true probability; naturally, this correction prior can be iteratively improved via variational methods.

6.3.2 Reasoning Principles

A mind exploiting reasoning principles warrants accurate estimations of probabilities, if used appropriately. Such reasoning principles include the *ignorance of irrelevant information* and *principle of indifference* (e.g., van Fraassen, 1989; Strevens, 1998; Kemp & Eddy, 2017). Consider the probability of

tossing a fair dice that lands "1" face up. Without having touched the dice, recalling past dice-tossing experiences, or mentally simulating dice tossing, it is possible that one can give a firm answer: the probability is 1/6. The ability to infer the correct probability in minimum time requires us to invoke the principle of indifference. This principle states that, without sufficient reason to assign any two events different probabilities, they should be assigned the same probability. Here, the fair dice has six faces, and we do not have any reason to believe these six probabilities are distinguishable. Given the fact that one face will have landed up for sure, the sum of these six probabilities must equal to one, and therefore the probability of "1" should be assigned a probability of one-sixth.

Like the variational methods, these reasoning principles are not necessarily at conflict with sampling to compete for control of behaviour. The brain may speed up the process of sampling with the help from reasoning principles. The most straightforward application of cooperation between sampling and reasoning principles is to ignore the irrelevant information. For example, when asked to estimate the length of daytime of a random place on a random day, it is sensible to ignore the information about the location's longitude because length of daytime does not vary with longitude. Then we could concentrate on drawing samples from places that share similar latitudes with the target location.

Indeed, an interesting question that merits future research could be about how the brain coordinates among many computational tools. To solve this question may require a computational complexity perspective (Chater & Vitányi, 2003; Papadimitriou, 2003; Vitányi & Chater, 2017; Gershman, Horvitz, & Tenenbuam, 2015; Bossaerts & Murawski, 2017). The reasoning principles basically discard *fake* complexity that, upon successful removal of fake complexity, no information can be lost for compression (i.e., lossless compression). However, approximation approaches, such as sampling and variational methods, deal with *actual* complexity where any form of data compression is inevitably lossy.

6.4 Envoi

The philosophy of sampling is deeply rooted in experimentation. Experiments are primary components of scientific method, which provide the basis for knowledge, test existing theories, and call for new theories. Likewise, mental sampling plays critical roles for cognition: it provides evidence for the brain to form valid hypotheses about reality. The sampling brain hypothesis is a promising research direction but definitely not the end of history. The potential applications of the hypothesis in fields such as psychiatry, financial markets, and artificial intelligence have yet to be fully explored. Nevertheless, I believe that there will be better theories and models to transcend the sampling brain hypothesis in the future. As a student of science, I hope this future arrives sooner.

References

Aitchison, L., & Lengyel, M. (2016). The Hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS Computational Biology*, *12*(12), e1005186.

Alonso, E., & Schmajuk, N. (2012). Special issue on computational models of classical conditioning guest editors' introduction. *Learning & Behavior*, *40*(3), 231-240.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

Aragones, E., Gilboa, I., Postlewaite, A., & Schmeidler, D. (2005). Fact-free learning. *American Economic Review*, *95*(5), 1355-1368.

Ariely, D. (2009). The end of rational economics. *Harvard Business Review*, *87*(7-8), 78-84.

Austerweil, J. L., Abbott, J. T., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems* (pp. 3041-3049).

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C., Saxe, R., & Tenenbaum, J. (2011, January). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33).

Balleine, B. W., & O'doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 201306572.

Bayes, T. & Price, R. (1763). An essay towards solving a problem in the doctrine of chances.

Beck, J., Pouget, A., & Heller, K. A. (2012). Complex inference in neural circuits with probabilistic population codes and topic models. In *Advances in Neural Information Processing Systems*(pp. 3059-3067).

Beierholm, U. R., & Dayan, P. (2010). Pavlovian-instrumental interaction in 'observing behavior'. *PLoS computational biology*, *6*(9), e1000903.

Bennett, D., Bode, S., Brydevall, M., Warren, H., & Murawski, C. (2016). Intrinsic valuation of information in decision making under uncertainty. *PLoS Computational Biology*, *12*(7), e1005020.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*(6013), 83-87.

Berkolaiko, G., Havlin, S., Larralde, H., & Weiss, G. H. (1996). Expected number of distinct sites visited by N Lévy flights on a one-dimensional lattice. *Physical Review E*, *53*(6), 5774.

Blaisdell, A. P., Gunther, L. M., & Miller, R. R. (1999). Recovery from blocking achieved by extinguishing the blocking CS. *Animal Learning & Behavior*, *27*(1), 63-76.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience*, *3*(1), 13.

Blanchard, T. C., Hayden, B. Y., & Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron*, *85*(3), 602-614.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, *1*(1), 121-143.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877.

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, *35*(7), 1011-1023.

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

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, *30*(2), 149-165.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, *114*(1), 80.

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological psychiatry*, *52*(10), 976-986.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119-126.

Bromberg-Martin, E. S., & Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. *Nature Neuroscience*, *14*(9), 1209.

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, *7*(11), e1002211.

Charpentier, C. J., Bromberg-Martin, E. S., & Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proceedings of the National Academy of Sciences*, *115*(31), E7255-E7264.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335-344.

Chater, N., & Oaksford, M. (Eds.). (2008). The probabilistic mind: Prospects for Bayesian cognitive science. OUP Oxford.

Chater, N., & Vitányi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346-369.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations.

Chow, J. J., Smith, A. P., Wilson, A. G., Zentall, T. R., & Beckmann, J. S. (2017). Suboptimal choice in rats: Incentive salience attribution

promotes maladaptive decision-making. *Behavioural brain research*, *320*, 244-254.

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, *395*(6699), 272.

Clayton, N. S., Bussey, T. J., & Dickinson, A. (2003). Can animals recall the past and plan for the future?. *Nature Reviews Neuroscience*, *4*(8), 685.

Cohen, J. D., McClure, S. M., & Angela, J. Y. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1481), 933-942.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463.

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*(2), 304-321.

Costello, F., & Watts, P. (2018). Probability Theory Plus Noise: Descriptive Estimation and Inferential Judgment. *Topics in Cognitive Science*, *10*(1), 192-208.

Costello, F., Watts, P., & Fisher, C. (2018). Surprising rationality in probability judgment: Assessing two competing models. *Cognition*, *170*, 280-297.

Coulom, R. (2006, May). Efficient selectivity and backup operators in Monte-Carlo tree search. In *International conference on computers and games* (pp. 72-83). Springer, Berlin, Heidelberg.

Courville, A. C., & Daw, N. D. (2008). The rat as particle filter. In *Advances in neural information processing systems* (pp. 369-376).

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294-300.

Crockett, M. J., Braams, B. R., Clark, L., Tobler, P. N., Robbins, T. W., & Kalenscher, T. (2013). Restricting temptations: neural mechanisms of precommitment. *Neuron*, *79*(2), 391-401.

Crupi, V., & Tentori, K. (2016). Noisy probability judgment, the conjunction fallacy, and rationality: Comment on Costello and Watts (2014).

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from?. *Cognitive Psychology*, *96*, 1-25.

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67-81.

Dave, A. S., & Margoliash, D. (2000). Song replay during sleep and computational rules for sensorimotor vocal learning. *Science*, *290*(5492), 812-816.

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal replay of extended experience. *Neuron*, *63*(4), 497-507.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876.

Deneve, S. (2008). Bayesian spiking neurons I: inference. *Neural computation*, *20*(1), 91-117.

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review*, *114*(1), 177.

Devenport, L. D. (1998). Spontaneous recovery without interference: Why remembering is adaptive. *Animal Learning & Behavior*, *26*(2), 172-181.

Devenport, L., Hill, T., Wilson, M., & Ogden, E. (1997). Tracking and averaging in variable environments: A transition rule. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*(4), 450.

Dickinson, A. (1996). Within compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology: Section B*, *49*(1), 60-80.

Dinsmoor, J. A. (1983). Observing and conditioned reinforcement. *Behavioral and Brain Sciences*, *6*(4), 693-704.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180.

Douglas, R. J. (1967). The hippocampus and behavior. *Psychological bulletin*, *67*(6), 416.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. (Eds.). (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216-222.

Dunn, R., & Spetch, M. L. (1990). Choice with uncertain outcomes: Conditioned reinforcement effects. *Journal of the experimental analysis of behavior*, *53*(2), 201-218.

Edwards, W. (1968). Conservatism in human information processing. *Formal representation of human judgment*.

Eichenbaum, H., & Cohen, N. J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function?. *Neuron*, *83*(4), 764-770.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*(3), 145.

Euston, D. R., Tatsuno, M., & McNaughton, B. L. (2007). Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science*, *318*(5853), 1147-1150.

Evans, J. S. B., & Over, D. E. (2013). *Rationality and reasoning*. Psychology Press.

Farrell, S., Wagenmakers, E. J., & Ratcliff, R. (2006). *1/f* noise in human cognition: Is it ubiquitous, and what does it mean?. *Psychonomic Bulletin & Review*, *13*(4), 737-741.

Fennell, J., & Baddeley, R. (2012). Uncertainty plus prior equals rational bias: An intuitive Bayesian probability weighting function. *Psychological Review*, *119*(4), 878.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659.

Fortes, I., Vasconcelos, M., & Machado, A. (2016). Testing the boundaries of "paradoxical" predictions: Pigeons do disregard bad news. *Journal of Experimental Psychology: Animal Learning and Cognition*, *42*(4), 336.

Foster, D. J. (2017). Replay comes of age. *Annual review of neuroscience*, *40*, 581-602.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230-1233.

Gao, J. B., Billock, V. A., Merk, I., Tung, W. W., White, K. D., Harris, J. G., & Roychowdhury, V. P. (2006). Inertia and memory in ambiguous visual perception. *Cognitive Processing*, *7*(2), 105-112.

Gardner, R. J., & Moser, M. B. (2017). Multiple mechanisms for memory replay?. *Science*, *355*(6321), 131-132.

Gershman, S. J. (2017). Predicting the past, remembering the future. *Current Opinion in Behavioral Sciences*, *17*, 7-13.

Gershman, S. J., & Beck, J. M. (2017). Complex Probabilistic Inference. *Computational Models of Brain and Behavior*, 453.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & behavior*, *40*(3), 255-268.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197.

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

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1-24.

Gershman, S., & Goodman, N. (2014, January). Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Gershman, S., Vul, E., & Tenenbaum, J. B. (2009). Perceptual multistability as Markov chain Monte Carlo inference. In *Advances in Neural Information Processing Systems* (pp. 611-619).

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452.

Ghahramani, Z., & Rasmussen, C. E. (2003). Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*(pp. 505-512).

Gigerenzer, G. (2001). Decision making: Nonrational theories. In *International encyclopedia of the social and behavioral sciences* (pp. 3304-3309). Elsevier Science.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451-482.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, *8*(4), 296-301.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). *1/f* noise in human cognition. *Science*, *267*(5205), 1837-1839.

Gipson, C. D., Alessandri, J. J., Miller, H. C., & Zentall, T. R. (2009). Preference for 50% reinforcement over 75% reinforcement by pigeons. *Learning & Behavior*, *37*(4), 289-298.

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, *55*(1), 96-135.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779.

Green, L., & Rachlin, H. (1977). Pigeons' preferences for stimulus information: effects of amount of information. *Journal of the Experimental Analysis of Behavior*, *27*(2), 255-263.

Griffiths, T. L., Lieder, F., & 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.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, *18*(12), 1069-1076.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263-268.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, *65*(5), 695-705.

Haberlandt, K., Hamsher, K., & Kennedy, A. W. (1978). Spontaneous recovery in rabbit eyelid conditioning. *The Journal of General Psychology*, *98*(2), 241-244.

Haefner, R. M., Berkes, P., & Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, *90*(3), 649-660.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the annual meeting of the cognitive science society*.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Hennequin, G., Vogels, T. P., & Gerstner, W. (2014). Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, *82*(6), 1394-1406.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46-54.

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral biology*, *12*(4), 421-444.

Hofstadter, D. R. (1995). Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought. Basic books.

Hollis, K. L. (1982). Pavlovian conditioning of signal-centered action patterns and autonomic behavior: A biological analysis of function. In *Advances in the Study of Behavior* (Vol. 12, pp. 1-64). Academic Press.

Hollis, K. L. (1997). Contemporary research on Pavlovian conditioning: A" new" functional analysis. *American Psychologist*, *52*(9), 956.

Hoyer, P. O., & Hyvärinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Advances in Neural Information Processing Systems* (pp. 293-300).

Hunt, L. T., Rutledge, R. B., Malalasekera, W. N., Kennerley, S. W., & Dolan, R. J. (2016). Approach-induced biases in human information sampling. *PLoS biology*, *14*(11), e2000638.

Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J., & Dayan, P. (2016). The modulation of savouring by prediction error and its effects on choice. *eLife*, *5*, e13747.

Jaynes, E. T. (2003). Probability theory: The logic of science. Cambridge university press.

Jenkins, H. M., & Boakes, R. A. (1973). Observing stimulus sources that signal food or no food. *Journal of the Experimental Analysis of Behavior*, *20*(2), 197-207.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, *37*(2), 183-233.

Kahneman, D. (2011). *Thinking, fast and slow*(Vol. 1). New York: Farrar, Straus and Giroux.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. *Punishment and aversive behavior*.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498-525.

Kehoe, E. J. (2006). Repeated acquisitions and extinctions in classical conditioning of the rabbit nictitating membrane response. *Learning & Memory*, *13*(3), 366-375.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223-232.

Kemp, C., & Eddy, C. (2017) A Toolbox of Methods for Probabilistic Inference. In *Proceedings of the annual meeting of the cognitive science society*.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20.

Kendall, S. B. (1974). Preference for intermittent reinforcement. *Journal of the Experimental Analysis of Behavior*, *21*(3), 463-473.

Kendall, S. B. (1975). Enhancement of conditioned reinforcement by uncertainty. *Journal of the Experimental Analysis of Behavior*, *24*(3), 311-314.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712-719.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge University Press.

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence*. CRC press.

Laborda, M. A., & Miller, R. R. (2013). Preventing return of fear in an animal model of anxiety: additive effects of massive extinction and extinction in multiple contexts. *Behavior Therapy*, *44*(2), 249-261.

Laplace, P. S. (1825). Quatrième Supplément a la Théorie analytique des probabilités. Imprimerie de Huzard-Courcier.

Laplace, P. S. (1829). Essai philosophique sur les probabilités. H. Remy.

Laplace, P. S. (2012). Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator (Vol. 13). Springer Science & Business Media.

Legge, E. L., Madan, C. R., Spetch, M. L., & Ludvig, E. A. (2016). Multiple cue use and integration in pigeons (Columba livia). *Animal cognition*, *19*(3), 581-591.

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.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322-349.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in Neural Information Processing Systems* (pp. 2690-2798).

Lin, L. J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, *8*(3-4), 293-321.

Liu, R., & Zou, J. (2017). The effects of memory replay in reinforcement learning. *arXiv preprint arXiv:1710.06574*.

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. *The Economic Journal*, *97*(387), 666-684.

Lubow, R. E. (1973). Latent inhibition. *Psychological Bulletin*, *79*(6), 398.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of comparative and physiological psychology*, *52*(4), 415.

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making*, *27*(2), 146-156.

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2015). Priming memories of past wins induces risk seeking. *Journal of Experimental Psychology: General*, *144*(1), 24.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning & behavior*, *40*(3), 305-319.

Ludvig, E. A., Sutton, R. S., Verbeek, E., & Kehoe, E. J. (2009). A computational model of hippocampal function in trace conditioning. In *Advances in Neural Information Processing Systems* (pp. 993-1000).

Ludvig, E. A., Zhu, J. Q., Mirian, M. S., Kehoe, E. J., & Sutton, R. S. (2017). Associative learning from replayed experience. *bioRxiv*, 100800.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579-2605.

MacKay, D. J. (1998). Introduction to Monte Carlo methods. In *Learning in graphical models* (pp. 175-204). Springer, Dordrecht.

MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.

Mackintosh, N. J. (1974). The psychology of animal learning. Academic Press.

Mackintosh, N. J. (1983). Conditioning and associative learning (p. 316). Oxford: Clarendon Press.

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin & Review*, *21*(3), 629-636.

Maei, H. R., & Sutton, R. S. (2010, March). GQ (λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the third conference on artificial general intelligence* (Vol. 1, pp. 91-96).

Marcus, G. (2009). Kluge: The haphazard evolution of the human mind. Houghton Mifflin Harcourt.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988).

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

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *bioRxiv*, 225664.

Matzel, L. D., Schachtman, T. R., & Miller, R. R. (1985). Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. *Learning and Motivation*, *16*(4), 398-412.

McDevitt, M. A., Dunn, R. M., Spetch, M. L., & Ludvig, E. A. (2016). When good news leads to bad choices. *Journal of the experimental analysis of behavior*, *105*(1), 23-40.

McRaney, D. (2011). You are not so smart. Dutton.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087-1092.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General*, *125*(4), 370.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*(3), 363.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Moore, J. W., & Blazis, D. E. (1989). Simulation of a classically conditioned response: A cerebellar neural network implementation of the Sutton–Barto–Desmond model. In *Neural models of plasticity* (pp. 187-207).

Moore, J. W., Choi, J. S., & Brunzell, D. H. (1998). Predictive timing under temporal uncertainty: the time derivative model of the conditioned response. *Timing of behavior: Neural, psychological, and computational perspectives*, 3-34.

Moore, J. W., Desmond, J. E., Berthier, N. E., Blazis, D. E., Sutton, R. S., & Barto, A. G. (1986). Simulation of the classically conditioned

nictitating membrane response by a neuron-like adaptive element: Response topography, neuronal firing, and interstimulus intervals. *Behavioural Brain Research*, *21*(2), 143-154.

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491-12496.

Napier, R. M., Macrae, M., & Kehoe, E. J. (1992). Rapid reacquisition in conditioning of the rabbit's nictitating membrane response. *Journal of Experimental Psychology: Animal Behavior Processes*, *18*(2), 182.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, *6*(4), 353-366.

Nie, X., Tian, X., Taylor, J., & Zou, J. (2017). Why adaptively collected data have negative bias and how to correct for it. *arXiv preprint arXiv: 1708.01977*.

Nilsson, H., Juslin, P., & Winman, A. (2016). Heuristics can produce surprisingly rational probability estimates: Comment on Costello and Watts (2014).

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in foraging bees: A simple explanation for risk averse behavior. *Neurocomputing*, *44*, 951-956.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

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

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, *92*(2), 530-543.

Papadimitriou, C. H. (2003). *Computational complexity* (pp. 260-265). John Wiley and Sons Ltd..

Pavlov, I. P. (1927). Conditional reflexes: an investigation of the physiological activity of the cerebral cortex.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*(1), 111-139.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29.

Pfister, J. P., Dayan, P., & Lengyel, M. (2010). Synapses with short-term plasticity are optimal estimators of presynaptic membrane potentials. *Nature Neuroscience*, *13*(10), 1271.

Prokasy Jr, W. F. (1956). The acquisition of observing responses in the absence of differential external reinforcement. *Journal of Comparative and Physiological Psychology*, *49*(2), 131.

Ramos-Fernández, G., Mateos, J. L., Miramontes, O., Cocho, G., Larralde, H., & Ayala-Orozco, B. (2004). Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). *Behavioral ecology and Sociobiology*, *55*(3), 223-230.

Rao, R. P. (2004). Bayesian computation in recurrent neural circuits. *Neural computation*, *16*(1), 1-38.

Rescorla, R. A. (2004). Spontaneous recovery. *Learning & Memory*, *11*(5), 501-509.

Rescorla, R. A. (2006). Deepened extinction from compound stimulus presentation. *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(2), 135.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current Research and Theory*, *2*, 64-99.

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as Lévy foraging. *Physica A: Statistical Mechanics and its Applications*, *385*(1), 255-260.

Rhodes, T., Kello, C., & Kerster, B. (2011, January). Distributional and temporal properties of eye movement trajectories in scene perception. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Roberts, W. A. (2002). Are animals stuck in time?. *Psychological Bulletin*, *128*(3), 473.

Roberts, W. A. (2014). Animal cognition. The Wiley Blackwell handbook of operant and classical conditioning, 393-415.

Roper, K. L., & Zentall, T. R. (1999). Observing behavior in pigeons: The effect of reinforcement probability and response cost using a symmetrical choice procedure. *Learning and Motivation*, *30*(3), 201-220.

Rosenthal, J. S. (2011). Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, *4*(10.1201).

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

Saeedi, A., Kulkarni, T. D., Mansinghka, V. K., & Gershman, S. J. (2017). Variational particle approximations. *The Journal of Machine Learning Research*, *18*(1), 2328-2356.

Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and cognition*, *112*, 98-101.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883-893.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Savin, C., & Deneve, S. (2014). Spatio-temporal representations of uncertainty in spiking neural networks. In *Advances in Neural Information Processing Systems* (pp. 2024-2032).

Savin, C., Dayan, P., & Lengyel, M. (2014). Optimal recall from bounded metaplastic synapses: predicting functional adaptations in hippocampal area CA3. *PLoS Computational Biology*, *10*(2), e1003489.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, *8*(9), 657.

Schmajuk, N. A., & Isaacson, R. L. (1984). Classical contingencies in rats with hippocampal lesions. *Physiology & behavior*, *33*(6), 889-893.

Schmaltz, L. W., & Theios, J. (1972). Acquisition and extinction of a classically conditioned response in hippocampectomized rabbits (Oryctolagus cuniculus). *Journal of comparative and physiological psychology*, *79*(2), 328.

Schwarting, R. K. W., & Busse, S. (2017). Behavioral facilitation after hippocampal lesion: a review. *Behavioural Brain Research*, *317*, 401-414.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, *14*(9), 425-432.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1-21.

Sharot, T. (2011). The optimism bias. *Current biology*, *21*(23), R941-R945.

Sharot, T., Riccardi, A. M., Raio, C. M., & Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. *Nature*, *450*(7166), 102.

Shlesinger, M. F., Zaslavsky, G. M., & Frisch, U. (1995). Lévy flights and related topics in physics. In *Levy flights and related topics in Physics* (Vol. 450).

Siegel, S. (1989). Pharmacological conditioning and drug effects. In *Psychoactive drugs* (pp. 115-180). Humana Press, Totowa, NJ.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354.

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., ... & Morritt, D. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098.

Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine learning*, *22*(1-3), 123-158.

Sissons, H. T., & Miller, R. R. (2009). Spontaneous recovery of excitation and inhibition. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(3), 419.

Skinner, B. F. (1963). Operant behavior. *American Psychologist*, *18*(8), 503.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 573.

Spetch, M. L., Belke, T. W., Barnet, R. C., Dunn, R., & Pierce, W. D. (1990). Suboptimal choice in a percentage-reinforcement procedure: Effects of signal condition and terminal-link length. *Journal of the experimental analysis of behavior*, *53*(2), 219-234.

Stagner, J. P., & Zentall, T. R. (2010). Suboptimal choice behavior by pigeons. *Psychonomic Bulletin & Review*, *17*(3), 412-416.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1-26.

Stigler, S. M. (2005). PS Laplace, Théorie analytique des probabilités, (1812); Essai philosophique sur les probabilités, (1814). In *Landmark Writings in Western Mathematics 1640-1940* (pp. 329-340).

Strevens, M. (1998). Inferring probabilities from symmetries. *Noûs*, *32*(2), 231-246.

Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning stochastic inverses. In *Advances in Neural Information Processing Systems* (pp. 3048-3056).

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, *17*(2), 129-144.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine Learning Proceedings 1990* (pp. 216-224).

Sutton, R. S., & Barto, A. G. (1987, July). A temporal-difference model of classical conditioning. In *Proceedings of the ninth annual conference of the cognitive science society*(pp. 355-378).

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement.

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

Sutton, R. S., Szepesvari, C., Geramifard, A., & Bowling, M. (2008). Dyna-style planning with linear function approximation and prioritized sweeping. In Ruishan Liu, James Zou (Eds.). *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 528–536. AUAI Press.

Swendsen, R. H., & Wang, J. S. (1986). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, *57*(21), 2607.

Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 193.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Thorndike, E. (1911). Animal intelligence: Experimental studies. Routledge.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4).

Thorpe, W. H. (1956). Learning and instinct in animals.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*(5), 727-741.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207-232.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

Van Fraassen, B. C. (1989). Laws and symmetry.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, *25*(2), 127-151.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*(3), 331.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and 1/f scaling. *Journal of Experimental Psychology: General*, *134*(1), 117.

Vanseijen, H., & Sutton, R. (2015, June). A deeper look at planning as learning from replay. In *International conference on machine learning* (pp. 2314-2322).

Vasconcelos, M., Monteiro, T., & Kacelnik, A. (2015). Irrational choice and the value of information. *Scientific reports*, *5*, 13874.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., & Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, *381*(6581), 413.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., Da Luz, M. G. E., Raposo, E. P., & Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, *401*(6756), 911.

Vitányi, P. M., & Chater, N. (2017). Identification of probabilities. *Journal of mathematical psychology*, *76*, 13-24.

Von Mises, R. (1957). Probability, Statistics, and Truth: 2d Rev. English Ed. Prepared by Hilda Geiringer. Allen and Unwin.

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.

Wagenmakers, E. J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of *1/f* α noise in human cognition. *Psychonomic Bulletin & Review*, *11*(4), 579-615.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*(1–2), 1-305.

Wasserman, E., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgement: The role of within-compound associations. *The Quarterly Journal of Experimental Psychology: Section B*, *51*(2), 121-138.

Weiss, C., & Disterhoft, J. F. (2015). The impact of hippocampal lesions on trace-eyeblink conditioning and forebrain–cerebellar interactions. *Behavioral Neuroscience*, *129*(4), 512.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*(5172), 676-679.

Wolpert, D. M. (2007). Probabilistic models in human sensorimotor control. *Human Movement Science*, *26*(4), 511-524.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS computational biology*, *6*(8), e1000871.

Wu, C. T., Haggerty, D., Kemere, C., & Ji, D. (2017). Hippocampal awake replay in fear memory retrieval. *Nature neuroscience*, *20*(4), 571.

Wyckoff Jr, L. B. (1952). The role of observing responses in discrimination learning. Part I. *Psychological Review*, *59*(6), 431.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis?. *Trends in Cognitive Sciences*, *10*(7), 301-308.

Zentall, T. R. (2005). Animals may not be stuck in time. *Learning and Motivation*, *36*(2), 208-225.

Zentall, T. R., Laude, J. R., Stagner, J. P., & Smith, A. P. (2015). Suboptimal choice by pigeons: Evidence that the value of the conditioned

reinforcer rather than its frequency determines choice. *The Psychological Record*, *65*(2), 223-229.

Zhang, S., & Sutton, R. S. (2017). A Deeper Look at Experience Replay. *arXiv preprint arXiv:1712.01275*.

Zhu, J. Q., Sanborn, A., & Chater, N. (2018). Mental Sampling in Multimodal Representations. In *Advances in Neural Information Processing Systems* (pp. 5749-5762).

Zhu, J. Q., Xiang, W., & Ludvig, E. A. (2017). Information seeking as chasing anticipated prediction errors. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.