INFORMATION SAMPLING, LEARNING AND EXPLORATION

Andra Geana

A DISSERTATION PRESENTED TO

THE FACULTY OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PSYCHOLOGY

Advisor: Jonathan D. Cohen

June 2015

Abstract

Our world, uncertain and rich in information, often presents us with many available choices, incomplete knowledge about most of them, and noisy feedback that we must make sense of. To make good decisions, we must extract information from our complex environments, form accurate representations of the available options, and perform efficient computations that help us choose the most goal-relevant actions. This thesis presents a series of studies investigating how humans learn information from the tasks they perform, and how they use that information to update representations, estimate the value of their actions, and adaptively adjust their behavior. Chapter 2 compared two conceptually different models of human learning strategies in an information-rich, probabilistic learning task, finding that humans are not Bayes-optimal when extracting the value of relevant features in noisy environments, and that it was possible to directly influence their performance by tailoring the information they received to their individual learning strategies. Chapter 3 delved further into the question of how people learn information about their available options, introducing the exploration-exploitation dilemma. Two experiments – one using a two-armed bandit task with a decision-horizon manipulation, and the other using a similar wheel-of-fortune design with an additional risk manipulation – suggested that people use exploration as a mechanism for acquiring information about unknown options, and that exploration strategies are affected by the decision horizon, risk and ambiguity. Chapter 4 examined the connection between information-seeking and exploration in terms of its effects on motivation. Five experiments showed that participants' boredom ratings depended on task informativeness, as well as the perceived opportunity cost of performing a task, and that higher boredom was correlated with increased exploration. A normative model was proposed, accounting for adaptive, boredom-driven exploration: when the environment has a global structure that needs to be learned, spending time locally maximizing reward must be balanced with the need for learning useful information; thus, boredom can occur when the current action (even if it is reward-maximizing) does not yield sufficient new information about the global environment structure. Overall, this work investigated the relationship between information, learning and exploration, and determined key factors that drive exploratory behavior in uncertain decision contexts, as well as the potential role of exploration as an adaptive information-sampling strategy.

Acknowledgments

First and foremost I would never have been able to get here without the help and guidance of my advisors, Jonathan Cohen and Yael Niv. So much of what I have learned during my time here, not only about science itself but about how to *be* a scientist, how to ask questions and how to interact with other researchers, I learned from them. I feel incredibly lucky to have worked with them, and will always be grateful for their advice and support that have brought me so far. I am extremely grateful as well to Alin Coman, whose encouragement and thoughtful feedback I could always rely on, and to Nathaniel Daw, for all his help and input that were invaluable to my modelling efforts. Many thanks as well to Jordan Taylor, for reading the first draft of this dissertation and providing insight that helped shape it into its final form.

I have also received tremendous support from our postdocs, particularly Bob Wilson who mentored me through essentially every single project I have ever done during my graduate work, and who showed a supernatural amount of patience for answering all my questions and getting me back on track when I floundered. Bob is my hero. Angela Langdon, my erstwhile – and lately, recurring – officemate and writing-coffee partner, has helped me so much through the writing process, and I am so glad to have had her and all our conversations and lunches. I would also like to thank Reka Daniel for her support and her friendship, and David Freestone, for so many helpful conversations and for sharing his invaluable Matlab expertise. Many thanks to everyone else in the Cohen and Niv labs as well, who have been wonderful colleagues and friends.

I would also like to thank my friends outside the lab, and everyone in my cohort. Particularly, Courtney Bearns, who has been my sounding board and my dissertation buddy for a very long time, and has generously talked me down from many ledges, fed me and made us tea and frequently reminded me to keep breathing. I'd like to reiterate how lucky I feel that we were in the same cohort. Thank you for all the puppy gifs. Similarly, thank you so much Ana-Maria Piso, for always being there, my second pea in our very dinky pod. Our phone chats always helped brighten my day; please don't forget you still owe me twenty-six ice creams. Many thanks to all my other friends who have supported me while working on this dissertation, either by talking to me about science, or by listening to me, by sending me care packages or writing me stories to read when I needed sanity boosters. You are all wonderful and I couldn't have done it without you.

It goes without saying that I will always, always be grateful for all the love and support from my family, who have always been there for me and whom I love very much. You are the most precious to me. Thank you.

Table of Contents

General Introduction

Making decisions in the real world is difficult – because the world is complex, uncertain and rich in information. We often have numerous alternatives to choose from (so making the choice itself is hard), we rarely have complete information about most of those alternatives (so it is difficult to know for sure which one is better), and often we get a lot of simultaneous feedback for an action, so we must figure out credit assignment and how to spread the information we've just earned among the options available. None of these are straightforward computations. But our brains – and in fact, many animal brains, as well – have evolved to help us with this kind of decisions. We are capable of complex learning, and given even a rich, noisy environment, we can extract sufficient statistical information from it to form reasonable representations of its structure. We can then use these representations to make choices and update our value estimates based on feedback, so as to make them even more accurate and in turn make even better choices. In short, when exposed to the complex world, we are capable of learning many of its useful statistics, and gearing our actions toward those that are most relevant to our goals.

But learning to determine what is and is not relevant to our goals is itself a challenge. Given the quantity and breadth of information usually available, selecting what we want to learn more about and what we can safely ignore often entails complex computations about reward probabilities, choice histories, and expected future values. The types of information we experience (Knock et al. 2005), as well as the order of information presentation (Ritter et al., 2007; Coenen, Rehder & Gureckis, 2013) determine what structure of the environment we learn, and a significant body of work, from psychology and neuroscience (Dias, Robbins & Roberts 1996; Kruschke 2006; Markant & Gureckis 2012) to computer science and machine learning (Sebastiani & Wynn 2010), has focused on pinning down the behavioral, computational and neural mechanisms involved in this type of learning. Studies have shown a wide array of possible strategies for parsing and sampling the large amounts of information available to us, but the exact mechanisms that underlie this process are not yet fully known. In chapter 2 of this work, I discuss some of these strategies in humans, and show that manipulating the available information can significantly impact learning, credit assignment, and our subsequent ability to choose the most rewarding options in the environment.

Even with a good information sampling strategy in place, however, the question of selecting the option with the highest reward is not always straightforward – and not just because of the multitude of options usually available, and the potential difficulty in differentiating between them. When choosing between multiple options, organisms must at the same time decide between at least two goals: one is earning as much immediate reward as possible – such as money, in the case of people, or food, if we consider foraging animals – but another goal is refining the representation of the environment to make sure that they are indeed choosing the best option. In order to do the latter, it is necessary to sample information from more than one option. This sampling might come at the cost of foregoing some immediate reward, but in the long-term it might lead to discovering better sources of reward, and thus a higher overall reward. This is a frequently encountered tradeoff that the literature commonly refers to as the explore-exploit dilemma (Krebs, Kacelnik & Taylor 1978; Kaelbling 1996; Cohen, McClure & Yu 2007). Although in terms of local reward maximization, exploring lower-reward options might seem suboptimal, in the long term, the information learned from exploratory choices often makes up for its cost. In chapter 3 of the present work, I discuss two different types of human exploration strategies, and show that humans use both these strategies when dealing with different types of uncertainty in the environment.

Indeed, information is so important, that it acquires value in and of itself, independent of reward. Actions that might yield less extrinsic reward (such as money, food, points in a game etc.), but contain information, can be seen as desirable under certain circumstances. Furthermore, prolonged exposure to environments that do *not* hold much information can become aversive (Hill & Perkins 1985; Patyn et al. 2008); indeed, a failure to properly engage our information processing systems could be an important factor underlying affective experiences such as boredom (Eastwood et al. 2012), which ultimately lead to demotivation and task disengagement. Under those circumstances, exploration arises as a frequent behavioral consequence: people are likely to abandon their current task, and switch to a different alternative in their environment. This raises the question of whether subjective experiences such as boredom might constitute a task-related signal of decreased informational content, and bias us toward exploration as a mechanism for encountering new, better sources of information. Chapter 4 of this work examines this question, and shows that increased perception of boredom can arise from environmental structures in which there is little useful information, and that exploration follows as a mechanism for seeking better information content.

The overall scope of this work was to investigate the relationship between information, learning and exploration. I aimed to study the value of exploration as related to information acquisition, determine some of the key factors that drive exploratory behavior in uncertain environments, and examine whether, in certain circumstances, disengaging from a rewarding task to explore the environment could represent an adaptive information sampling strategy.

Chapter 1: Background

1.1. Information Sampling and Credit Assignment in Complex Environments

Most living organisms in our world are often faced with a wide variety of sensory input. A jungle songbird, for instance, lives in a constant symphony of sounds, and must learn their meaning. Most of these sounds are not vital to survival, and can be ignored (such as the sound of a nearby river, or distant cars from the highway), but others, such as the alarm calls of nearby birds or the mating calls of conspecifics, must be attended. A human analog is the cocktail party effect (Arons 1992). At a cocktail party, most input – glasses clanking, music playing, strangers talking –is not important, but a friend's voice may be among the cacophony, or you might be trying to locate a person of interest. Humans and other animals constantly face this type of situation: performing even the simplest of daily activities, we must contend with an abundant stream of incoming data in the form of faces, voices, colors, time commitments, social cues etc. Cognitive constraints make it impossible to attend to and process all environmental stimuli equally well. So learning to differentiate the relevant features from the background noise is crucial.

Evolution can help by tuning neurobiology to innately relevant features. Frogs' auditory neurons, for instance, are sensitive to only certain frequencies (Singh & Theunissen 2003), and the auditory neurons of some species of songsbirds become highly specialized for the frequency ranges of viable mates (Margoliash 1986). In most cases, however, animals do not seem to benefit from an innate mechanism to select the relevant features in their rich environments, and so they must learn to associate stimuli with outcomes, and then decide which stimuli are relevant. The literature refers to this issue as the "credit assignment" problem: the difficulty of attributing the outcomes in one's environments to the specific stimuli or actions that actual generated those outcomes. Proper credit assignment is a key component of any learning process, as organisms cannot learn without building accurate representations of stimuli-outcome contingencies.

One problem with solving credit assignment is that our daily environments are extremely rich. From a purely rational perspective, an ideal learner who has access to all the information in the environment should also use all that available information to learn all the true correlations and contingencies among the numerous cues (Kruschke, 2006). An ideal Bayesian observer, therefore, will use its entire history to compute the probabilities, regardless of the dimensionality of the problem; mathematically, this type of learning allows the agent to maintain an accurate representation of all the cues in its environment, and always choose the ones with the highest expected value. In practice, however, that framework breaks down. For organisms with limited time and limited cognitive resources, learning about every single cue is at best time-inefficient, and at worst computationally intractable, and thus the exhaustive ideal observer strategy is rarely the most effective option.

The fact that animals still manage to successfully navigate their complex environments, despite being unlikely to use a full Bayesian ideal observer model, suggests that they have developed alternative strategies for learning the causal structure of their environment. Given the time and computational demands of most representation learning tasks, one desirable property of such alternative learning algorithms is the ability scale well to domains with many irrelevant cues. Theoretical and experimental results in machine learning show that parsing the space into smaller subsets of features and selectively focusing on some of those subsets significantly increases the speed of learning without unduly harming generalization or accuracy (Blum & Langley, 1997), and work in the cognition literature suggest that people do indeed employ learning strategies that reduce the set of features to which they must attend (Shepard, Hovland & Jenkins 1961; Kruschke 2006).

Information sampling and representation learning

What are the computational strategies that humans use to learn a representation for a given task? It goes without saying that trial-and-error learning depends on the information that the learner can access, and not only what is learned, but also the speed of learning can be significantly affected by this experienced information (Nelson et al. 2010). Indeed, work in machine learning and information theory has established how information in any given task might be optimally selected so as to maximally discriminate between competing hypotheses and accelerate learning (optimal experimental design, Sebastiani & Wynn 2000).

Although human learning does not always mirror these optimal strategies, the ability to choose which information to sample has been shown to improve learning: when participants were allowed to choose which piece of information they wanted to see next, their learning of category boundaries was better (Gureckis & Markant 2012; Markant & Gureckis 2014). Furthermore, even in the absence of this “active learning” option, the type of information presented can still impact what is learned, for instance in speech motor learning (Knock et al 2000), and the order in which information is presented can also make a significant difference (Ritter et al 2007). Additionally, different information sampling patterns can in fact predict significantly different decisions, even when the sampled information ends up being equivalent (Hills & Hertwig, 2010) – which suggests that even in the same space of available information, search strategies and experienced information have a large influence on humans' learning and ultimate decision policies.

The work in the second chapter of this dissertation proposes a novel method for manipulating information presented to participants as tool for investigating the computational processes underlying representation learning.

1.2. Uncertainty, Information, and The Explore - Exploit Tradeoff

Imagine you are driving home after having watched a big football game in a nearby city. The roads are reasonably crowded, since all the other fans are driving home as well – but you know that if you stay on the highway and take the exit you know in thirty miles, you will definitely be home in about an hour. On the other hand, as you survey the traffic situation you wonder if it’s not worth getting off the highway early, and trying one of the smaller side-roads that are sure to have less traffic. The choice here is clear: do you stay on the highway and take the well-known road home, or do you get off and try to find a shortcut, which might end up saving you time, but it might also end up getting you lost?

This kind of scenario is not restricted to humans: every organism in a natural environment is frequently faced with multiple alternatives, and must choose how to allocate its time among them. In humans, the choice might be between deciding whether to take a well-known route home or look for a shortcut, or between ordering a favorite food or trying something new off the menu at a restaurant. In foraging animals, bees for instance, the choice might be between searching a nearby flower patch for nectar, or flying further from the hive in search of other patches (Gallistel 1990). In either scenario, choosing which action to perform and how long to spend on it influences the amount of benefits (food, money etc.) that an organism receives, the energy it must expend, and the risks it might face (Caraco 1980). The question that follows, then, is: with limited time resources and a wide variety of potential actions, how should we best choose to spend our time?

From a purely economic point of view, abandoning the known, well-establish highway route in favor of taking a completely unknown side-road might make us poor decision-makers (and drivers). We are foregoing a certain positive outcome (getting home at a known time), and instead allocating our time to a series of actions of unknown or questionable benefit. However, anyone who has ever faced this scenario knows that the decision is often not that straightforward.

The Exploration - Exploitation Tradeoff

Going back to the driving home example: is it better to take the highway and know for sure that you will be home in an hour, without getting lost? Or is preferable to try to find a faster shortcut on the back roads, though it is not as certain what potential benefits you will gain from the latter action? The foraging bee faces a similar choice: if it flies further from the hive, it might find a patch with better nectar, or a higher replenishing rate. But it might also find nothing, and thus spend its time and energy for no reward.

These are both examples of what the literature refers to as the "exploration - exploitation tradeoff": the tradeoff between choosing a certain resource alternative (the highway route, the familiar nearby food patch), and searching the environment for other options, with uncertain benefits (calling a friend, more distant food patches). This issue is studied across fields, from animal cognition (Gallistel et al., 2007), to ecology (Caraco 1980), economics (Banks & Sundaram 1994), or reinforcement learning (Kaelbling et al 1996). Nevertheless, due to the complex nature of realistic environments, precisely analyzing exploration and exploitation in all of their different contexts is often impossible, and many questions about exploratory patterns of behavior are difficult to answer.

The literature on the exploration - exploitation dilemma is organized along two main directions. Ample theoretical work from economics, statistics, and machine learning deals with the formal, mathematical foundations of this problem: precise formalizations of the choice environments (Whittle 1980), computation of explicit and simulation-based optimal solutions (Gittins 1979; Whittle 1988;), the development of efficient computational algorithms that regulate the balance between exploration and exploitation (Auer et al. 2002; Yi, Steyvers & Lee 2009; Tokic 2010). However, despite these numerous analytic (algorithmic) approaches, no universal solution exists to date – as the current solutions are always constrained by assumptions that may or may not apply to real world situations of interest, and are not always practical (i.e., likely to be implementable by organisms). More recently, this problem has also raised interest in the cognitive neuroscience community, and a new research direction emerged to examine the question of how animals and humans actually negotiate the explore - exploit tradeoff, and what cognitive and neural mechanisms drive exploratory behavior (Aston-Jones & Cohen 2005; Cohen, McClure & Yu 2007; Behrens et al 2007; Frank et al. 2009).

How do organisms explore and exploit? Evidence of Adaptive Behavioral Adjustments.

To reiterate the above definition, the exploration - exploitation tradeoff refers to the conflict between choosing an action with certain, known benefits, and searching the environment by choosing other options, with uncertain and less immediate benefits. As this type of situation occurs so frequently, it is not surprising that it has been the object of experimental studies for several decades. Early animal studies, for instance, use time-based reinforcement schedules (variable interval schedules, Ferster & Skinner 1957), and later, response-based reinforcement schedules (variable ratio schedules, Loveland & Herrnstein 1970) to study the phenomenon. In humans, participants could be asked to choose between different bets with varying win probabilities (Pratt 1964; Jepma & Nieuwenhuis 2011), to play different virtual slot machines (Daw et al. 2006; Behrens et al 2007; Steyvers, Lee & Wagenmaker 2009), play Go or similar games (Gelly & Wang 2006), or to perform timing and other perceptual tasks (Frank et al. 2009). Any of these tasks can be seen from an exploration - exploitation perspective.

One highly robust finding regarding explore/exploit behavior on a variety of choice tasks is that organisms can adjust their behavior to different task structures. Pigeons show markedly different response rates and response patterns when they are choosing between variable interval schedules and variable ratio schedules (Ferster & Skinner 1957). Rats show different exploration patterns when choosing between two rich bandits than when choosing between two scarce bandits, even if the relative bandit values are the same (Reed, Schachtman & Hall 1988). Rats and people playing non-stationary bandits modify their exploration rates if the environment variability changes (Otto et al. 2010; Behrens et al 2007). Much of the early theoretical work on time allocation in animals focused on finding a law, or principle of behavior, to account for these different responses under different task conditions. Herrnstein (1961) first developed the matching law to describe the global result that animals allocate responses and time to alternatives proportional to their reward (Herrnstein 1961; Baum 1974; Baum 1979). More recent work has also shown that matching holds in humans, as well, (Davison 1988; Logue, Forzano & Tobin 1992), and that humans and animals can efficiently allocate their responses to a drifting reward rate, and they do so in real time (Heyman 1982; Gallistel 2005, 2007; Daw et al. 2006; Rushworth & Behrens 2008), and that abrupt changes in the reward rate are optimally detected and adjusted to (Gallistel, 2001; Courville, Daw & Touretzky 2006; Nassar et al. 2010).

All these findings suggest that animals and humans are sensitive to the specific structures of their decision problems, and can successfully adjust their exploration - exploitation balance to respond to specific environmental changes and maximize overall reward. These results have been found across a variety of tasks; however, there is one particular type of problem that is particularly well-suited for studying the exploration - exploitation tradeoff: the multi-armed bandit problem. The next section describes this, and discusses the current theoretical framework for solving it optimally.

Exploratory decision-making algorithms in the multi-armed bandit problem

The exploration - exploitation tradeoff has generally been treated in the literature within the framework of bandit problems (Robbins, 1952). Having initially borrowed its name from an old term used to describe a slot machine (see fig. 1a for an illustration of a real "one-armed bandit"), the multi-armed bandit problem has gained popularity in several areas of research as one of the simplest non-trivial problems in which one must handle the tradeoff between actions which yield immediate certain rewards and actions (such as acquiring information) which might yield benefits later (Whittle, 1980).

An n-armed bandit problem refers to a decision task in which there is a set of *n* response alternatives (referred to as 'bandits', see fig. 1b), with different reward rates. On each trial, the decision maker must choose a bandit, after which they receive feedback about the amount of reward earned for that choice. The decision maker's task is to incorporate this feedback to make a series of choices among the different bandits that maximizes the overall reward. Its sequential nature is a key feature of the bandit problem, and it is what makes multi-armed bandits an ideal framework for the study of exploration - exploitation behaviors. Indeed, the very notion of time allocation (and the trade-off between choosing familiar options with known benefits, and searching unfamiliar options with unknown benefits) relies upon on the idea that the decision-maker has some limited amount of time and energy to invest, and is free to allocate between several response options as it sees fit, choosing each option as frequently as it wants.

The structure of the bandits can vary depending on the generative process underlying the rewards from each bandit. Frequently, the rewards generated by a bandit come from an underlying distribution (Gaussian, binomial etc.) with certain mean and variance parameters. These parameters can remain constant over time (in which case the bandits are referred to as "stationary", cf. Gittins, 1989), or they can drift or change abruptly (making the bandits non-stationary). The bandits can sometimes be rich, with very high reward rates, or sometimes be scarce, with very low reward rates. Figure 1B illustrates several different types of bandits and their corresponding reward structures.

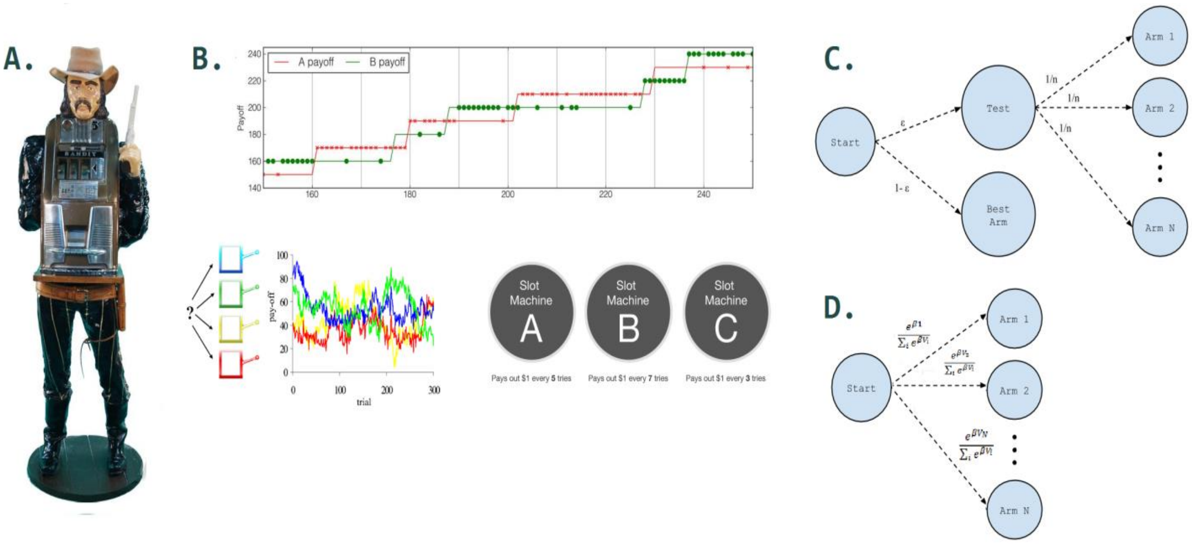

A.
B.
A payoff
B payoff
Payoff
240
220
200
180
160
140
160
180
200
220
240
?
pay-off
100
80
60
40
20
0
0
100
200
300
trial
Slot Machine A
Slot Machine B
Slot Machine C
Pays out $1 every 5 tries
Pays out $1 every 7 tries
Pays out $1 every 3 tries
C.
Start
ε
1-ε
Test
Best Arm
1/n
1/n
1/n
Arm 1
Arm 2
Arm N
D.
Start
$\frac{e^{\beta V_1}}{\sum_i e^{\beta V_i}}$
$\frac{e^{\beta V_2}}{\sum_i e^{\beta V_i}}$
$\frac{e^{\beta V_N}}{\sum_i e^{\beta V_i}}$
Arm 1
Arm 2
Arm N

Figure 1: Bandit problems. A: "Original" one-armed bandit slot machine. B: Examples of different bandit structures (top – leapfrog bandit, Knox et al. 2011; bottom left – drifting bandits, Daw et. al 2006; bottom right – example of fixed bandits). C: The ε-greedy algorithm. D. The softmax algorithm.

Different bandit payoff structures lead to different exploration - exploitation patterns. Tasks with non-stationary bandits, for instance, generate higher exploration rates than similar tasks with stationary bandits (Otto et al 2010). Binary bandits generate different choice patterns from Gaussian bandits, and rich bandits different patterns from scarce bandits (Caraco 1980). This once again speaks to the idea that organisms are capable of adapting their behavior to reflect the specific problem they are solving. It is then reasonable to consider the question of optimal allocation: is there a best way for a decision-maker to allocate its responses in a given context?

Optimal Exploration in Bandit Problems

The existence of a task-specific optimal policy for balancing exploration and exploitation to maximize expected payoff over time has been widely researched in bandit problems (Zhang, Xu & Callan 2003; Bogacz et al. 2006; Steyvers, Lee & Wagenmakers 2009; Lee et al. 2010). In real-world situations, it is impossible to obtain a closed-form optimal solution because the number of choice options is too high, and details about the time horizon and the specific reward structures of each option, and their stationarity, can never be fully known. However, constraining the environments by making a conservative set of assumptions makes it possible to identify several types of strategies that maximize total payoff.

The first explicit solution to a multi-armed bandit problem was developed by Gittins and colleagues (Gittins & Jones 1974, Gittins 1979), for a finite number of stationary bandits (bandits whose reward rates never change, such a food patch that never depletes). They modeled the multi-armed bandit problem as a Markov decision process (a systems that transition between states based on transition probabilities for each state) with an infinite time horizon (that is, the process goes on forever), and exponential discounting of future rewards. Thus, assuming *n* bandits, Gittins showed that it was possible to calculate allocation indices, *vi*, for each individual bandit B*i*, as a function of only the decision time t and the state of the bandit at that time. These indices *vi* – now known as Gittins indices – were calculated by solving the problem of optimal time allocation between each bandit and a theoretical bandit with a standard fixed value λ.

Following this computation, the strategy proposed by Gittins was to simply choose, at every decision time t, the bandit with the highest index. This makes it a "greedy" policy – as it always prescribes choosing the highest-valued option; the difference from other greedy policies lies in its optimality. Whittle (1980) used a dynamic programming approach to show that the Gittins policy is indeed optimal for stationary multi-armed bandit problems. In a later paper, Whittle also extended the Gittins-index-based optimal solution to so-called restless bandits -- that is, bandits whose underlying reward structures changes regardless of whether or not the decision-maker chooses to play them (Whittle, 1988).

An alternative, but related, framework for computing optimal strategies in multi-armed bandit problems was proposed by Lai & Robbins (1985). Their model quantifies the "regret" of a given decision policy as the difference between the actual reward obtained by employing that policy for the past *n* plays, and the maximum reward that could have been obtained by always playing the bandit with the highest reward. They computed a class of allocation indices for each bandit, referred to as "upper confidence indices", and offered a closed-form solution for the best possible regret that could be obtained after n plays. However, in order to compute the allocation indices, their strategy required the entire sequence of past rewards for each bandit, and the computational became very high as the time horizon increased. Auer and colleagues (2002) showed that it was possible to extend and improve this framework by setting better upper confidence bounds (UCB) on the regret. Their policies showed equal or better performance, were much faster and simpler to implement even for high numbers of bandits, and could be extended to account for non-stationary bandits (Auer & Ortner 2010).

One problem with all these strategies that rely on calculating allocation indices is that their highly intense computational requirements make them imperfect candidates for real organisms' decision mechanisms. This mirrors a similar problem in learning comprehensive representations of the environment, described in section 1 of this chapter: algorithms that operate over the entire space of available information are far too complex and time-consuming to make good candidates for actual human strategies. It is here that cognitive psychology and neuroscience work diverged from theories inspired by the fields of economics, statistics and machine learning: as interest in the question of human performance on bandit tasks grew, it became increasingly more important to develop solutions to the problem that could be implemented in a fast, efficient manner and that afforded psychological plausibility. Yi, Steyvers and Lee (2009) used a particle filter model to solve the multi-armed restless bandit problem. Unlike the index-allocation policies described above, particle filter models, a class of algorithms closely related to sequential Monte Carlo methods, are simulation-based rather than explicit closed-form solutions to the bandit problem[1]. They have gained popularity in recent years, as part of a class of solutions that relies on Bayesian estimation of the bandit values (Doucet, de Freitas & Gordon 2001). Particle filters are particularly useful for restless bandits because they employ a sequential on-line inference method, that is, they estimate and update values based on prior observations and incoming available data. This method relies on having a set of beliefs about the environment ("particles") at each decision time-step, then updating the current set of particles based on incoming data. Particles that describe the environment well propagate through the decision process, while particles that don't get replaced. Overall, the set of all current particles describes an estimate of the underlying bandit generative processes, and, because only one set of particles needs to be maintained

[1] Particle filter models bear certain similarity in implementation to genetic algorithms, and indeed, recently, particle filter/genetic algorithm hybrids have been proposed (such as the ‘genetic filter’ Park et al., 2007); but the two types of strategies are generally considered different, most notably due to their sampling mechanisms: PFs have uniform resampling and don’t usually dynamically adapt the number of samples, whereas GAs ‘crossover’ and ‘mutation’ operations make for more varied sampling. (Kwok , Fang & Zhou 2005)

at a given time, the computational requirements are far lower than in other sequential or explicit methods (Yi, Steyvers & Lee 2009).

Animals Can Be Optimal Explorers on Bandit-Like Tasks

So far, this section has presented the bandit problem, and described several different theoretical approaches to obtaining optimal solutions in multi-armed bandit problems. One immediate question that follows the optimality discussion is: can animals (including humans) actually optimally solve this type of problem? Krebs, Kacelnik and Taylor (1978) offer one early answer to this question, in a study that showed that foraging birds' behavior in a simulated two-armed bandit problem approximated the relative exploration - exploitation rates consistent with an optimal Gittins policy. Similar results were found in human studies as well: Steyvers, Lee and Wagenmakers (2009) showed that humans could approximate an optimal model on a multi-armed bandit task. However, not all participants' behavior on their task was consistent with optimal performance, leading the authors to suggest that perhaps different individuals have different perceptions of the task structure. This again relates back to the representation learning framework described in section 1: depending on what participants learned about the task environment, their exploration/exploitation patterns could easily lead to suboptimal performance if their representations did not coincide with reality.

Bandit problems have received significant attention over the past few decades, and they have been studied both theoretically, in fields such as economics, statistics and machine learning (Gittins & Weber 1989; Macready & Wolpert 1998; Tokic 2010), and empirically in psychology and neuroscience studies (Daw et al. 2006; Steyvers et al.

2009; Auer & Orten 2010). The theoretical work on optimal solutions to this problem has allowed for a more informed examination of the question of how real organisms negotiate time allocation and the exploration - exploitation trade-off. The theory described in this section provided several potential starting points for answering this question; the next section describes a series of cognitive neuroscience studies that propose various strategies for modeling human exploratory behavior.

Modeling Human Exploration - Exploitation Behavior on Bandit Tasks

The bandit problem described in the previous section is psychologically interesting because it captures the tension between exploration and exploitation present in many real-life decision-making contexts. Decision-makers must satisfy their goal of obtaining rewards, which requires exploitation, while simultaneously trying to learn about the available alternatives, which requires exploration. (Zhang, Lee & Munro, 2009). Studying human performance on bandit problems addresses several questions of interest, including how people search for information, how they incorporate the information they get, and how they adjust their behavior to achieve their goals.

There are countless empirical studies examining human behavior in bandit problems, and a variety of proposed choice models that might account for the way people allocate time and responses in complex environments. These various models differ in the way they incorporate past choices and outcomes into current decisions: they might assume that people maintain updated reward rates for all options (Tennenbaum, Griffiths & Kemp 2006), show graded forgetting of past experiences (Pan, Schmidt & Wickens 2005), show strong trial-by-trial effects (Li, Levi & Klein 2004), and more (Steyvers, Lee

& Wagenmakers 2009). The models may also differ in their treatment of exploration, assuming different sources and mechanisms (such as internal random noise, externally-generated noise, or directed information-seeking), and different underlying functions to describe exploration.

ε-greedy and softmax: two popular choice models of human exploration

One choice function frequently used to model exploration - exploitation behavior is known as the ε-greedy algorithm (Sutton & Barto 1998). This model assumes exploration is undirected -- that is, the decision-maker exploits the option likely to be the most rewarding most of the time (with probability 1–ε), while occasionally exploring a completely random choice (with probability ε/n, where n is the number of available choices; see figure 1c), according to the following rule:

$$P(choosing B_i) = \begin{cases} 1 - \frac{\varepsilon}{n}, if B_i isthemostrewardingbandit \\ \frac{\varepsilon}{n}, otherwise \end{cases} \quad (1)$$

This algorithm can capture an exclusive strategy (if the value of ε is zero), or a fully random strategy (if the value of ε is n/(n+1)). What it cannot do, however, is modulate exploration differently for the different options.

A different choice model, known as the softmax rule (equation 2, figure 1d), allows for separate exploration patterns for each option, by assuming that people choose each option proportionally to its relative reward. This algorithm is conceptually similar to a matching strategy: the relative response rate on each option is proportional to the relative reward of that option, scaled by a gain parameter. This gain parameter β essentially specifies the degree to which we balance exploration with exploitation. A gain value of 0, for instance, would mean that the decision-maker is equally likely to choose any of the available options. The larger the value for β, the more the decision-maker is biased toward exploitation.

$$P(choosingB_i) = \frac{e^{\beta V_N}}{\sum_i e^{\beta V_i}} \qquad (2)$$

Both the ε-greedy and the softmax choice functions have been widely used to model human exploration - exploitation behavior, though in multi-armed bandit problems, the latter has been found to afford more flexibility and thus be a better choice for human data (Daw et al. 2006), and so it is the choice function that I will be using to model human data in the following chapters.

Information-Sampling and Exploration: Exploratory Decisions under Uncertainty

It is reasonable to think that people's strategies for balancing exploitation and exploration are based on an interplay between reward magnitude and uncertainty (Caraco 1980; Preuschoff, Bossaerts & Quartz 2006; Daw et al. 2006), as exploitation is aimed at maximizing reward, while exploration is a potential tool for minimizing uncertainty by sampling less-rewarding alternatives for information. Recently, an important body of work inspired by the economics literature has emerged in cognitive neuroscience to investigate how uncertainty regulates exploratory behavior, and how it relates to the neural underpinnings of decision-making on this type of problem. The following section discusses two frequent sources of uncertainty – risk and ambiguity – their impact on human decision-making, and their potential interaction.

Risk and Exploratory Decisions

The intuitive idea that people prefer higher rewards to lower rewards does break down under certain scenarios. An investor choosing between opening a simple savings account (low expected reward but a known outcome) or investing the stock of a new company (higher expected reward but an uncertain outcome) might prefer the option with the lower expected reward. This suggests that, beside reward magnitude, there are other factors at play in people's decisions. In this particular example, the factor that modulated the investor’s preferences was risk.

The impact of risk on people's decisions has been studied extensively in the economics literature for many decades. Early accounts of decision-making under risk relied heavily on utility theory (von Neumann & Morgenstern 1944), calculating the utility of choosing an option as the sum of each possible outcome of that choice, weighed by the probability of the outcome. This utility framework broke down under certain scenarios: people seemed to prefer certain options to risky options, even if the weighted sum of the rewards was equal, or even when it was worse for the certain options; this lead researchers to propose that people underweight risky outcomes (Kahneman & Tversky 1986; Abdellaoui et al. 2007). The overwhelming finding from the risky decision-making literature is that people are generally risk averse. This was found to depend, to some extent, on the individual: Pratt (1964) offers a utility model of risky choice that first defines the idea of one agent being more risk-averse than another; later studies confirm that risk preferences differ among individuals (Wolf & Pohlman 1983; Mahoney et al. 2011). Overall, however, risk aversion has been a robust finding that spans decades of economic and psychological experiments (Bossaerts and Plott, 2004; Holt and Laury, 2002; Preuschoff, Bossaerts & Quartz 2006)

Risk-aversion was first linked with the issue of exploration and exploitation in an important ecology paper (Caraco 1980) that proposed that the degree of risk-aversion depended on the structure of the decision problem. Caraco suggested a foraging model that not only predicted that time allocation on an alternative would increase with reward magnitude, but that the amount of variance (or risk) in the reward would also play a role. He argued that different environmental conditions (such as scarcity) could generate different reactions to increased variability of the environment: resource-rich environments, for instance, would lead to increased risk-aversion, while resource-poor environments would make the decision-maker increasingly less risk-averse, as its need to obtain reward became more acute. This was an early model that incorporated sensitivity to higher variance into the decision strategy by choosing a utility function that was consistent with either risk aversion or risk-seeking, depending on the task environment.

Ambiguity and Exploratory Decisions

One prevalent interpretation of the widely-observed risk aversion is that people do not like uncertain outcomes (Preuschoff, Bossaerts & Quartz 2006). However, uncertainty in decision-making can also take another form, different from risk. That form is ambiguity. While in a risky choice, the decision-maker knows the possible outcomes, and their probabilities, and the uncertainty stems from the stochastic nature of the choice, in ambiguous choices, the decision-maker does not know the probabilities of the multiple possible outcomes (and sometimes, does not even know all the possible outcomes -Hsu et al. (2005); Bach, Seymour & Dolan (2009)). Ambiguity makes it impossible to directly calculate the expected reward associated with a choice, because the outcome probabilities are unknown. This introduces a kind of uncertainty that is different from risk, a second-order uncertainty (Bach et al. (2011)).

Elsberg (1961) showed that ambiguity poses a challenge to classical choice theory, by showing that, when faced with uncertainty in their environment, people preferred risky outcomes to ambiguous outcomes, to such an extent that they made series of contradicting choices that violated the axioms of rational choice. Similar findings were replicated in later studies (Becker and Brownson, 1964; Curley and Yates 1985; Camerer & Weber 1992), leading to a general agreement in the economic literature that people are ambiguity averse. However, posing the problem in an exploration - exploitation framework reveals a different perspective: if the tension between information-seeking and reward-seeking is present, the adaptive strategy might in fact bias the decision-maker toward the ambiguous option, rather than away from it, as choosing the ambiguous option would in fact gain more useful information (Daw et al. 2006; Cohen, McClure & Yu 2007).

Under this framework, the key feature that would shift people from being ambiguity-averse to being ambiguity seeking is the utility of acquiring information. In single-trial tasks (all the economics papers mentioned above have this structure), there is no benefit from acquiring information from the ambiguous options. In a bandit problem, however, due to the sequential nature of the task, acquiring information about ambiguous bandits can in fact be beneficial, as it leads to more informed choices later on (Meyer & Shi 1995). Therefore, in a task involving sequential decisions, it might in fact be adaptive to explore ambiguous options (Cohen, McClure & Yu 2007).

As first hinted at by Caraco (1980), it is very likely that people incorporate different sources of uncertainty differently into their decision processes, which may bias them either toward exploration (seeking out uncertain options to gain information) or away from it. All the findings discussed above constitute ample evidence that there are different potential mechanisms at work in human exploration and that both risk and ambiguity play important roles. Human risk- and ambiguity-preferences are clearly differentiable (Hogarth & Einhorn 1990; Bossaerts et al. 2009; Levy et al. 2009), but the interplay between risk and ambiguity when it comes to the exploration-exploitation tradeoff has been mostly studied in the domain of organizational behavior (March 1991; Gupta, Smith & Shalley 2006), or financial decisions (Uotila et al. 2009). In cognitive neuroscience, the interaction of risk and ambiguity has been examined primarily in single-shot economic games with no notable exploration component (Huettel et al. 2006; Hsu et al. 2009; Levy et al. 2009).

The work presented in the chapter 3 addresses some of these outstanding questions, and examines in depth people's exploratory decisions under uncertainty, focusing both on the different mechanisms of exploration employed by humans playing a two-armed bandit task, and on shedding further light on the interplay of risk and ambiguity in exploratory decisions.

1.3. Information and Task Engagement: Exploration as an Adaptive Response to Boredom

Failures of task engagement: Boredom, from an affective to a cognitive phenomenon

Let us call for a moment on an earlier example: imagine you are driving home from the football game in a nearby city. Now imagine that you've been in the car for a while, and to pass the time as you make your way through highway traffic, you've been listening to the radio. It's a station you like, and it plays songs you're familiar with and enjoy listening to. Yet, after fifteen, twenty, twenty-five minutes of listening to the same station play the same kind of songs, you might feel like pressing the tuner button on your radio, and letting it browse other stations. This is not because you've suddenly begun to dislike rock or classical music or whatever the station is playing – rather, after a long time listening to the same variety of music, you simply want to listen to something else.

From an economic point of view, the decision to press the tuner button and browse away from your favorite station might strange. You are actively going against your established preferences, and foregoing known benefits – a station that plays songs you *know* you will like – in favor of letting the radio cycle through a series of other options of unknown or questionable value. Yet if an economist walked in right now and asked you why you are making such a seemingly irrational choice, you would probably have no trouble answering. You were bored of listening to the same kind of songs for thirty minutes, and you wanted to listen to something else.

The notion of boredom might seem straightforward and intuitive to anyone who has ever had to take long drives, fill out taxes, wait in line at busy registers, or sit through two-hour business meetings, but it has intrigued economists, psychologists and ecologists alike for more than a century. Boredom is undoubtedly real and widely observable throughout history (as early as 1890, William James makes reference to a notion of "Tædium, ennui, Langweile, boredom… words for which every language known to man has an equivalent"), and across cultures (Sundberg et al. 1991, for instance found similar boredom behaviors in young adults spanning four continents). But despite its prevalence, little is understood about the origin and function of boredom, or about the mechanisms that underlie it.

Historically, boredom has been studied in the context of affective disorders because of its role in anxiety, neurosis and depression (Bergler 1945, Fenichel 1951). Studies have linked the subjective experience of boredom, as well as a "boredom proneness" character trait, to personality and mood disorders (Vodanovich, Verner & Gilbride 1991), addictive behaviors (Blaszczynski, McConaghy & Frankova 1990), and even physiological symptoms (Farmer & Sundberg, 1986; Sundberg et al. 1991). In this clinical context, boredom is associated with negative outcomes such as higher school dropout rates, low achievement, and job dissatisfaction (Drory 1982, Watt & Vodanovich, 1999). This view, however, is increasingly being complemented by a more cognitive perspective. As the behavioral correlates of boredom easily lend themselves to cognitive interpretation (in the above example, for instance, getting bored while filing the reimbursement forms can affect a series of cognitive processes: we might pay less attention to the task, take longer to do it, make more errors, or switch frequently between tasks), the drive has emerged to investigate the impact of boredom on human cognitive processes.

Building on earlier work that correlated boredom with cognitive demand (London & Monello 1974; Pattyn et al. 2008), task monotony and difficulty (Hill & Perkins 1985) and effort to engage with the environment (Hamilton 1981; Harris 2000), a series of recent studies interpret boredom as a cognitive construct rather than a primarily affective state indicative of psychopathology. Notably, Eastwood et al. (2012) frame boredom as a failure to engage with the environment due to insufficient arousal (for both internal and external reasons), and suggest that the phenomenon is best examined in the context of attention, attentional failures, and executive control. Through this cognitive interpretation, boredom could be an adaptive mechanism, as some of its behavioral byproducts – such as higher distractability, the tendency to give up sooner, or increased randomness in performance (Wallace, Vodanovich & Restino 2003, Watt & Hargis 2010) – that are generally considered negative, might in fact be useful.

Exploration as a behavioral consequence of boredom: A potentially adaptive response to insufficient information?

The tendency to switch away from a task and increased randomness in performance are both good examples of behavioral correlates of boredom that might serve an adaptive function. As discussed in detail in section 1.2, the decision-making literature has established that making random choices, can be useful in a number of ways. When operating in an environment about which we do not have full knowledge (as is the case with most real-life situations), exploring it by occasionally selecting random options can help us discover better strategies and often lead to more overall reward (Cohen, McClure & Yu 2007). Giving up on a task can also be adaptive, if the benefits we gain from the task have fallen below the costs involved in doing it (Wrosch et al. 2003) – and indeed it has been hinted that boredom might be a way to signal increasing opportunity cost (Charness, Kuhn & Villeval 2012; Kurzban et al. 2013).

This notion of an adaptive role for boredom has been mentioned previously in the reinforcement learning literature by Schmidhuber (1997), who proposed a learning model that included a "boredom unit" that computed estimated future change in prediction errors in order to track how much information was left to learn. This unit penalized the value of an action if the sum of future prediction errors was small – meaning that learning had reached an asymptote. According to Schmidhuber's proposed model, increased boredom occurred when there was little possibility for learning more from the current action – either because it had been fully learned, or because it was too random to allow learning – and it led to an increased probability of switching away.

This framework suggests that when we become bored with our current circumstances, we may in fact be driven to explore our environment more, abandon options that have become unsatisfactory and perhaps discover better strategies for gaining reward. Based on the existing theories about boredom arising as a consequence of insufficient information, the increased exploratory drive in this case could reflect that the information content of current option has fallen far enough that it is time to switch away. In some ways, this is similar in concept to the marginal value theorem in the foraging literature (Charnov, 1976). The MVT proposes that a foraging agent chooses when to leave a current patch and explore others based on a comparison between its current rate of reward from the patch it's in, and an estimate of the average (global) rate of all available patches. When the agent's current reward dips below the overall rate for the environment, MVT dictates that it should leave and explore another patch, since the current one is not as valuable as it could be. The exploration that occurs as a consequence of boredom could be adaptive in a similar way, if the agent were tracking the amount of useful information derived from the current task, and decided when it was time to seek a new, richer source of information.

A strategy close to this idea was proposed in the machine learning literature by Simsek & Barto (2005), who showed that simulated agents with the capacity for becoming bored were better learners in a complex grid world. Their "bored" agents explored the artificial environment discovering how certain actions led to certain outcomes, but after being exposed to the same action-outcome contingency too many times (i.e., that contingency had been learned well and there was no more information to be gained about it), they began to devalue it and specifically explore other actions. By this mechanism, they were able to discover more complex action sequences faster than agents who did not have this capacity for boredom, and their average reward rate was higher.

These findings all indicate that not only can boredom be seen as an important component in a learning and decision-making task, but it can in fact serve a useful function. However, to date, these kinds of results have been proposed only theoretically (Schmidhuber 1997; Simsek & Barto 2006; Eastwood et al. 2012; Kurzban et al. 2013), and the precise link between information and boredom has not yet been studied in humans, nor has the connection between boredom and exploration been precisely pinned down. Chapter 4 of the present work discusses several experiments investigating how information structure affects the perception of boredom in humans, how this perception of boredom depends on the other options available in the environment, whether boredom signals a decrease in informativeness of the current task and how exploration could be seen as an adaptive response to that signal.

Chapter 2: Manipulation of Available Information Impacts Human Representation Learning

To study the computational processes by which people learn the relevant features in their environment, the present work proposes a novel method of *causal model comparison*. Participants played a probabilistic learning task that required them to identify one relevant feature among several irrelevant ones. To compare between two models of this learning process, I ran each model alongside the participant during task performance, making predictions regarding the values underlying the participant's choices in real time. To test the validity of each model's predictions, I used the predicted values to try to perturb the participant's learning process: I crafted stimuli to either facilitate or hinder comparison between the most highly valued features. A model whose predictions coincide with the learned values in the participant's mind is expected to be effective in perturbing learning in this way, whereas a model whose predictions stray from the true learning process should not. Indeed, results showed that in our task a reinforcement-learning model could help or hurt participants' learning, while a Bayesian ideal observer model could not. Beyond informing us about the notably suboptimal (but computationally more tractable) substrates of human representation learning, this manipulation suggests a sensitive method for model comparison, which makes it possible to change the course of people's learning in real-time.

Introduction

We live in a rich, complex environment, in which we are constantly bombarded with a wide variety of sensory input. Even an action as simple as walking down the street carries with it a large volume of low-quality information in the form of people we see, places we walk by, cars, colours, voices, noises, emotional content etc. Intuitively, one would imagine that given sufficient resources, it is best to always represent every aspect of the environment so that any detail can potentially be acted upon. However, the "curse of dimensionality" (Bellman, 1957) posits that task representations that involve unnecessary stimulus dimensions will not afford efficient learning and decision making, where efficiency is measured in the number of examples needed to learn the task. In particular, an increase in the number of dimensions of the problem (in our case, the dimensions of the environment that the brain may represent) implies that the learner needs to collect exponentially larger quantities of data to learn to solve the problem. If we want learning to be feasible it is therefore both computationally optimal and a practical imperative to represent tasks with as compact a representation as possible.

What are the computational strategies that humans use to learn a representation for a given task? To address this question, I tested participants on a multidimensional trial-and-error choice task, in which only one dimension was relevant to predicting reward (Wilson & Niv 2012, Niv et al., 2015). To test the explanatory power of different models of learning dynamics, I developed a method that compares two models against each other in terms of their causal effects on behaviour. Specifically, I used each model to manipulate participants' learning in real time, and asked which model was more effective in changing behavior. This is at the same time an intuitive measure of how well a model captures participants' strategies, and it constitutes evidence that it is possible to use model predictions to impact learning in real-time, by manipulating the stimuli that are presented to the participant.

It goes without saying that trial-and-error learning depends on what information is available to the learner. Indeed, work in machine learning and information theory has established how information in any given task might be optimally selected so as to maximally discriminate between competing hypotheses and accelerate learning (optimal experimental design, Sebastiani & Wynn 2000). Although human learning does not always mirror these optimal strategies, judicious choice of information has been shown to improve learning, for instance of category boundaries (Gureckis & Markant 2012) or speech motor learning (Knock et al 2000). Moreover, the order in which information is presented is relevant to determining what is learned (Ritter et al 2007). I thus set forth to use our candidate models to manipulate the timing and availability of information in such a way as to aid or hinder participants' learning trajectory.

This kind of effort to manipulate learning, however, is heavily dependent on having a good model of how participants structure and update their representations of the environment. How to compare and select a 'best' model for a complex cognitive process is not trivial (Pitt, Myung & Zhang 2002, Cutting et al 1992): models that fit the data better on some common goodness-of-fit measures may not fit better on other such measures; models that posit very different processes may perform similarly in terms of average fit (Townsend 1990, Rust et al. 1995); and a model that seems to describe behaviour better might do so because of a more flexible function form or different numbers of parameters, and not necessarily because it better captures the underlying cognitive processes (Busemeyer & Wang, 2000). I therefore developed an *interventional* method for model comparison.

I used the two candidate models to predict in real time what hypotheses a participant might be testing, and to design stimuli that will make it easier or harder to distinguish between the competing hypotheses. The reasoning was that a model that does not capture the participants' beliefs about the available stimuli would not be effective at such a manipulation of learning. In contrast, a model whose predictions are well-matched to the cognitive processes underlying participants' behaviour should make it possible to manipulate learning in real time. Therefore, this work shows not only how differential information presentation can significantly alter what representation of the environment people learn, but also that it is indeed possible to examine people's underlying learning mechanisms, by presenting them with helpful or unhelpful information based on representations inferred using different models, and testing which manipulation affects their learning.

In particular, while participants played the multidimensional probabilistic learning task, I inferred their value representations in real-time using either a Bayesian or a reinforcement learning model. The comparison of these two models relates to a long-standing question in cognitive psychology, namely, the extent to which humans resort to optimal Bayesian decision strategies, compared to suboptimal – but more intuitive and computationally more efficient – heuristics (Steyvers, Lee & Wagenmakers, 2009). Despite much work suggesting that the human brain is Bayes-optimal (Körding & Wolpert, 2004; Beierholm et al. 2009), and in line with our previous findings (Niv et al., 2015), the manipulation here was only effective when based on predictions of the reinforcement learning model.

Our ability to manipulate the learning process both precisely and in real time consists of a proof of concept for the new proposed model-testing tool, and is a step in the right direction in terms of development of individualized tools to improve learning in general.

Methods

Participants

25 participants (16 females) recruited from the Princeton University undergraduate community gave informed consent and were compensated $12 an hour plus a performance bonus of up to $5 depending on their final score in the task. The average pay was $15. Study materials and procedures were approved by the Princeton University Institutional Review Board.

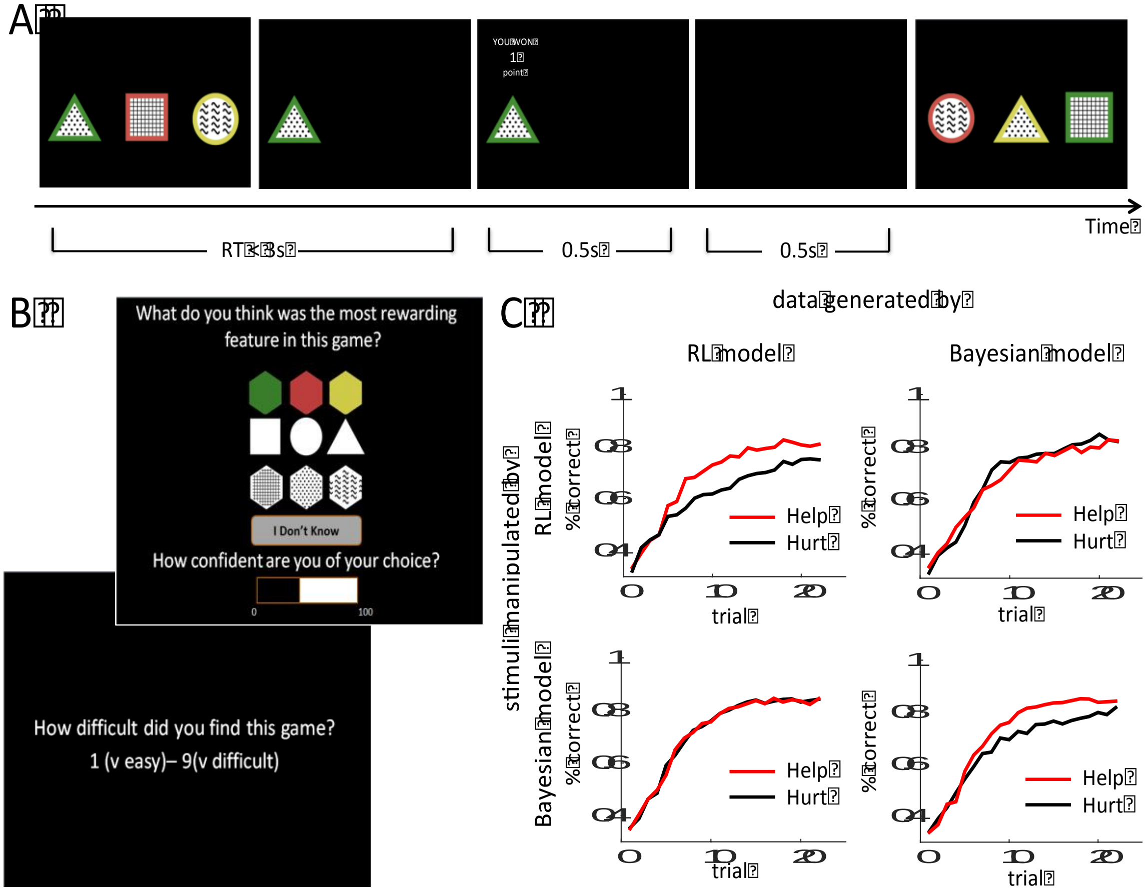
Task

Participants played a probabilistic learning task. Each trial involved choosing one of three compound stimuli displayed on the screen (see figure 2a). Each stimulus comprised of three features defined on three dimensions: a color (red, yellow or green), a shape (triangle, square or circle), and a texture (dots, plaid or waves). No two stimuli could share the same feature, i.e., there was only one red stimulus, one triangle etc., per trial. Choosing a stimulus resulted in immediate feedback, in the form of either one or zero points. Participants were instructed to try to obtain as many points as possible.

The task was designed so that, of the three dimensions of a stimulus, only one dimension was relevant to determining reward. Within that dimension, one feature was the target feature—choosing the stimulus that contained this feature led to one point 75% of the time, and zero points 25% of the time. Choosing any other stimulus resulted in one point only 25% of the time. To maximize their score, participants therefore had to aggregate over previous choices and outcomes to learn which feature is the target feature. Participants were explicitly instructed about these aspects of the task structure in advance.

The task consisted of 52 'games'. Participants were informed that the target feature would not change within a game, but would change between games. Ends of games were explicitly signalled on-screen. The first 12 games (referred to as the baseline phase, described below) included 30 trials each, while the remaining games (the manipulation phase) consisted of 36 trials each, for a total of 1800 trials. After each game, participants were asked how difficult they found the game (on a scale from very easy (1) to very difficult (9); figure 2b), and to identify the target feature in that game. They could select any of the nine features, as well as an "I don't know" option. If they did select a feature, they were also asked to rate how confident they were about their choice.

After the baseline phase, participants took a one-minute break, during which I used their baseline phase data to fit the free parameters of the two candidate models I would later test in the manipulation phase. The remaining 40 games of the task comprised of the manipulation phase, in which I manipulated stimuli according to predictions from each model, to either help or hurt participants' learning (see below).



What do you think was the most rewarding feature in this game?

How confident are you of your choice?

How difficult did you find this game?
1 (v easy)– 9(v difficult)

Figure 2: The Dimensions Task. A: *Example trial - stimulus presentation, choice, positive feedback, new stimulus presentation.* ***B:*** *Query screens - difficulty ratings, identifying correct feature, confidence ratings.* ***C:*** *Simulations of model-based manipulation. The manipulation was effective ("Help" improves performance as compared to "Hurt") only when stimuli were manipulated according to predictions from the same model that generated the choice data (top left, bottom right panels).*

Modelling

The two models compared here represent two different ways of thinking about human representation learning. The first is a Bayesian model that assumes statistically optimal updating of the posterior probabilities of each feature being the target feature, and the second model uses reinforcement learning principles to update values via trial and error.

Both models compute the value of a compound stimulus by estimating values of individual features and combining them: the Bayesian model estimates, for each feature, the posterior probability that it is the target feature, while the reinforcement-learning model learns the values of each feature based on prediction errors. In both models, current values of stimuli depend on the history of choices and rewards.

Bayesian Model

The Bayesian model tracks the posterior probability that each feature *f* is the target feature *f**. At the end of each trial, the posterior is updated by combining the prior (i.e., the posterior from the previous trial) and the likelihood of the observed data if *f* were the target feature. The prior depends on the history of choices *C* and rewards R, $D_{1,t-1} = \{C_{1:t-1}; R_{1:t-1}\}$, from the beginning of the game and up until the current trial (not inclusive). The likelihood depends on the reward probabilities imposed by the experimenter; for instance, the likelihood of a win if the chosen stimulus contains the target feature is 0.75.

At the beginning of the game, the prior is initialized at 1/9 (all features are equally likely to be the target feature). After each trial, the posterior is updated according to:

$$P(f = f^*|D_{1:t}) \propto P(R_t|f = f^*, C_t)P(f = f^*|D_{1:t-1}) \quad (1)$$

The value of a stimulus S is then calculated as the probability of obtaining a 1 point reward for choosing that stimulus on the current trial t,

$$V(S) = P\left(R = 1\middle|S, D_{1,t-1}\right) = \sum_{f\in S}\left[P(R = 1|f = f^*)P\left(f = f^*\middle|D_{1,t-1}\right)\right] + P(R = 1|f^* \notin S)\left(1 - \sum_{f\in S} P\left(f = f^*\middle|D_{1,t-1}\right)\right) \quad (2)$$

where $P(R = 1|f = f^{*}) = 0.75$ for all features contained in S, and $P(R = 1|f^{*} \notin S) = 0.25$. The model can be considered an ideal observer because it maintains a full probability distribution over the identity of f* and updates this distribution in a statistically optimal way.

To afford this model the same temporal locality as the reinforcement model (described below), this model also allowed some degree of decay for all feature posteriors toward a uniform value of 1/9

$$Ᵽ(f = f^{*}) = (1 - d)P(f = f^{*}) + d * 1/9 \tag{3}$$

and used Ᵽ instead of P in equations (1) and (2) above. Although suboptimal, the decay component has been shown to significantly improve the models' fit to behavioural choices in our task (Niv et al. 2015).

Finally, the model assumed that the probability of choosing stimulus S_i on each trial is proportional to the value of the stimulus, according to the softmax probability choice function:

$$P(chooseS_i) = \frac{e^{\beta V(S_i)}}{\sum_{j=1}^{3} e^{\beta V(S_j)}} \tag{4}$$

The positive-valued inverse-temperature parameter β sets the level of noise in the decision process, with large β result in near-deterministic choice of the highest value option, while small β result in high decision noise and more random decisions. In all, this model has two free parameters, β and the decay rate d.

Reinforcement Learning Model

This model takes advantage of the fact that, in our task, features determine reward independently, and uses reinforcement learning to learn values for each of the nine features. The values of all features were initialized at zero at the beginning of each game; on each trial, values of the three features of the stimulus that was chosen were then updated according to

$$V_t(f) = V_{t-1}(f) + \eta\big(R_t - V_{t-1}(S_{chosen})\big)\forall f \in S_{chosen} \tag{5}$$

where η represents the learning rate, and $\big(R_t - V_{t-1}(S_{chosen})\big)$ is a prediction error – the discrepancy between the actual reward on the current trial and the reward that was expected based on choosing this stimulus.

Based on previous findings (Niv et al., in press), this model also included a decay of the values of the unchosen features to zero:

$$V_t(f) = (1-d)V_{t-1}(f)\forall f \notin S_{chosen} \tag{6}$$

with *d* a free parameter determining the decay rate. As in the Bayesian model, action probabilities were determined by a softmax probability function on stimulus values (equation 4). The RL model thus had three free parameters, the learning rate η, the softmax inverse temperature β and the decay rate *d*.

Model Fitting

At the end of the baseline phase (described above), participants were given a one-minute break while the computer fit both models to their data. Free parameters of both models were set for each participant separately, and were selected so as to maximize the likelihood of the data from the baseline phase (12 games and a total of 360 trials), by using the Matlab routine fmincon and fitting the data five times, initializing at different random starting points. Parameter values from the run that obtained the best likelihood for the data were then used to fully specify the model for the manipulation phase of the experiment, in which each model was used to track feature values in real-time.

Real-Time Manipulation

In the remaining forty games of our task, the aim was to help or hurt learning by manipulating stimulus presentation. Specifically, I manipulated how the available features (colors, shapes and textures) were combined into three different stimuli, so as to either make available or obscure information about which feature is more likely to be the target feature. This manipulation relies on the idea – common to both candidate models – that while playing the task, participants update values for each feature, with the goal of ultimately learning which is the most rewarding feature. These feature values carry predictions regarding the reward associated with each feature, and thus can be seen as 'hypotheses' as to which is the target feature.

For every manipulated trial, to help learning, the highest-valued feature in each dimension was selected, and each of the three stimuli presented to the participant was designed to include only one of these three highest-valued features. This manipulation facilitates hypothesis testing, allowing participants to test one high-value feature independently of the other two. Conversely, to hurt learning, the three highest-valued features (one in each dimension) were combined into a single stimulus, thus preventing the participant from assigning credit for the feedback to just one of the three competing features. Both types of manipulation can potentially impact the rate of learning in the game, but only as long as the inferred values are close to the participants' actual values (Figure 2c).

In each manipulated game, I manipulated every other trial from the fourth to the thirtieth trial for a total of fourteen manipulated trials, using one of the two models and either helping or hurting learning throughout the game. Because this manipulation affects not only learning, but the likelihood to make the correct choice on the current trial, to measure learning I analysed choices only in non-manipulated (neutral) trials, in which stimuli were constructed so as to specifically not include all three highest-valued features in the same stimulus, nor separate them into three different stimuli. (Therefore, in non-manipulated trials, one stimulus always contained exactly two of the highest-valued features –and these trials did not overlap with either the helping or the hurting manipulation.)

Each of the forty games in the manipulation phase consisted of thirty-six trials, with fourteen manipulated and twenty-two neutral trials. The last six trials in each game were not manipulated so as to allow measurement of steady-state learning at the end of the game.

Results

To understand the dynamic process of learning a compact and sufficient task representation in a multidimensional environment, I tested human participants on a probabilistic 3-dimensional choice task. While they played the task, I used the real time trial-by-trial predictions from two competing models to manipulate the presented stimuli so as to help or hurt learning. Success of the causal manipulation would attest to congruence between the model and participants' learning strategies.

As shown in Figure 3, the manipulation had a significant effect on learning only in those games in which stimuli were manipulated using the reinforcement-learning (RL) model (figure 3a), and did not alter learning in games in which stimuli were manipulated using the Bayesian model (figure 3b). For RL-manipulated games, average performance on the last six trials of games (all non-manipulated) in the Help condition (red line) was significantly better than in the Hurt condition (black line).

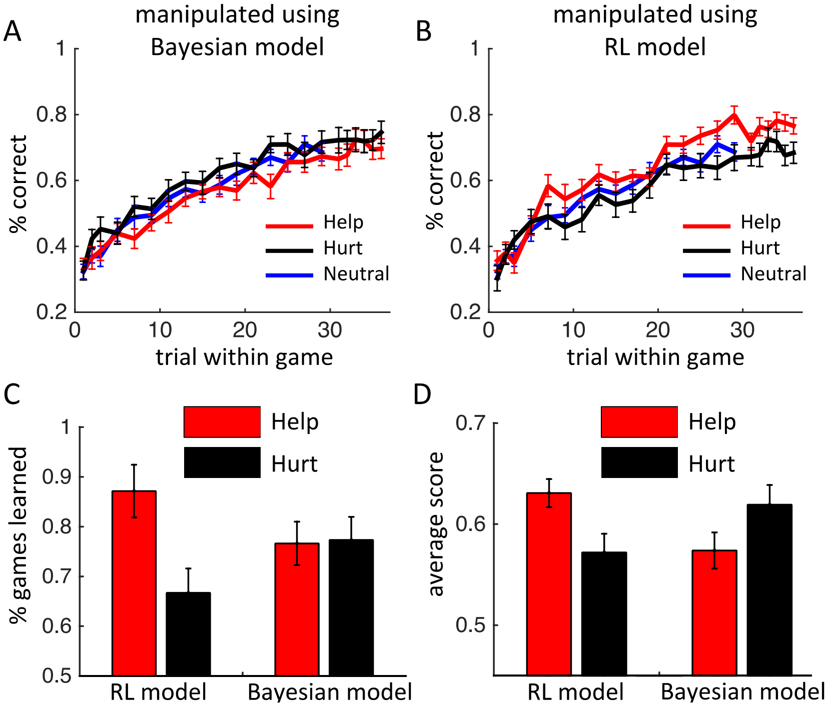


***Figure 3:** Model-based manipulation of stimuli affects learning only when using predictions from the RL model, not from the Bayesian model. **A:** Learning curves for Help (red) and Hurt (black) conditions overlap when the manipulation used predictions from Bayesian model. **B.** When the manipulation used predictions from the RL model, learning curves for the Help condition show significantly better learning at the end of the game as compared to the Hurt condition. Blue line:data from the baseline phase. Similar effects of the manipulation are also seen in **C.** the overall number of learned games and **D.** the average score across games. Error bars: SEM.*

A two-way Performance x Manipulation repeated-measures ANOVA yielded a significant interaction (*F*(1, 48) = 5.22, *p* = 0.02) with no main effects, and the average performance at the end of the game was significantly higher in the Help than the Hurt condition for RL-manipulated games (one-sided paired t-test, *t*(24) = 2.25, *p* = 0.01), but did not differ between conditions for the Bayesian-manipulated games (*t*(24) = -0.96, *p* = 0.82).

Participants did not learn all the games, that is, for some games, they could not identify the correct feature when probed at the end of the game (Figure 2b). The real-time manipulation had an impact on the total number of games learned (figure 3c): here too a two-way Model x Manipulation repeated measures ANOVA showed a significant interaction (*F*(1,48) = 7.23, *p* = 0.009) with no main effects. When the RL model was used to manipulate stimuli, the number of learned games differed significantly depending on whether the game was designed to help or hurt learning (*F*(1, 24) = 12.64, *p* < 0.01). On average, the Help condition resulted in an approximately 30% increase in the number of learned games. Conversely, when the Bayesian model was used to manipulate stimuli, there was no difference between the help and hurt conditions (*F*(1, 24) = 0.77, *p* = 0.73).

A similar pattern was observed in the average score per game (figure 3d). For games manipulated using the RL model, the average score (number of correct choices) was higher in the Help condition than in the Hurt condition (paired t-test; *t*(24) = 2.72, *p* = 0.011). In contrast, scores for the Help and Hurt conditions were similar when the Bayesian model was used to manipulate the stimuli (*t(*24) = -1.91, *p*=0.07). The impact of the stimulus manipulation was also evident in participants' difficulty (figure 4a) and confidence (figure 4b) ratings. For RL-manipulated games, participants rated ‘help’ games as easier compared to the ‘hurt’ games (*t(*21) = -2.23, *p* = 0.03, paired t-test). This effect was not present in the games manipulated using the Bayes model (*t(*21) = 0.734, *p* = 0.47, paired t-test).

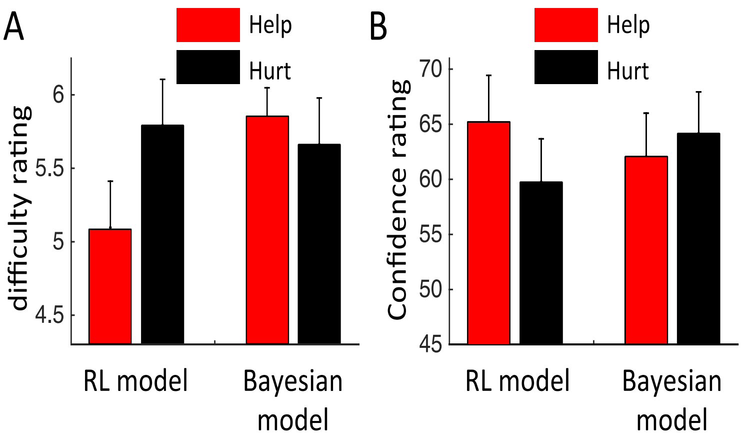


Figure 4: *Difficulty (**A**) and confidence (**B**) ratings reflect that the manipulation was effective in RL-manipulated games, not in games manipulated using the Bayesian model. Error bars: SEM. Three subjects were excluded from this analysis, two because they consistently did not rate difficulty/confidence, and one who reported reversing the difficulty scale for most of the experiment*

A similar pattern was seen in the confidence ratings (figure 4b), where confidence in RL-manipulated ‘help’ games was rated as significantly higher than confidence in the ‘hurt’ games ($t(21) = 2.57, p = 0.01$), but ratings in games manipulated using the Bayesian model were not significantly different across conditions ($t(21) = -1.26, p = 0.22$).

Discussion

Using a multidimensional choice task, the present work investigated the computational strategies by which humans determine what dimensions of the environment are relevant for obtaining reward, and which can be safely ignored. The assumption underlying this work is that naturalistic tasks require such a representation learning process: in any given scenario, only a subset of information is relevant to the task at hand, and, moreover, the particular environmental dimensions that are relevant to one task are not necessarily relevant for another. For instance, the color of cars is irrelevant for crossing the street, but relevant for hailing a taxi, and the identity of the shops across the street is irrelevant to both tasks (but of course not for the task of navigating to the coffee shop).

To compare between qualitatively different accounts of how humans may learn what dimension of the environment is relevant for the current task, I showcased a method that compares two learning models by attempting to use each model in a causal, real-time manipulation of participants' learning. That is, I used each model to predict what hypotheses participants might be testing at each point in time, and manipulated stimuli to either help or hinder comparison of these hypotheses. This method is related to the concept from educational computing of "intelligent tutors" (Beck, Wolf & Beal, 2000; Anderson et al., 2003), as it uses the same idea of feedback tailored to an individual's ongoing learning process. However, the method we used here is novel not only in the way it specifically infers a formal model of the participants' learning as they perform the task (as opposed to using their responses to measure how closely they adhere to the correct, 'expert knowledge' model, as most intelligent tutor systems do) – but it is also a novel way of doing model comparison, as this model-based manipulation can affect learning only to the extent that a model indeed captures participants' underlying learning process.

Results showed that when stimuli were manipulated based on a reinforcement-learning model, games designed to help learning resulted in faster and more complete learning than games designed to hurt the learning process. In contrast, manipulating games using a Bayesian model had no significant effect on learning. This method thus provides a stringent measure of how well each model captures people's strategies, and at the same time, the current findings provide evidence that it is indeed possible to impact representation learning in real time, by manipulating the stimuli that people have access to.

This method is related to the framework of “optimal experimental design” in which experiments are designed so as to optimally elicit information about the process under investigation (Sebastiani & Wynn 2000, Atkinson, Doney & Tobias 2007). Normative statistical principles from Bayesian inference can, in some cases, be used to select an experimental design that will best resolve the details of participants’ underlying cognitive processes (e.g., set the free parameters of a model of the process under scrutiny; Rafferty, Zaharia & Griffiths 2014). One way to optimize the present manipulation would be to choose, on each trial, the stimulus configuration that would allow participants to glean the maximum (or minimum) amount of information regarding the identity of the target stimulus, assuming participants’ underlying cognitive processes matched each of the candidate models. This manipulation, while more normative than the one presented in this work, is more computationally intensive and, importantly, relies on further assumptions regarding the optimality of participants’ actions. In particular, if participants are not selecting stimuli in an effort to maximize information (e.g., because they are also maximizing reward), this alternative manipulation may not be more effective. It is due to this interaction between information acquisition and reward acquisition that I assessed performance only in non-manipulated trials – as the ‘help’ manipulation, while affording better information, made it more difficult to obtain reward on manipulated trials than did the ‘hurt’ manipulation.

Rather than assume that the highest valued features correspond to the hypotheses that the participant is comparing, another alternative is to infer these hypotheses using Bayesian inference. Such a method has previously been used in association with the dimensions task, and shown that recent choice history can effectively identify tested hypotheses (Wilson & Niv, 2012). However, in that work, inference was only optimal if one assumed that participants test one hypothesis at a time—an assumption that is incompatible with the current results. If participants were indeed focusing only on one hypothesis (feature) at any point in time, then neither of the manipulations would have affected learning.

The current findings join others (Eckstein et al., 2004) in suggesting that human learning is not always Bayes-optimal, and in particular, that humans do not solve the difficult task of representation learning in a Bayes-optimal way (see also Wilson & Niv, 2012; Niv et al., in press). These findings stand in contrast to multiple demonstrations of Bayes-optimality (Doya et al., 2006) in perceptual decision making (Gold & Shadlen, 2002; Knill & Pouget, 2004), motor control (Trommershäuser et al. 2003, 2005) multimodal integration (Körding & Wolpert, 2004), reasoning (Oaksford & Chater, 1994) and even in setting meta-learning parameters for reinforcement learning (Behrens et al. 2007, Yu 2007). However, whereas Bayesian inference may be computationally feasible (and indeed, simple) in scenarios that can be reduced to a several-alternative forced-choice decision (Gold & Shadlen, 2002) or a choice between lotteries (Wu et al., 2009), representation learning in natural environments places much heavier computational demands on the learning system. In particular, the dimensionality of the environment is essentially unbounded (given that dimensions such as previous actions and events can be, and frequently are, relevant for task performance), and whereas feedback is often available for one’s actions, the environment does not provide any supervision regarding one’s representations.

To overcome these difficulties, results here suggested that humans might use a suboptimal but computationally more tractable strategy based on reinforcement learning. However, it is worth noting that the present work only compared two very different models, partly as a proof-of-concept for our novel model-comparison method. It is entirely possible – in fact, likely – that the feature-level reinforcement-learning model suggested here also falls short of fully capturing participants’ learning strategies. Future applications of this method will hopefully delineate more precisely the workings of representation learning in the human brain.

Finally, the “dimensions task” lends itself easily to manipulation of presented information. Work on “active learning” (Cohn, Ghahramani & Jordan, 1995) in categorization and perceptual estimation tasks has used a related manipulation, effectively allowing participants to design their experiment optimally (Kruschke, 2008; Castro et al., 2008; Gureckis & Markant, 2009; Markant & Gureckis, 2014; Juni et al., 2011). Some adjustments will likely be needed in order to apply this model-comparison method to other task structures, though there is reason to be optimistic as to the method’s wider applicability (Nelson et al. 2010).

In sum, this chapter has described a real-time manipulation of information presented to participants, and has suggested that basing this manipulation on predictions of different models can allow for a new, sensitive and *causal* means of model comparison. Using this method and a reinforcement-learning model, the work here shows that human representation learning can be improved or hampered. Beyond the implications for effective, individual-difference-sensitive model selection, such ‘access’ to participants’ mental strategies suggests exciting applications, particularly in the domain of education and tailoring the flow of information toward individual learning.

Chapter 3: Exploration Strategies and the Interplay of Decision Time, Risk and Ambiguity

In this chapter, I present two experiments that examined how information-seeking drives exploration. The first experiment[2] tested participants on a two-armed bandit task with different ambiguity conditions, under different decision horizons. Results showed that people used two distinguishable types of exploration – random (characterized by an increase in decision noise) and directed (information-driven) exploration – and that the length of the decision horizon affects both. The second experiment tested a different set of participants on a 'wheel of fortune' task, designed to clearly separate risk and ambiguity and examine how they impact the two different exploration strategies. The results suggested that the presence of ambiguity in the environment drives people to explore in order to acquire more information and reduce the ambiguity. Conversely, a higher risk level in the environment increases exploration by increasing decision noise and making people less sensitive to the reward values of the available options. These findings imply that different sources of uncertainty impact exploration differently, and may shed light on the mechanisms behind people's use of exploration to acquire information from the environment.

[2] The work I describe for this experiment has been published, in Wilson et al. (2014)– see Acknowledgments section for details, and Reference section for full citation

Introduction

When you go to your favorite restaurant, do you always order the same thing, or do you try something new? Sticking with an old favorite ensures a good meal, but if you are willing to explore you might discover something better. This simple conundrum, deciding between something you know and something you do not, is commonly referred to as the exploration - exploitation dilemma (Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). Whether deciding on a meal, a vacation destination or a life partner, this problem is an important one to solve.

All adaptive organisms face this fundamental tradeoff between pursuing a known reward (exploitation) and sampling lesser-known options in search of something better (exploration). Exploration can be most beneficial in the presence of environmental uncertainty: when the range and benefits of all reward options are not fully known, exploration can lead to the discovery of new, better resources and an ultimately higher overall reward. However, uncertainty can take many forms – most prominently, risk (known uncertainty) and ambiguity (unknown uncertainty) are two clearly differentiable sources of uncertainty that often appear together – and it is unclear how different sources of uncertainty impact people's exploratory behavior. Furthermore, theory suggests at least two strategies for solving the explore - exploit dilemma: a directed strategy in which choices are explicitly biased towards information seeking, and a random strategy in which decision noise leads to exploration by chance. The mechanisms and factors involved in each type of exploration are not yet fully known.

This chapter presents work investigating the extent to which humans use these two exploration strategies, and examining the influence of important environmental factors such as decision horizon, risk and ambiguity on human exploration.

Experiment 3.1: Humans use directed and random exploration to solve the explore-exploit dilemma

In our variant of the two-armed bandit paradigm – that we refer to as the "horizon task" – participants made explore-exploit decisions in two contexts that differed in the number of choices that they would make in the future (the time horizon). Participants either were allowed to make either a single choice in each game (horizon 1), or six sequential choices (horizon 6) giving them more opportunity to explore. By modeling the behavior in these two conditions, it was possible to measure exploration-related changes in decision making and quantify the contributions of the two strategies to behavior. Results suggested that participants were more information seeking and had higher decision noise with the longer horizon, suggesting that humans use both strategies to solve the exploration-exploitation dilemma. It was thus possible to conclude that both information seeking and choice variability can be controlled and put to use in the service of exploration.

Methods

Participants

31 Participants (20 women; mean age 19.7 years, min: 18 years, max: 24 years) were recruited from the Princeton student population. They received course credit for taking

part in the experiment and gave informed consent; the study was approved by the Princeton University Institutional Review Board.

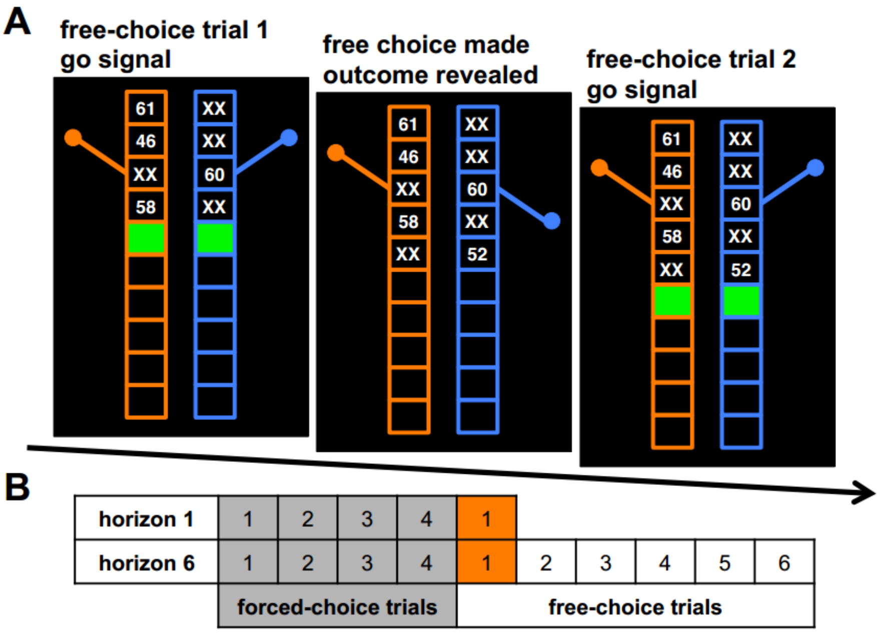
Procedure

Participants played 320 games (in 4 blocks of 80 games) of our horizon task (see Figure 5A). Each game lasted either 5 or 10 trials and the two game lengths were interleaved and counterbalanced such that there were 160 games of each length. In each game, participants made repeated decisions between two options. Each option paid out between 1 and 100 points that was sampled (rounded to the nearest integer) from a Gaussian distribution with a fixed standard deviation of 8 points. The generative means of the underlying Gaussians were different for the two options and remained stable within a game. In each game, the mean of one option was set to either 40 or 60 points and the mean of the other was set 7 relative to the mean of the first, such that the difference between the two was sampled from 4, 8, 12, 20 and 30. Both the identity and the difference in means were counterbalanced over the entire experiment.

Participants were instructed in the task via a set of illustrated onscreen instructions. These conveyed explicitly that the means of the two options were constant over a game and that the variability in the options was constant over the entire experiment. Participants were told to maximize the points they earned and that one option was always better on average. Choice and outcome history in each game remained onscreen inside each of the slot machines (Figure 5A). After a particular option was played, the reward on that trial was added to the slot machine, while the corresponding space for the unplayed option was filled with an ‘XX’.

The first four trials of each game were forced-choice trials, in which only one of the options (cued by a green square inside the next available space) was available for participants to choose. These forced-choice trials were used to manipulate the information participants had about the two options from experience (Hertwig et al. 2004) before their first free choice, while maintaining their active engagement in the task. The four forced-choice trials set up two information conditions: ‘unequal information’ (or [1 3]) in which one option was forced to be played once and the other three times, and ‘equal information’ (or [2 2]) in which each option was forced to be played twice. Crucially, this manipulation ensured that participants were exposed to a specified amount of information about each option regardless of how rewarding it was. Furthermore, the relative amount of information provided about each option was independent of the relative difference in their means.

After the forced-choice trials, participants made either 1 or 6 free choices (Figure 5B). At the beginning of each game, the number of upcoming free-choice trials (i.e., the horizon) was indicated by the length of the slot machines (Figure 5A), which contained an empty space awaiting the outcome from each of the subsequent trials.



horizon 1	1	2	3	4	1					
horizon 6	1	2	3	4	1	2	3	4	5	6
	forced-choice trials				free-choice trials					

Figure 5: Two-armed bandit horizon task. A: Example trial. B: Structure of forced-choice and free-choice trials in short horizon (horizon 1) and long horizon (horizon 6). Orange highlighted cells represent the first free choice trials, the only ones used for data analysis.

Model Fitting

A simple logistic model was used to fit the behavior on the first free-choice trial. This model computes a value Qa , for each option, a, and makes probabilistic choices based on these values. In particular Qa is the weighted sum of the expected reward Ra , information Ia , and spatial location sa,

$$Qa = Ra + \alpha Ia + Bsa$$

where α denotes the information bonus and B the spatial bias.

Assuming that the values for each option are perturbed by logistic noise with variance σ d and the model chooses the option with highest perturbed value, then the probability of choosing option a over option b is

$$pa = 1 \ 1 + exp \ Rb - Ra + \alpha(Ib - Ia) + B(sb - sa) \ \sigma \ d \ \Big/ \Big(\Big/ \ \Big) \Big/ \ \Big/$$

The expected rewards *Ra* and *Rb* were set as the observed mean of the outcomes of the example plays for options a and b respectively, which assumes that participants have a linear utility function and weigh each outcome equally in the decision.

The information *Ia* was defined such that when option b was more informative than a in the unequal condition $Ib - Ia = +1$, and $Ib - Ia = -1$ when b is less informative than a. In the equal condition $Ib - Ia = 0$. This choice of Ia allows us to interpret the information bonus as the indifference point of the choice curves in Figure 6A. Similarly, the location variable sa was defined such that $sb - sa = +1$ when b is on the right and a is on the left $sb - sa = -1$ when b is on the left, and a is on the right. By fitting this choice function to participants' data, it was possible to estimate the information bonus α , the bias B, and the magnitude of decision noise σ, separately for each participant in each information and horizon condition.

Results

The probability of choosing one of the options, option A, on the first free-choice trial, was computed as a function of the difference in means of the samples observed on the forced plays. By convention, option A was defined differently in the two information conditions. In the unequal condition it was the more informative option (i.e., the option played only once in the [1 3] forced-choice trials). In the equal condition, because both options are equally informative, option A was the option on the right hand side of the screen. The resulting empirical choice curves along with the average fits from the model are plotted in Figure 6A and B for the unequal and equal conditions respectively.

In all conditions the probability of choosing option A on the first free choice increased as a function of the difference in mean between the two options. Furthermore, for that choice, increasing the horizon from 1 to 6 increased the probability of choosing the more informative option in the unequal condition. For example, in the horizon 6 condition, even when the mean of the more informative option was 8 points lower than the alternative (-8 on the x-axis), it was still chosen 50% of the time. This change in the indifference point – the point at which participants 11 were equally likely to choose either option – is indicative of directed exploration driven by an information bonus. That is, on the first free-choice trial in the horizon 6 condition, participants behaved as though the more informative option had greater value.

In addition to the shift in the indifference point, there was also a change in the slope of the choice curves with horizon (Figure 6A and B). Curves in horizon 1 were steeper than those in horizon 6 for both information conditions. This change in slope is consistent with random exploration induced by an increase in decision noise. That is, participants' choices on the first free choice trial became more random and hence less correlated with the difference in means as the horizon increased. The model fit confirmed these informal observations, as shown in Figure 6C-E. Consistent with the choice curves, there was a highly significant increase in information bonus between horizons 1 and 6 (t(29) = 5.05, p < 10-4). Likewise a repeated measures ANOVA found a significant increase in decision noise with horizon (F(1,119) = 65.97, p < 10-8) in addition to a small main effect of information condition (F(1,119) = 5.06, p < 0.05) but no interaction between horizon and information condition (F(1,119) = 0, p = 0.96). Furthermore, the effect of horizon held for almost all subjects (Figure 6F-H) with 25 out of 30 showing an increase in information bonus in the long horizon condition, and 28 showing a similar increase in decision noise (observed in both of the information conditions).

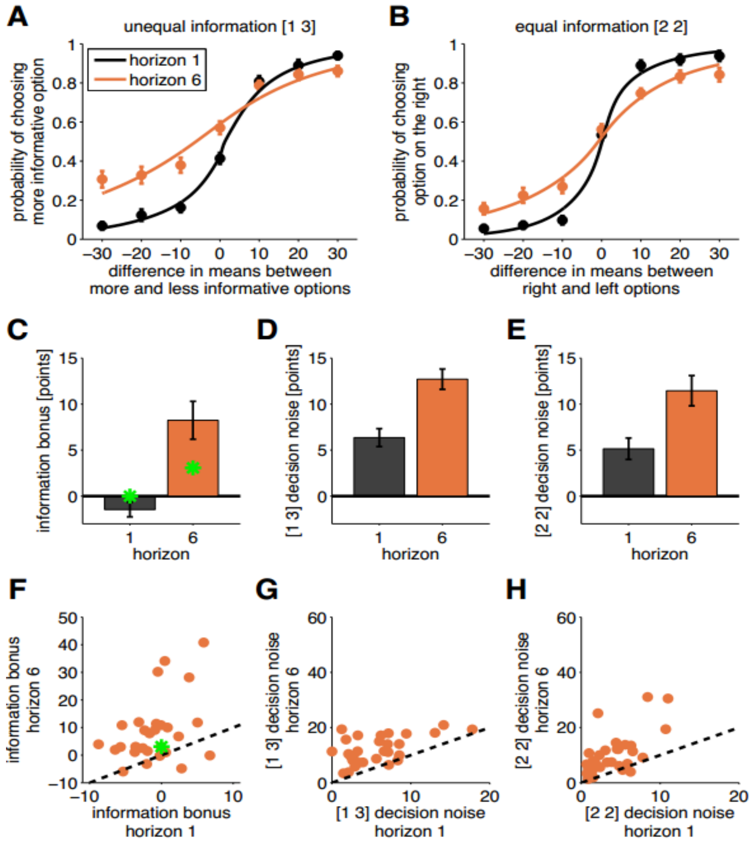


Figure 6: A,B: Choice curves for long horizon (orange line) and short horizon (black line) in the unequal information condition (A), and the equal information condition (B). C-E: Participants showed more exploration for long horizon compared to short horizon, both in the form of directed exploration (higher information bonus, panel C) and random exploration (more decision noise, panels D and E). This consistently holds true for almost all participants (panels F-G: each data point is one participant).

To test whether the change in information bonus with horizon was consistent with theories of optimal exploration, the optimal information bonus was computed for each horizon condition in the task by running the same fitting procedure on simulated choice data from the optimal model. The optimal information bonus, plotted as the green stars in

Figures 6C and 6F, bears a qualitative resemblance to the estimates of the information bonus for 12 the human participants, although quantitatively, participants appear to exhibit a greater information bonus than the optimal value. It is worth noting that, in this task, optimal performance is associated with zero decision noise, which is clearly quite different from the behavior shown by humans (Figure 6D, E).

Discussion

This work investigated the extent to which humans use directed and random exploration to solve the exploration-exploitation dilemma. The results indicate that humans use both strategies, with both an information bonus and decision noise increasing between horizons 1 and 6.

For directed exploration, results showed that, when the horizon is longer, humans exhibit an information bonus that effectively increases the value of the more informative option, making sampling that option more likely. From a theoretical perspective this result is not surprising. It is well known that information has real value for long horizons (Gittins & Jones, 1974; Gittins, 1979), and a carefully calibrated information bonus insures optimal or near-optimal exploration in many settings (Bubeck & Cesa-Bianchi, 2012). Experimentally, however, previous results on directed exploration have been mixed, with some studies finding evidence for this strategy (Meyer & Shi, 1995; Banks, Olson, & Porter, 1997; Frank et al., 2009; Steyvers, Lee, & Wagenmakers, 2009) and others failing to do so (Daw et al., 2006; Payzan, LeNestour & Bossaerts, 2011). It is possible that one reason for these mixed results is the subtle confound between reward and information that arises in sequential choice tasks and makes directed exploration both hard to observe and difficult to confirm. The horizon task described here removed this confound on the first free-choice trial by manipulating reward and information before subjects made a free choice. This made it possible to unambiguously identify directed exploration on that trial.

For random exploration, findings showed both an increase in decision noise between horizons 1 and 6 and a decrease in decision noise over the course of the horizon 6 games. Such noise-driven exploration has a long history in statistics (Thompson, 1933) and machine learning (Watkins, 1989; Bridle, 1990; Sutton & Barto, 1998) where its simplicity allows the strategy to be applied to situations in which the optimal information bonus is hard to compute. Thus, random exploration driven by decision noise may represent a reasonable adjunct to the theoretically optimal, but costly computations required to quantify the information bonus and may furthermore, even meliorate any losses when the information bonus is wrong. In this light, the use of random exploration by our participants may reflect an effort to compensate for their incorrect setting of the information bonus.

While our results demonstrate the qualitative existence of directed exploration, and the influence of decision horizon on both directed and random exploration, we did not investigate how participants might be biased toward one exploration strategy over the other, or how the two might interact. More generally, investigating this potential interaction between random and directed exploration is a question of interest. Furthermore, the design here only takes into account ambiguity as the main source of uncertainty, and does not consider the impact of risk, or outcome variability, on either strategy.

The next section presents a different study that addresses these questions, and investigates how risk influences people's exploratory decisions in the presence of ambiguity.

Experiment 3.2: Risk and ambiguity affect exploration in a sequential Wheel of Fortune task[3]

It is likely that people incorporate risk and ambiguity differently into their decision processes. Several studies argue that this distinction is important enough for the brain to maintain separate representations of risk and ambiguity (Yu & Dayan 2005, Huettel et al. 2006), that different neural substrates respond to risk (orbito-frontal cortex, striatum) and ambiguity (lateral prefrontal cortex), and that different neurotransmitters are released in the presence of risk (acetylcholine) and ambiguity (norepinephrine). Despite this important distinction, there have not been many attempts to establish how risk and ambiguity independently influence the degree to which people explore, or whether they impact random and directed exploration differently. Many studies that examine the neural and behavioural responses to risk and ambiguity pay little attention to the role of exploration, and instead search for evidence that people have separate risk- and ambiguity preferences (Hsu et al. 2005, Huettel et al. 2006). Studies investigating people's exploration strategies, on the other hand, often focus heavily on the effect of relative reward magnitude, and either disregard the influence of ambiguity, or pool risk and ambiguity together under the broader factor of uncertainty (Behrens et al. 2007, Daw

[3] This work has previously been presented in poster format to the Society for Neuroscience conference (Geana, Wilson & Cohen - Washington, DC 2014), and the Reinforcement Learning and Decision-Making conference (Geana, Wilson & Cohen - Princeton, NJ, 2013). Part of this work has also been presented during a symposium talk at the CoSyNe conference in March 2014. Please see “References” section for full references.

et al. 2006). To bridge this gap, I designed a task that permitted a clear separation and empirical manipulation of risk and ambiguity, and examined how these two factors impact exploratory behaviour.

I used the horizon manipulation from experiment 3.1 – that made it possible to distinguish between conditions in which exploration was adaptive (long horizon), versus non-adaptive (short horizon) – to test how risk, ambiguity and decision horizon interact to influence people's exploratory choices in a probabilistic decision-making task in which participants chose between two wheels of fortune. The results showed that both decision noise and information-seeking are affected not only by the length of the decision horizon (which impacts the value of acquiring information), but also by the level of risk in the environment. Our findings suggest that the mechanisms underlying exploratory behaviour are sensitive to different kinds of uncertainty, and that risk and ambiguity might interact to modulate the mixture of random and directed exploration in people's decisions.

Methods

Participants

30 participants (16 females) were recruited from the Princeton University undergraduate community. All participants gave informed consent and were compensated either with course credit, or $12 an hour plus a performance bonus of up to $5 depending on their final score in the task. The average pay was $15. The experiment was approved under the rules of the Princeton University Institutional Review Board.

Behavioural Task

Participants played multiple games of a sequential decision task that required them to choose between two virtual wheels of fortune (figure 7A). The rewards (number of points) that each wheel could pay out were written on slices of the wheel. On each trial, participants saw the two wheels and made a choice; the chosen wheel would spin a random number of times, and participants received a reward equal to the number of points written on the slice that landed under the pin. The task was split into games that consisted of either one trial or five trials, with a brief self-timed pause between games.

The mean reward of each wheel was given by the mean of the numbers on the slices. Throughout the task, participants were exposed to a range of means from 35 to 65 points. Risk in this wheel of fortune task was operationalized as the variance of the numbers on a wheel, and could take five different values that ranged from no risk (all numbers were the same) to high risk (the variance of the numbers was high). Ambiguity was operationalized as the number of question marks on the wheel, and could be either zero (all the wheel slices remained uncovered) or four (four slices were covered by the question mark). One wheel, referred to as the ambiguous wheel, always had four of the reward values covered with a question mark, and the other wheel – the certain wheel – did not. When participants chose the ambiguous wheel to spin, and it stopped on one of the covered slices, that slice would become uncovered: the computer generated a value for it based on a Gaussian distribution that maintained the variance of the rewards on the wheel. That number of points was displayed on the slice, and it remained visible for the rest of the trials in that game.

The decision horizon here was manipulated as the length of a game: each game consisted of either one choice between the two wheels (horizon 1), or five sequential choices (horizon 5) using the same pair of wheels. The number of choices left in a game was always displayed on the screen (figure 7A). The wheels remained the same within a game, but they changed between games. These changes were always signalled.

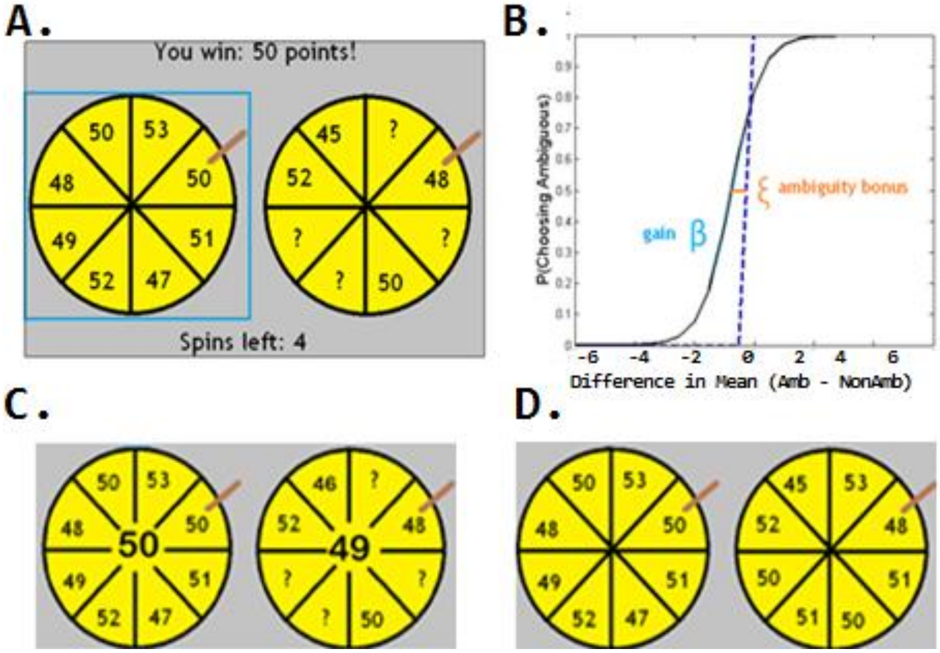


Figure 7: Wheel of Fortune task. ***A****: Sample task screen. The blue rectangle around the left wheel shows that it was chosen. The wheel stopped on the slice with the number '50', earning the participant 50 points for that spin.* ***B****. Example choice curve generated using model. The black curve is described by parameters β (slope) and ξ (centre), which in the choice model correspond to gain (inverse decision noise) and ambiguity bonus. The dotted blue line shows what the choice curve would be if the participant chose only based on reward, with no decision noise. Control tasks.* ***C****: to control for the computational demand of calculating means on wheels with different risk, I ran a version of the task in which the mean was displayed at the center of the wheel.* ***D****: to verify that the presence of ambiguity impacted how participants handled risk, I ran a version of the task in which there was no ambiguity*

Model and Analysis

I modelled participants' behaviour by assuming that they assigned a value $V(W_i)$ to each wheel W_i, based on the mean of the uncovered slices, an estimate of the mean of the covered slices, and a fit ambiguity bonus:

$$V(W_i) = \mu_{uncoverd} + \mu_{covered} + \xi \qquad (1)$$

and made their choices using a softmax function that assigns a choice probability $P(W_i)$to each option, according to:

$$P(W_i) = \frac{1}{1+e^{\beta * \Delta V_i}} \qquad (2)$$

where β is the gain parameter that determines how sensitive the choice probabilities are to the values, and ΔV_i is the difference in value between the two means. The value of an option was modelled as the expected value of the rewards (both the known, uncovered slices in the wheel of fortune task, and the unknown, covered slices). In addition to mean reward, an ambiguity bonus parameter, ξ , was added to the value computation, to quantify how much reducing the ambiguity about that particular option is worth in units of value.

The β and ξ parameters define a choice curve for each participant (figure 7B) that estimates, for any given difference in mean reward between the two wheels, how frequently the participant chose the ambiguous wheel (i.e., explore). The gain parameter in the choice model was used to quantify decision noise, a measure of random exploration, in participants' behaviour. The ambiguity bonus parameter was used as a measure of information-seeking, or directed exploration.

Results

Participants' choice curves, collapsed across risk level, show that they were generally sensitive to the mean reward values of the two wheels, with their exploration (probability of choosing the ambiguous wheel) increasing as the mean of this wheel increased relative to the unambiguous one (figure 8A, left panel). There was also an observed effect of decision horizon on exploration, with the choice curve for the long horizon (Horizon 5, orange line) being shifted to the left compared to those for the short horizon (Horizon 1, black line). This effect was not present in the control study 2, which removed ambiguity (figure 8A, left panel; for details on the control task, see figure 7D). Overall average exploration on the first trial of a game was significantly higher in the horizon 5 condition than in the horizon 1 condition ($P_{exp,horizon1} = 0.38 \pm 0.15, P_{exp,horizon5} = 0.56 \pm 0.19$). This also holds true when separating the data by risk level (figure 8B). A two-way repeated-measures ANOVA showed a main effect of horizon on overall exploration, with exploration in Horizon 5 being consistently higher, for all risk levels but one. [*F*(4,58)=2.76, *p* = 0.03]

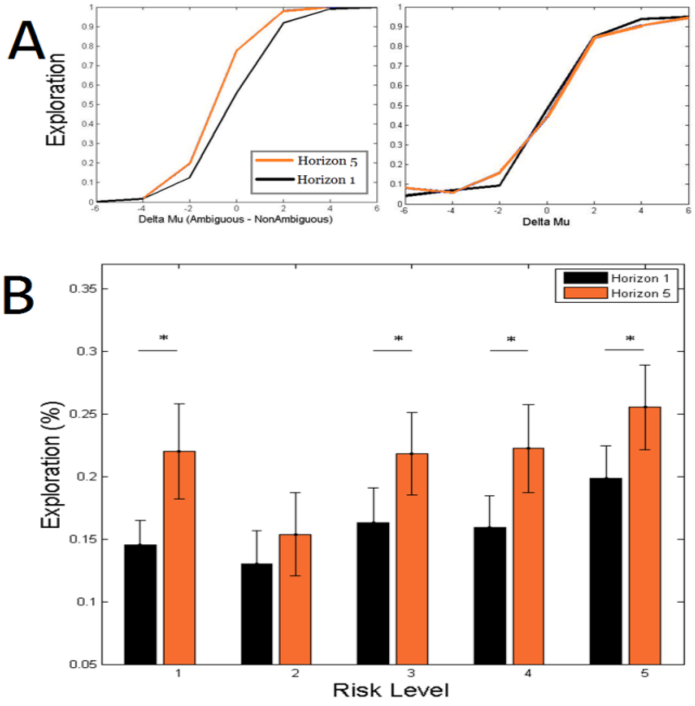


Figure 8. A: Choice curves show that probability of exploring is sensitive to the difference in mean between the ambiguous and the non-ambiguous wheel. Left panel: choice curve for horizon 5 (orange line) are shifted to the left compared to horizon 1 (black line). Right panel: choice curves overlap in control task with no ambiguous wheel. B: Overall percentage exploration is significantly higher in long horizon (orange bars) than short horizon (black bars), across risk levels. Error bars: SEM

Plotting choice curves for each individual risk level showed a significant effect of risk on exploration: the choice curves became flatter as risk increases, and they shifted to the right (figure 9A). This is equivalent to a change in both curve parameters, with decision noise (curve slope) increasing, and information bonus (curve center) decreasing. Figure 9B shows that a similar pattern was found in the choice curves in control task 1 (mean of each wheel displayed at the center of the wheel – see fig 7C), but not in control task 2 (no ambiguous wheel – see fig. 7D).

A one-way ANOVA showed a significant effect of risk level on both decision noise (figure 9C, left panel) and ambiguity bonus (figure 9C, right panel) in horizon 5 (F(4,24) = 5.62, p<0.01). There was also a significant linear trend (F(1,4) = 21.76, p = 0.01) indicating that as risk increased, decision noise increased proportionately, while ambiguity bonus decreased proportionately (F(1,4) = 64.75, p < 0.01); furthermore, the increasing decision noise and decreasing ambiguity bonus were negatively correlated ($r = -0.93, p = 0.01$). Conversely, in horizon 1, the one-way ANOVA revealed no significant effects (F(4,145)=0.10, p = 0.98), and neither the linear trend analyses nor the correlation were significant for either of the two parameters ($r = -0.39, p = 0.51$).

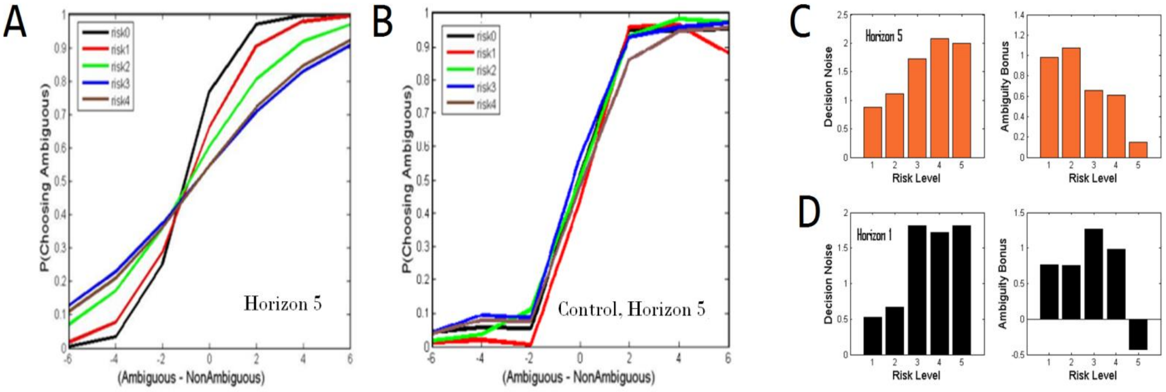


Figure 9 A. Choice curves showing the probability of choosing the ambiguous wheel as a function of the difference in mean rewards, across risk levels, for decision horizon 5. B. Choice

curves from a control task in which both wheels were non-ambiguous superpose. C. Decision noise (left panel) increases as a function of risk, in horizon 5, while ambiguity bonus (right panel) decreases proportionately. D. Decision noise and ambiguity bonus in horizon 1. The linear trend here is not significant.

Discussion

To investigate the impact of different types of uncertainty on human exploration, the experiment described above used a repeated-choice wheel of fortune task which clearly separated reward, risk and ambiguity, and manipulated participants' decision horizon between games.

First, results showed that decision horizon impacts exploration patterns: participants' choice curves were shifted to the left in the long horizon compared to short (reflecting higher information-seeking), and they explored more in the long horizon in all risk conditions. These findings are consistent with the idea that the degree to which ambiguity drives exploration is strongly correlated with the value of information (Cohen, McClure & Yu 2007, Behrens et al. 2007), and replicate earlier data (Wilson et al. 2014) regarding the effect of decision horizon on both random and directed exploration. These findings also strengthen the claim that people use exploration adaptively, increasing it when they have a longer window of opportunity in which to use the information they acquire by exploring.

Furthermore, by including a risk manipulation, this experiment allowed an examination of the separate contributions that risk and ambiguity make to exploratory decision-making. It revealed that higher risk levels were associated with a higher decision noise parameter in the choice function, suggesting that risk might influence the strength of people's preference for the more rewarding option by modulating decision noise. A potential interpretation for the increase in decision noise with higher risk levels could be related to the increased computational demands involved in calculating the mean reward of a wheel, when these means are more broadly distributed. However, a control task run with identical risk levels (but no ambiguity) showed no significant effects on decision noise. This suggests that it is not simply the computational load that causes increased decision noise in higher-risk conditions. Rather, the presence of ambiguity (and thus the potential benefit of exploration) seems to be necessary for risk to affect exploration.

Interestingly, the ambiguity bonus decreased with risk level, suggesting that people become less information-seeking when the environment becomes more variable. This effect, however, was only observed in the long-horizon condition, when information acquired by exploring could be usefully employed in future choices. This, along with the increasing decision noise in high risk conditions, indicates a potential trade-off between random and directed exploration, with more risky environments driving increased randomness and decreased information-seeking, while more ambiguous environments bias people toward information-seeking. This pattern was observed in most participants, and, while the experiment controlled for the hypothesis that it was solely due to increased computational load of higher-variance condition (control tasks 1 and 2), it is still possible that the tradeoff of random and directed exploration reflects an optimal strategy that takes into account computational demands. Specifically, it has been suggested (Wittman et al.2008) that, when the cost of explicitly computing action values becomes too high relative to the benefits earned from a simpler strategy (such as random exploration), it is more efficient to abandon the computationally intense strategy in favour of the simpler one. Thus, the increasing decision noise that parallels a decreasing ambiguity bonus when the environment becomes riskier could reflect a cost-effective strategy that ensures that exploration still happens despite increased difficulty of computing its value.

The current results are also consistent with findings in an infinite-horizon version of the bandit task, in which a similar tradeoff between random exploration and directed exploration is observed across the subject population for different environment hazard rates (Wilson & Cohen, 2014). Notably, those studies also found evidence of individual differences in the random/directed exploration balance in individual participants; furthermore, it has previously been suggested that the balance between exploration (of all types) and exploitation can also vary within a population, with different individuals biased toward different patterns of exploratory behavior (Badre et al, 2012; Frank et al, 2009). While the current work does not address the question of individual differences in adjusting exploratory strategies in response to risk and ambiguity, it is entirely possible that those differences exist – particularly as they relate to individual risk preferences and trait impulsivity scores (Niv et al., 2002; Kayser et al, 2014). When examining how humans are adaptively adjust their random-directed exploration strategies, both internal, personality-related, and external environmental demands should be taken into account.

In our wheel task, the decision to adjust between random and directed exploration strategies should also depend on how people represent variance, and how they infer the values of the hidden slices on the ambiguous wheel. Participants were instructed in detail with regard to what they should expect in terms of risk and relative reward between the two wheels. However, it is possible that their internal model did not entirely reflect the instructions they received. Participants' internal representations here were described a simple model that updated values based on the means of the two wheels, but more complex models that also update the inferred variance could shed further light on how risk impacts exploration. Previous studies have suggested ways in which risk (as variance) might impact quantities similar to decision noise (Smith et al. 2009; Heilman et al. 2010). However, to date, none have taken into account ambiguity and information seeking in the same computations – making this a critical future line of study.

The work presented here provides evidence on the role of two separate sources of uncertainty in human random and directed exploration, and suggests a possible tradeoff between a low-demand, efficient strategy (favouring random exploration) and more intensive computation of the value of exploration (favouring directed exploration), as the risk level in the environment increase. The current findings suggest that both risk and ambiguity should be taken into account when modelling exploratory behaviour, and future work is required to describe a more precise mechanism of how these two factors interact to affect exploration.

Chapter 4: Information and Task Engagement:

Humans use Exploration as an Adaptive Behavioral Response to Boredom

Introduction

The notion of boredom might seem straightforward and intuitive to anyone who has ever had to fill out taxes, wait in line at busy registers, or sit through two-hour business meetings. However, despite its prevalence as a human phenomenon, little is understood about the origin and function of boredom, or the exact mechanisms that underlie it. Building on earlier work that correlated boredom with cognitive demand (Pattyn et al. 2008), task monotony and difficulty (Hill & Perkins 1985) and effort to engage with the environment (Harris 2000), our current work suggests that the phenomenon is best examined in the context of attention, information-processing and executive control. In line with recent efforts in cognitive psychology (Eastwood et al 2012), we frame boredom as a failure to engage with the environment due to insufficient motivation, both internal (i.e., ‘this current task is not sufficiently interesting or informative’) and external (i.e. ‘this task does not bring enough benefits to make it worth staying engaged in it’).

Under this framework, some of the behavioral byproducts of boredom that are generally considered negative – such as the tendency to give up sooner or increased randomness in performance (Wallace, Vodanovich & Restino 2003, Watt & Hargis 2010), can in fact be adaptive. In previous chapters of this work I have shown that, when operating in an environment about which we do not have full and complete knowledge (as is the case with most real-life situations), exploring it by occasionally selecting random options can help us discover better strategies and often lead to more overall reward (see studies 3.1, 3.2). Furthermore, there is ample evidence in the literature (Payne, Bettman & Luce 1995; Kurzban et al 2013) that giving up on a task can in fact be optimal if the benefits we gain from it have fallen below the costs involved in doing it, including the costs associated with missed opportunities to identify and/or pursue more remunerative options.

Therefore, boredom with current circumstances may reflect a drive to explore the environment, abandoning options that have become unsatisfactory and perhaps discovering better options for gaining reward. This notion of an adaptive role for boredom has been mentioned previously in the reinforcement learning (Schmidhuber 1997) and machine learning literatures (Simsek & Barto 2005), in studies that suggested that simulated agents with the capacity for becoming bored were better explorers and more efficient learners in complex environments.

This chapter describes a series of experiments that tested these hypotheses in people (rather than artificial agents, as has been the case previously), and investigated whether and how boredom correlates with the amount of information extracted from the environment, and how it impacts people's exploration. I also propose the first normative model, to date, that considers boredom-induced exploration through an information-sampling account, and shows how quitting on a "boring" task early can in fact be adaptive for long-term reward.

Experiment 4.1: Boredom and Information: People rate tasks as more boring when they can't acquire enough information

The work presented in this section examined how boredom correlates with the amount of useful information that can be gained by continued engagement in the current task – i.e., the amount of available information that people can use for learning the structure of the environment.

Using change in prediction error as a marker for learning (the higher the prediction errors, the more there is left to learn in the environment), I found that, in line with the previous theory from reinforcement learning, people report higher boredom self-ratings if they can derive no useful information from the task, compared to lower boredom ratings when they are still learning and improving their performance. When our participants played a simple computer game in an environment in which they already knew all the information (so there was nothing left for them to learn), or an environment that was completely random (so they could not learn its structure), they reported being more bored. In contrast, when they played the task in an environment in which they could acquire useful information as they played, they were more engaged. Thus, this work tested and confirmed a previous theory- and simulation-based framework on the relationship between information/learning and boredom, showing that it does indeed hold true in human participants, as well.

Methods

Participants were asked to predict numbers generated by a virtual machine (for a similar design, see Nassar et al.2010). On each trial, they made their predictions by adjusting a vertical slider (the "prediction slider", see fig 10A) between 0 and 100 to reflect the next value that the virtual machine would generate. After they adjusted the slider, they pressed the space key to confirm their prediction, and the machine generated the number for that trial; after that, participants could make their prediction for the next trial. Games in the task consisted of thirty trials, and changes between games were signaled to the participants; there were twenty-four games in total, with the task lasting approximately one hour.

The critical experimental variable was the difference between the participants' prediction and the actual generated number, which was referred to as the Prediction Error (PE). Participants were rewarded based on these prediction errors: the smaller the error (i.e., the closer their prediction was to the actual number), the more points they received.

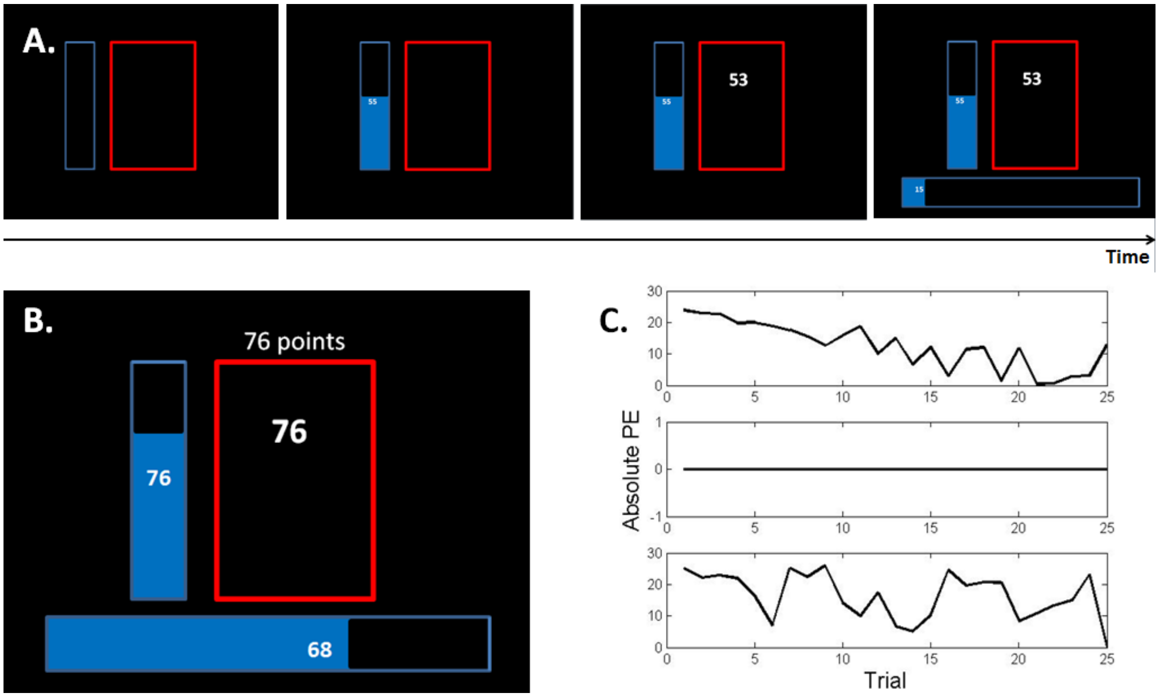


Figure 10: Number prediction task. A: the progression of a trial, from the onset of the stimuli (number generating machine – red rectangle; prediction slider: blue rectangle), to the participant adjusting the slider, to the machine generating the number for that trial (53), and the boredom slider appearing on the screen, for the participant to adjust. B. An example of the last screen in a trial in the Certain condition. C: Example pattern of prediction errors in each of the three conditions. Gaussian(top panel) – PE is reduced as participant advances in the game. Certain (middle panel) – PE is zero throughout the game. Random (bottom panel): PE remains high regardless of position in the game.

The machine generated numbers according to an underlying distribution, which differed between conditions. In the “Gaussian” condition, numbers were generated from a Gaussian distribution with a fixed mean and standard deviation; however, each number was not displayed until after the participant recorded their guess. In the “Certain” condition, numbers were generated from a Gaussian, and displayed on the screen before the participant made their response (see figure 10B). In the “Random” condition, numbers were generated uniformly between 0 and 100, but again not displayed until after the participant recorded their guess. Therefore, the underlying distribution of the number-generating machine was such that participants had to either learn the generative process to gradually reduce their prediction error (in the Gaussian condition), or they were already told the next number and did not need to learn anything to make perfect predictions (the Certain condition), or the numbers were randomly generated and participants could not reduce their prediction error (in the Random condition). Figure 10C shows these three PE patterns.

The purpose of this study was to test the relationship between informativeness in a task (operationalized as the ability to change prediction errors by continuing to do the task) and boredom. Therefore, in addition to predicting the upcoming number on each trial, participants were also asked to self-report their level of boredom several times throughout a game. Every fourth trial (for a total of ten times throughout the game) they were asked to self-report boredom by adjusting another slider (the "boredom slider") at the bottom of the screen (figure 10A).

Results

As predicted, the "Certain" condition elicited the highest boredom average overall ratings in all participants ($M_{Certain} = 80.39, M_{Random} = 63.14, M_{Gaussian} = 41.27$, repeated measures ANOVA, $F(2, 60) = 5.03$, $p = 0.01$). The ratings for the Certain condition were also consistently higher than for the Gaussian and Random conditions in both early trials (first six games) versus late trials (last six games), as shown in figure 11B, and for the average ratings within a game. The Gaussian condition, by contrast, was consistently rated as the least boring; the Random condition was rated in-between the other two, though this effect did not reach significance.

There was also a significant observed main effect of time on boredom ratings: for all three conditions, later ratings were significantly higher than earlier ratings, both within a game , and across the entire session in early versus late trials (figures 11A and 11B, two-way repeated measures ANOVA, $F(2,18) = 13.39$, $p < 0.01$).

Absolute prediction errors were computed for each game (as the absolute value of the difference in participants' prediction from the number generated on each trial) and the average change in prediction error in a game was computed the average difference between PEs on consecutive trials. These values were then binned for changes in PE, and the average boredom ratings corresponding to those games were calculated (regardless of which condition those games were in – although, as explained in the methods, the participants were only able to significantly reduce their PE in the Gaussian condition). As shown in figure 11C, there was a significant negative correlation between the change in prediction error and the boredom ratings ($R^2 = 0.58, p = 0.004$)

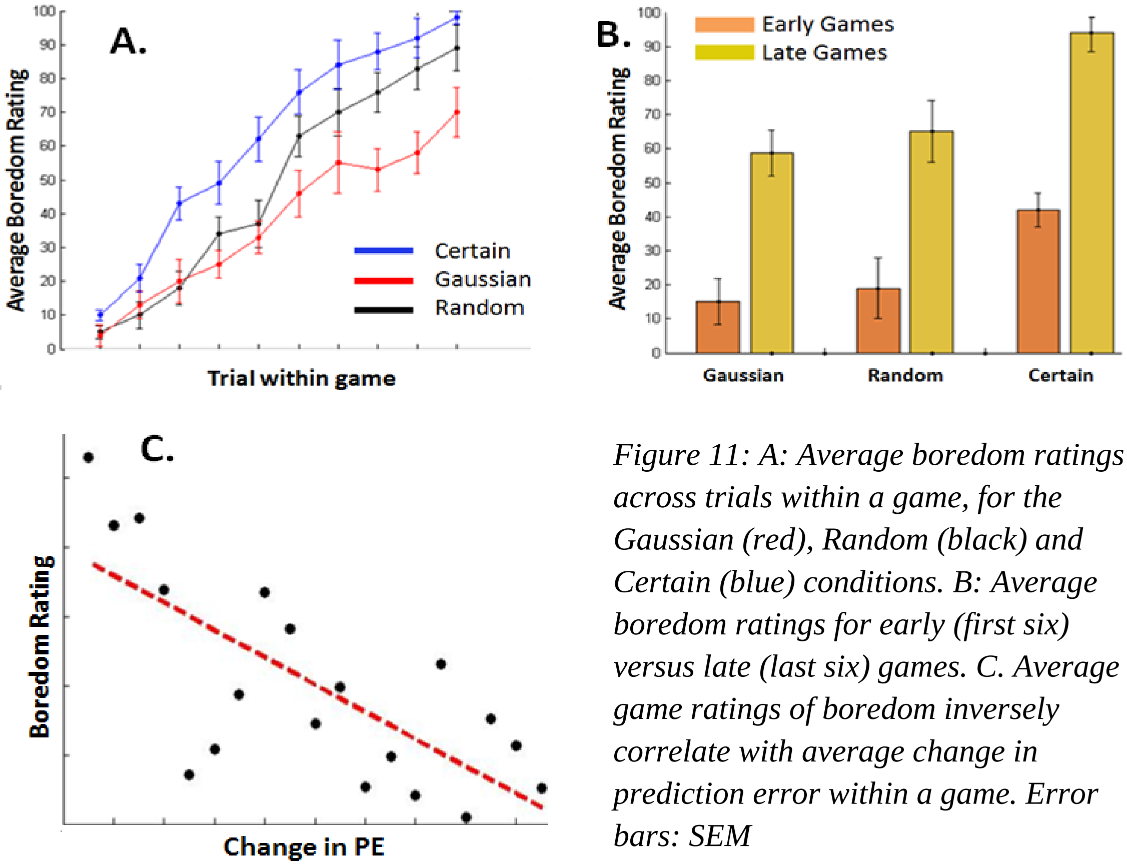


Figure 11: A: Average boredom ratings across trials within a game, for the Gaussian (red), Random (black) and Certain (blue) conditions. B: Average boredom ratings for early (first six) versus late (last six) games. C. Average game ratings of boredom inversely correlate with average change in prediction error within a game. Error bars: SEM

Discussion

The current experiment showed evidence that the ability to change prediction error in a learning task correlates with boredom. This correlation was first suggested in theoretical reinforcement learning work by Schmidhuber (1997), and later tested in artificial multidimensional environments with boredom-capable learning agents (Barto & Simsek 2005; Simsek & Barto 2006). To our knowledge, however, this is the first demonstration of such a correlation in human data, with a direct operationalization of boredom as people’s self-ratings.

Our results suggest that the amount of information that can be learned from a task is linked to how boring the task is perceived: the Certain condition – i.e., the one in which there was no useful information to be learned by performing the task, because all the information was already given to participants – elicited the highest boredom ratings (figure 11), while the Gaussian condition, in which it was possible to improve predictions by learning the underlying number-generating distribution, was rated as the least boring. This is consistent with previous theories on 'too much or too little information' causing suboptimal levels of arousal (Csikszentmihalyi 1990; de Rijk, Schreurs & Bensing 1999), as well as with the notion of "desirable difficulty" – i.e., the notion that there a certain amount of effortful information-processing actually helps learning (Bjork, 1994; Richland et al. 2005). It was shown here that giving participants all available information and removing the need for them to learn anything lead to high boredom. that we

The ratings in the Random condition started out closer to the Gaussian ratings (as can be seen in the average within-game ratings, fig 11A), but ended close to the Certain ratings. This is likely due to the fact that it took several trials for participants to acquire enough evidence to learn that the numbers were being generated randomly. Accordingly, the shift in boredom ratings within Random games from lower to higher about half-way through is consistent with the theory that too little information is as likely to elicit boredom as too much information , since the participants end the Random games reporting similar levels or boredom as in the Certain games. However, they reach those levels of boredom only after acquiring sufficient experience to confirm the distribution of numbers is random and that new trials hold no useful information that could help reduce their prediction errors.

It should be noted here that while the current task design makes similar predictions for participants' boredom in the Random and Certain conditions, the former might require more "effort" – in terms of the cognitive demands of forming a representation of the underlying number distribution. It is likely that if that type of effort alone were aversive enough to elicit the high boredom ratings in the Random condition (rather than the lack of informativeness), the ratings in the Gaussian condition might also be higher – since participants also need to engage similar cognitive processes to learn the Gaussian distribution. However, it is difficult to quantify how the subjective effort of representing a random distribution compares to the effort of representing a Gaussian distribution. Furthermore, studies have suggested that people might in fact seek out more effortful tasks if it means avoiding idleness (Hsee, Yang & Wang, 2010, Navarro & Osiurak 2015). In addition to that, the Certain condition entails a different type of effort, related to increased control demands of staying on task (Kurzban et al. 2013). This design does not permit an in-depth analysis of the different types of "effort" involved in this kind of learning task, and further work is needed to investigate exactly how effort interacts with the informativeness to elicit boredom.

Interestingly, not only did participants rate the Certain condition as the most boring, but they also tended to find it aversive: post-task briefings included frequent comments on the dislike they felt as soon as they started a new game and realized it was a Certain game. This is consistent with Schmidhuber's model that assumed a penalty for the inability to acquire new useful information, as well as with previous qualitative boredom models (Hill & Perkins 1985) that more frequently predict high aversion to contexts in which too much is known, rather than too little. It should be noted here that the latter half of the Gaussian games was very similar to the Certain games in nature: by then, participants had obtained enough data to make a strong guess about the mean of the underlying number distribution, and they no longer adjusted their predictions by much. Thus, their change in PE resembled the Certain condition – yet their average ratings late in the game were still lower than the late-game ratings for the Certain condition. The most likely explanation for this is an anchoring effect, as the task design of incrementally adjusting the slider to reflect their boredom ratings probably prevented participants from making disproportionately large adjustments in the later trials of the Gaussian games.

Finally, in addition to a significant effect of informativeness on boredom ratings, there was also a notable effect of time: within a game, average ratings increased from the first trials to the last trials, and across the entire session, later games were rated as significantly more boring than early games. Therefore, the longer participants spent playing the task, the more boring it became. This type of effect has been observed in many other experimental contexts (London & Monello 1974; Damrad-Frye & Laird 1989), and most probably reflects novelty – that is, informativeness at the level of the task itself – as well as informativeness on a trial-to-trial basis, both of which diminish with the passage of time. It is, however, interesting to note that the effect of time seemed similar in all there conditions (the two-way repeated-measures ANOVA in figure 11A revealed no interaction). This suggests that this component of boredom-related effects may reflect primarily task-level effects, or other time-on-task effects (such as fatigue, or perhaps an increased perceived opportunity cost of staying with the task, as suggested previously by Kurzban et al. 2013; Bench & Lench 2013), above and beyond effects of information related to the individual trials of the task.

In summary, this study provided empirical evidence that participants' self-reported boredom ratings are closely related to the possibility for information acquisition in their current task, an effect that is consistent with previous theoretical work suggesting that too much or too little information can be associated with an increased perception of boredom. However, the current task design did not include an explicit exploration component – while people could show some degree of "exploration" by, for instance, making inaccurate predictions in the Certain condition (the increased error rate in that condition is consistent with existing theories of boredom and vigilance/fatigue – see Patyn et al., 2008), they could not truly switch away from their boring contexts by exploration. Study 4.2 used a slightly modified design of the number prediction task to allow such exploration, and investigate whether low-information conditions that showed an effect on boredom perceptions here also influence participants' exploratory behavior.

Experiment 4.2: Boredom and Exploration: People are willing to take an income penalty to switch away from boring contexts

This experiment further examined the extent of this increase in exploratory behavior in response to boredom. Using the same three information conditions that elicited differential boredom levels in participants in Study 4.1, I modified the number-prediction in order to allow participants to determine, on their own, whether they wished to persist in the task or change the task conditions. This afforded a more direct examination of the relationship between informativeness and exploration, which could then be related to boredom by comparing the findings with those of Experiment 4.1. Results showed that the information-poor conditions, even when they were rich in reward, did not successfully engage participants. This suggests that task informativeness might play a role in the value computations that drive exploration.

Methods

Participants played a variant of the number-prediction task presented in Study 4.1. Just as in the previous version, each trial required them to adjust a vertical slider between 0 and 100 to predict the next number that would come up on the screen. The closer their prediction was to the actual number generated by the virtual machine, the more points they received. Also similarly to Study 4.1, the number generating process within a game stayed constant. However, unlike in our previous study, the games here did not have a fixed length. Rather, participants were told that a game could go on for up to one hundred trials, but they could choose to end it earlier and move on to a new game at any time by pressing the "reset" button to the right of the screen (figure 12). If they pressed the "reset" button, they would see a brief inter-game screen, and then start a new game with a new number-generating process. Participants were informed that the task would take approximately fifty minutes, regardless of how many games they went through in that period: the task finished at the end of the first game after the fifty-minute time period was up.

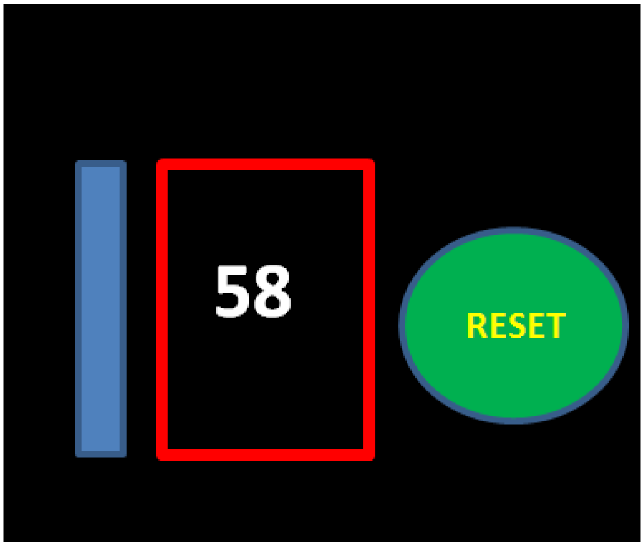


Figure 12: Experimental design, a variant of the number prediction task in Study 4.1. Participants had to predict the number that would appear inside the red rectangle on each trial. They used the mouse to drag the vertical blue slider on the left and make their predictions. At any point during a game, they could press the green RESET button on the right, to end the current game and move on to a new one.

There was no boredom slider as in Experiment 4.1, since the aim of this experiment was to measure exploration in the same three conditions in which our previous experiment had already found different boredom levels. In all other respects, the task was the same as in Experiment 4.1: the "Gaussian" condition, the numbers were drawn from a Gaussian distribution with a fixed mean and standard deviation; in the "Random" condition, they were drawn from a uniform distribution; and in the "Certain" condition, they were again drawn from a Gaussian, but the upcoming number was displayed on the screen before each trial (see fig 10B). After a game ended (because the participant pressed the "Reset" button, or after 100 trials), the next game would be drawn from any of the three conditions, with equal probability.

Results

No participants chose to stay in any game for the entire duration of one hundred trials; all pressed the "reset" button to move on to a new game well before the total number of possible trials in the current game had elapsed. There was a significant difference, however, in the average number of trials spent in a game before choosing to switch: most participants spent significantly longer on games in the Gaussian condition than in either of the other two conditions (repeated measures ANOVA, *F*(2,69) = 7.04, *p* < 0.01; fig 13A). There was no significant difference on games in the Certain condition versus in the Random condition (paired t-test, *t*(23) = -1.29, *p* = 0.206). However, as shown in fig 13C, the probability of switching away after the first trial of a new game was significantly higher for the Certain games than either of the other two conditions (*F*(2,69)=9.22, *p*<0.01); as expected, the values for Random and Certain were statistically indistinguishable (*t*(23) = 0.11, *p* = 0.91).

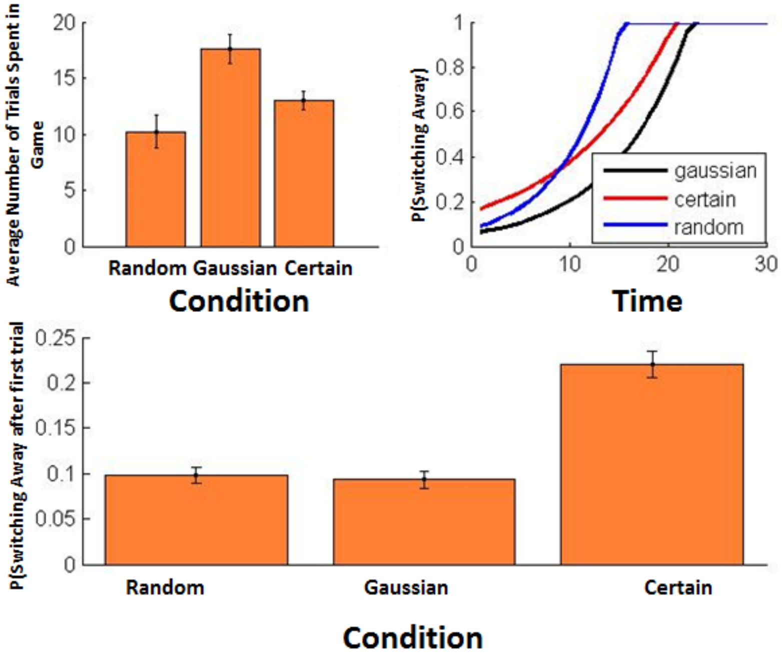


Figure 13: A. Average trial spent in game, in each of the three information conditions. B. Fit logistic probability of switching away from a game as a function of the trial in game. C. Probability of switching away after the first trial.

Finally, participants were more likely to want to switch away from all types of games later in the experiment than earlier: as shown in figure 13B, the average in-game stay was shorter for the last ten games than for the first ten games.

Discussion

The results of this experiment suggest that boredom carries a “penalty” – participants playing a number-prediction task similar to Study 4.1 were willing to take a point loss in order to quit boring games early, even though that strategy was not optimal. This is because performance was necessarily (and obviously) highest in the Certain condition, in which participants were told the answer on every trial. Thus, the optimal strategy would have been to stay in the high-reward, Certain games until the end, and always switch away from the Gaussian and Random games until finding another Certain game. However, none of the participants showed that pattern of behavior, choosing instead to switch away from Certain games *more* frequently than from the other two conditions.

Note that it was also optimal to quit Random games early: the information-sampling opportunities were almost as sparse as in the Certain condition, and since reward was proportional to the accuracy of the predictions, this was at the same time a low-reward and a high-boredom condition. As discussed in Study 4.1, participants’ willingness to stay for any amount of time in the Random games was likely due to the fact that they needed a few trials (i.e. data points) to confirm that they were indeed playing a Random game. Nevertheless, they switched away from Random games significantly sooner than from Gaussian games, around the same number of trials as from Certain games. The fact that participants showed similar patterns for the low-reward, high-boredom Random condition as for the high-reward, high-boredom Certain condition strongly suggests that they were taking into account more than just the extrinsic reward (i.e. the number of points which ultimately translated to payment) when making their decisions.

It is possible that participants' would have been more likely behave in a reward-optimal manner if, instead of being told that they had to perform the task for a fixed time duration, they had been given a fixed number of trials, or a fixed number of games, on which to get as much reward as possible. This is suggested by observations that compelling people to do monotonous tasks for fixed periods of time without the option to quit leads to more errors (Patyn et al 2008). It is also not clear, from this experiment alone, just how much they would have willing to give up to avoid boredom. It is possible that the amount of reward they sacrificed in this task by switching away from the Certain games was not sufficiently high (as it amounted to approximately $1 - $2) to motivate them to submit to the boredom associated with the Certain trials for longer. However, pilot data suggests that even higher amounts of reward cannot fully compensate for the effects for boredom – we offered participants as much as $40 – about three times the normal amount – to do a boring, monotonous vigilance task that offered no useful information, and they still quit much earlier than optimal from a purely reward-based point of view.

There are many observed instances of seemingly suboptimal behavior, in which humans and animals forego local maximization but are actually using a strategy that maximizes long-term reward (Krebs, Kacelnik & Taylor 1978; Stevyers Lee & Wagenmakers 2009). While under the current task design, quitting early in the Certain condition was not in fact optimal either locally or globally, the conditions used here were the same conditions that elicited a U-shaped curve of boredom ratings in Experiment 4.1, and quitting times mirrored that curve. It is therefore possible that a sustained experience of boredom in the task overall caused participants to penalize even the high-reward conditions, leading them to suboptimal exploratory choices.

In summary, Experiments 4.1 and 4.2 used a number prediction task to manipulate the degree of informativeness in different task conditions, and observe its effect on humans' self-reported boredom (4.1) and exploratory behavior (4.2). Participants reported greater boredom in contexts that were less informative (4.1) and were willing to sacrifice monetary value to avoid those contexts. Other studies have suggested that boredom emerges not only as a consequence of the properties of the current task (such as informativeness, as tested here), but also as a consequence of the overall environment in which the individual performs the task. This interplay between local task structure and global environment structure is particularly relevant when examining how boredom might signal a decision-maker to switch from a current task to exploring the environment. The next section discusses an experiment examining the idea that boredom serves as a form of opportunity cost signal that weighs the current task against other available options.

Experiment 4.3. Boredom and Opportunity Cost: People rate tasks as more boring, and show more exploration, when there are more interesting tasks available

The notion that motivation and task engagement patterns can emerge from global properties of the environment (rather than the local task currently being performed) precedes the work in human cognitive boredom. As early as 1954, Fowler discusses "the facilitating effect of irrelevant sources of drive on exploratory behavior" in rats, suggesting that the very awareness of diverse stimuli in the environment (even if those stimuli are not actually available for engagement) can prompt more switching between the available options and earlier leaving times when new options are made accessible. Motivation theories have also suggested the notion of an “optimal arousal” level, which organisms strive to achieve by balancing their engagement with various sources of arousal in the environment (Csikzentymihalyi 1990; Aston-Jones & Cohen 2005).

More recently, Kurzban et al (2013), building on previous literature on motivation, proposed that the subjective experience of boredom results from a computation of opportunity cost. Under their theoretical framework, boredom could be a signal that maintaining attention on the current task is becoming increasingly effortful and unrewarding compared to other options in the environment. This idea of a cost/benefit analysis underlying motivation and boredom has also been suggested in other work (Larrick, Nisbett & Morgan 1993; Eastwood et al. 2012). However, given that the interest in quantitative, cognitive accounts of boredom has only recently gained momentum, this ideas has not yet been systematically examined in humans.

The current experiment tested whether it was possible to change people’s perceptions of the “interestingness” of a task by manipulating the availability of other more or less attractive options in the environment. Results suggested that not only do people rate the same task differently based on their other available options, but while they are performing the task, their exploration levels are significantly higher when they are more bored.

Methods

Participants

40 participants recruited from the Princeton University undergraduate community gave informed consent and received course credit for participating in this experiment. Study materials and procedures were approved by the Princeton University Institutional Review Board

Procedure

Participants played a task that consisted of two parts (referred to as part A and part B, figure 14), played in order. They were told at the beginning of the experiment about both parts, and what each part would entail. They were also regularly reminded about part B while playing part A (they received three reminders, every five minutes, for the twenty-minute duration of part A).

All participants played the same task for part A: the two-armed bandit task described in Study 3.1 (figure 5A, chapter 3). They played seventy-two games of either horizon 1 (a total of five trials: four forced-choice trials, one free trial) or horizon 6 (a total of ten trials, four force-choice, six free trials), and were instructed to choose between the two options so as to maximize their score. Every few games, participants received a query screen that asked them to assess task-related factors such as difficulty, estimating the average number of points they received and other measures (see Appendix for the full list of questions). Among these questions, there were regular queries about their level of interest in the game, which they were asked to rate a total of six times while playing the bandit task (part A).

Part B differed between participants, and for each involved one of a set of tasks that had been previously rated as more or less interesting by a different sample of participants in a brief task-rating study. As noted above, participants were instructed from the start that they would be performing this second part after part A of the experiment was finished, and were reminded about it three times during part A. Each participant was assigned to one of four conditions, each of which involved a different task to perform during part B. 10 participants watched "CrashCourse" YouTube video (previously rated as a highly interesting task); 10 participants counted the number of words in a two-page mathematical typography article (a task previously rated as highly boring); 11 participants played a simple color-matching game (previously rated at medium levels of interestingness), and 9 participants played another round of a bandit task very similar to the one played in part A. Figure 14 shows the four possible conditions.

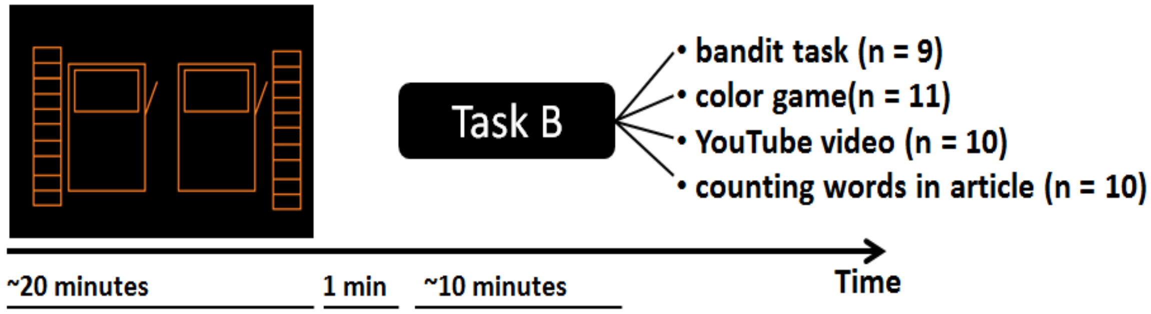


Figure 14: Experimental paradigm for the opportunity cost study. Participants performed the bandit task (task A) for approximately 20 minutes. They had a break of ~1 minute, then performed a second task (B) which could be one of four options.

At the end of both parts A and B, participants received a brief questionnaire that included the Boredom Proneness Scale (BPS, Farmer & Sundberg 1986), for a measure of their trait boredom proneness. For the full list of questions in the scale, see Appendix.

Results

The ratings for the bandit task in part A, which all participants played for the first twenty minutes of the experiment, were significantly different between conditions. Depending on whether the task for part B was one rated as highly interesting (the YouTube video), highly boring (the word-counting task), or something in-between, participants rated the interestingness of the bandit task differently. A one-way ANOVA showed a significant main effect of condition (not including the control condition, F(2,26) = 9.07 , p < 0.01), with participants who performed the bandit task before watching the video rating it as more boring than those participants who performed the task before counting the words in the mathematical typography article (paired t-test, *t*(1,15) = 5.21, *p* < 0.01). The ratings for the simple color game condition, as well the control, fell in the middle range (fig 15C).

Exploration in the bandit task was defined as in Study 3.1 – choosing the ambiguous bandit. Participants here replicated the results from studies 3.1 and 3.2 regarding the impact of decision horizon on overall exploration: figure 4a shows that the decision curve for the long horizon (red) is flatter and shifted to the right compared to the curve for the short horizon (black), which suggests higher exploration for the long horizon, both in the form of decision noise and of information bonus.

Looking at the average exploration within the task revealed a significant correlation between the participants' ratings of the task and amount of exploration: the higher the boredom rating, the more likely participants were to explore (figure 15D). This pattern was only observed in the long decision horizon, and not in the short horizon (*F* (1,33) = 6.01, *p* = 0.02 for Horizon 6, *F*(1,33) = 1.88, *p* = 0.17 for Horizon 1), and it was observed for both fit decision noise and the fit information bonus (which were both significantly higher in the long horizon with high boredom ratings, fig 15C).

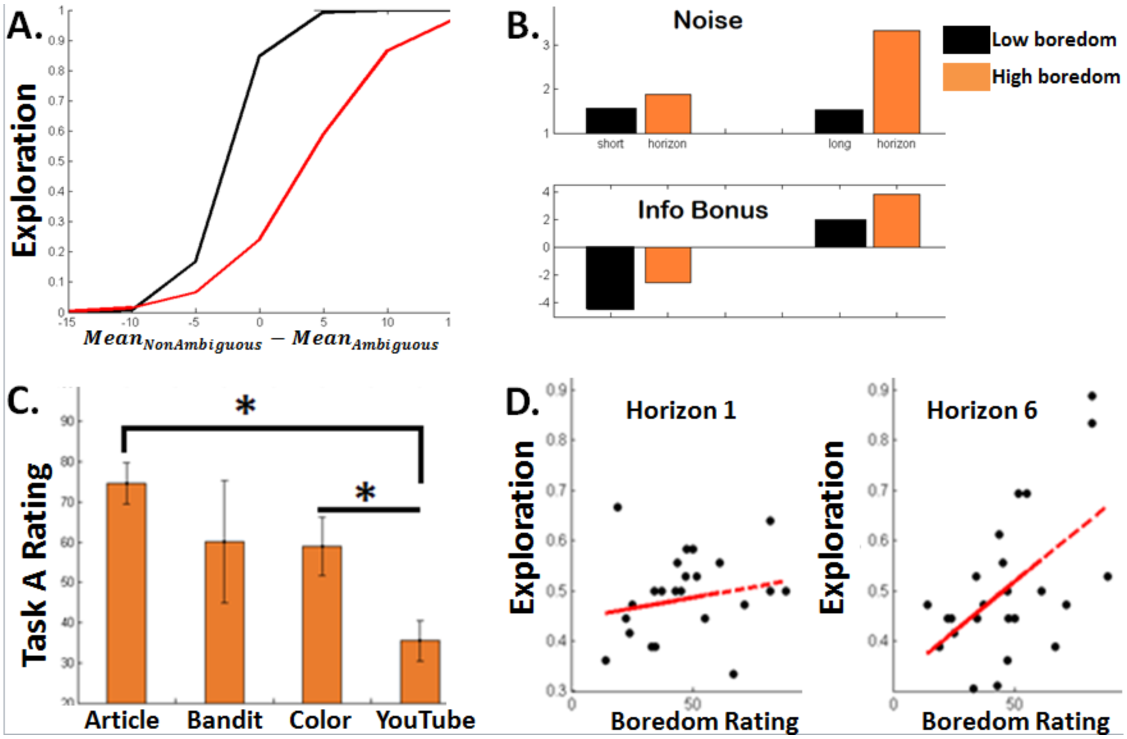


Figure 15: More exploration in long horizon (A). Higher boredom (per participant) correlated with higher exploration (B, D). Interestingness ratings of bandit task differ depending on second task (C). Error bars: SEM.

Discussion

The results of Experiment 4.3 indicate that it was possible to manipulate participants' perceptions of the interestingness of the bandit task based solely on the context in which they were given the task. All participants played the same task for the same period of time, but those participants who had been told that after finishing the bandit task they would watch a YouTube video rated it as significantly more boring than those who had been told that they would have to count the words in an article. This is consistent with previous findings regarding the effect of increased available stimulation on relative motivation (Fowler 1954), as well as the theoretical framework proposed recently by Kurzban et al (2013) that links the perception of boredom with opportunity cost.

A second important result was the strong correlation between participants' boredom and their exploratory behavior in the two-armed bandit horizon task: participants who rated the bandit task as more boring also showed significantly higher exploration of the ambiguous bandit (fig 15D). Increased exploration in response to a boring situation has previously been suggested in the literature (Fowler 1954; Litman & Spielberger 2003; Cohen, McClure & Yu 2007), but this is the first report of a correlation between a quantitative operationalization of boredom (as self-reported ratings) and a measure of exploration.

Interestingly, the association between exploration and boredom was only observed in the long horizon condition (horizon 6). As discussed in chapter 3, the value of information acquired from directed exploration can only be optimally used in the long horizon condition (Wilson et al. 2014). Therefore, the fact that the correlation between boredom and exploration was only observed in horizon 6 suggests that this condition engaged a potentially adaptive, information-sampling-based mechanism that maybe have driven the increase in exploration as a consequence of higher perceived boredom.

Experiment 4.1 provided evidence that the perception of boredom is related to informativeness in the task, operationalized as the utility of information acquired on each new trial. Experiment 4.2 showed that the low-information contexts (such as the Certain and Random conditions), in addition to eliciting higher perceived boredom in participants, also elicited higher exploration. Experiment 4.3 suggested that the wider context of the situation in which the task is performed may also play a role (i.e., what else there is to do beyond the current task), and again showed that humans explore more in high-boredom situations. In the section that follows, I described a model that formalizes boredom in terms of an information-sampling process, and accounts for switching away from boring conditions (exploration) as a potentially adaptive response.

4.4. An Information-Sampling Model of Boredom and Exploration

The work described so far in this chapter established three important findings relating to task disengagement and subsequent exploration in the context of information sampling. First, I showed that humans' levels of reported boredom correlate in a non-monotonic fashion with the amount of useful information they can extract from the environment (Experiment 4.1), such that when there is too much, or too little available information, they become more bored. Secondly, I showed that human's perception of how boring a particular task is can be modulated based on what else is available in the environment while they are performing that task (Experiment 4.3). This strongly suggests that when people compute the value of staying with the current action, they take into account some measure of relative value between the local and global environments. Furthermore, I also showed that when boredom levels are high (due to low information content), people show more exploratory behavior and switching away from the current task – both in situations when that behavior is useful (Experiment 4.3.), and even when it is suboptimal (Experiment 4.2).

Some of these findings have previously been suggested in theoretical work in reinforcement learning, machine learning, and psychology (Schmidthuber 1997; Barto & Simsek 2004; Simsek & Barto 2005; Kurzban et al. 2013). However, to date, they have not been incorporated into a single integrated model that provides a unified, quantitative account for all the observed phenomena. Previous cognitive models of boredom are either entirely theoretical (Eastwood et al. 2012; Kurzban et al.2013), or they rely on qualitative predictions and unsystematic operationalization of boredom (Hill & Perkins 1985; Perkins & Hill 1986). This section, building on our experimental data and theoretical support from the machine learning and optimal foraging literatures, proposes a simple model that normatively accounts for how people's observed task disengagement patterns, and subsequent increased exploration, can emerge as adaptive responses to certain types of information structures in the environment.

Model Assumptions

As some of the basics of this model were rooted in optimal foraging work, the model was envisioned to operate in an environment consisting of local reward patches that offered different reward rates, with the model agent free to either stay within a patch to reap reward (exploit), or switch away to search for other patches (explore). This structure resembled foraging environments (Charnov 1976; Pyke 1984; Kacelnik & Brunner 2002), and therefore also lent itself easily to the use of dynamic programming and Gittins indices to calculate optimal policies for when the agent should switch away from a current patch (Gittins 1979).

The structure of each patch mirrored the experimental structure in studies 4.1 and 4.3, i.e. a stochastic environment in which the agent could earn rewards by making accurate predictions. In this framework (just as in the two number-prediction task studies), each patch had a hidden distribution with mean μ_i and standard deviation σ_i. On each time step spent inside the patch, the agent had to make a prediction relating to this distribution. The agent's reward r_i was proportional to the accuracy of the prediction (for a similar task design, see Nassar et al.'s (2010) "estimation task"), according to

$$r_{t,i} = \rho - PE_t, \tag{1}$$

where ρ represented the maximum amount of reward that an agent could earn (if its predictions were fully accurate), and PE_t represented the prediction error, computed as the difference between the agent's prediction *Pr* and the actual number generated in the patch on time step *t*:

$$PE_t = Pr_t - N(\mu_i, \sigma_i) \tag{2}$$

Thus, the longer an agent spent in a patch, the better it could estimate the underlying patch distribution (as it experienced more data points from that distribution), and the more accurate its predictions could become. This marks an important difference to most foraging environments, as under our assumptions the patch actually became more rewarding with the passage of time, rather than depleting. In this way, the environment resembles a learning task more than a foraging one, and is closer to the reinforcement-learning framework in which repeated trials in the same environment lead to increased performance (Sutton & Barto 1998).

An agent could spend as long as it wanted exploiting a patch, but each patch had a fixed chance of termination λ, meaning that on every time step the patch would end with probability λ, and continue with probability (1 – λ).

One essential assumption of our model is that all local patches were connected under a higher-level, global structure. In other words, the underlying patch distribution parameters μ_i and σ_i came from a global distribution with a (fixed) grand mean M and standard deviation S. This is a property of many real-life environments, in which humans sequentially sampling different "patches" learn about the local structure while simultaneously learning about an overarching global structure (Diuk et al 2013). For instance, when going apple-picking, we learn about the quality and availability of fruit in each individual tree we choose to pick from (so we could choose to move from a smaller, poorer tree to a better one), but at the same time we are also learning about the overall qualities of the orchard, so next time we go apple-picking me might choose an altogether different orchard.

A more practical example, which reflects our model assumptions even more faithfully, is job training procedure after starting a new job. Many companies have started requiring their employees to rotate teams and responsibilities for a long training period after being hired; under this procedure, an employee has the chance to learn information about each different position he or she tries out, and the longer they spend in a position, the better (presumably) they become at it, and their returns increase while their need for learning about that position decreases. Sometimes, an employee might choose to switch out of a position early, if they feel that they are stagnating and no longer challenged, and move to a new project with new challenges. At the same time, while they are experiencing each individual position, they are also simultaneously learning global information about the company culture, its goals, its management etc., and thus become better about knowing what to expect in their next position within the company.

This was exactly the type of framework our present model assumed. Exploiting a local patch obtained increasing local reward (fig. 16A), but exploring many local patches helps the agent learn the global structure faster. There was also a global reward *R* associated with learning the global mean *M*. Depending on goals, therefore, it could be optimal to quit a local patch before its ending time (even though it was yielding a high reward) and move on to a lower-reward patch that contained better information about the global mean. This strategy resembles the idea of "early stopping" or "optimal stopping" in neural network training (Sarle, 1995; Prechelt, 1998): when there is a danger of overfitting, which would impair generalization, it is best to keep track of how much each new data point contributes to generalization error, and stop training (even if that means not making full use of the available training set) when the new data points start hurting generalization instead of helping.

Our model tracked several quantities of interest as an agent exploited a patch with the above structure. First, at each time step it computed an estimate of the local mean for patch *i* at time *t*, $\mu_{i,t}$, as the average of all data points $x_{i,t}$ observed in that patch up to the current time:

$$\overline{\mu_{i,t}} = \frac{\sum_t x_{i,t}}{n} \qquad (3)$$

which can also be written in terms of the prediction error *PE* and a learning rate of 1/n, as

$$\overline{\mu_{i,t}} = \overline{\mu_{i,t-1}} + \frac{1}{n} * PE_t \tag{4}$$

(Given the structure of the task, the optimal prediction at any time step was the current estimate of the mean, $\overline{\mu_{i,t}}$, and our model assumed that the agent would always predict that mean)

In addition to tracking the mean estimate for the patch, the model also tracked a variance estimate of the local patch,

$$\sigma_{i,t}^2 = \sigma_{i,t-1}^2 + \frac{1}{n}\left(\frac{(n-1)*PE^2}{n} - \sigma_{i,t-1}^2\right) \tag{5}$$

which allowed computation of how informative each new data point was, in terms of how much it could reduce variance about the local patch. As the above equation shows, the informativeness of each new data point decreased proportionally to *1/n* (see fig 16b).

The model also tracked an estimate of the global mean and variance M and S, in terms of the history of visited patches. Each final mean estimate, μ_i, for the distribution within a patch served as an additional data point for inferring the grand mean M , in the same way that each within-patch data point served to estimate μ_i (and the variance of the global distribution was similarly computed using the μ_i values).

Crucially, the model assumed that upon first entering a new patch, the initial prediction regarding the distribution of that patch (essentially, the prior, before any data points from that patch were observed) was set to the current estimate of the global mean *M*. This provided a way to quantify the value of information in each patch, in terms of expected reward, as the estimated improvement in initial predictions on future patches. That is, the better the estimate of the global mean, the better the agent could do, on average, when entering a new patch. This is because the mean for each patch was drawn from a distribution centered on the global mean, and thus the optimal initial guess (prior) for a given patch was the global mean. Thus, at each time step *t*, the value of acquiring one extra data point in the current patch *i* could be estimated in terms of how much it improved future predictions (i.e. how much closer it moved them to *M*), relative to how much sampling a new patch would improve future predictions; that is:

$$V_{i,t} = \frac{1}{N(n-1)} * PE_{i,t} - \frac{1}{N-1} * PE_{i+1,t} \qquad (6)$$

where *N* was the current number of patches exploited so far, *n* the current data points observed in the current patch, $PE_{i,t}$ the next estimated prediction error within the current patch, and $PE_{i+1,t}$ the next estimated prediction error assuming that the agent explored a new patch.

This relative value between staying (exploiting) and switching (exploring) depended therefore on the current position within the game (*n*), the current position within the patch (*N*), and the two variance estimates for the patch mean ($\overline{\sigma^2}$) and the global mean ($\overline{S^2}$), as those variance estimates were used to compute the two prediction errors of interest in the above equation. Given a fixed number of available patches, the assumption that the agent could not return to a patch once it switched away, and values for the rewards *ρ*, *R* and the termination probability λ, our model used dynamic programming to compute the value of the two possible actions – staying and switching – at each time point in the game, based on the states defined by the four quantities: n, N, $\overline{\sigma^2}$, and $\overline{S^2}$.

Results

Figure 16 shows our model results. Compared to an agent that exhausts all available trials in a patch, our “information-sensitive” agent that leaves a patch depending on the relative informativeness of an extra data point within the patch versus a data point in a new patch showed faster learning of the global mean (i.e., learned it in a shorter number of trials, figure 16C). Under certain model parameters (see appendix for details on parameter calibration), it also earned higher overall reward, and it most cases it at least matched if not surpassed the average reward rate of a foraging model that only took into account reward (fig. 16D).

In addition to leading to faster learning of the global environment structure, taking into account the relative value of information had another adaptive consequence, in the form of faster change detection in variable environments. Figure 16E shows that, in environments with a non-zero hazard rate, our model was able to detect and adjust to changes in distribution parameters faster than the model that did not take into account information. (This effect was mitigated at very high hazard rates, which translated into almost random environments – in those cases, learning was nearly impossible given that distribution parameters changed on almost every trial.) Furthermore, in non-stationary environments, our model also earned higher average reward (fig. 16F).

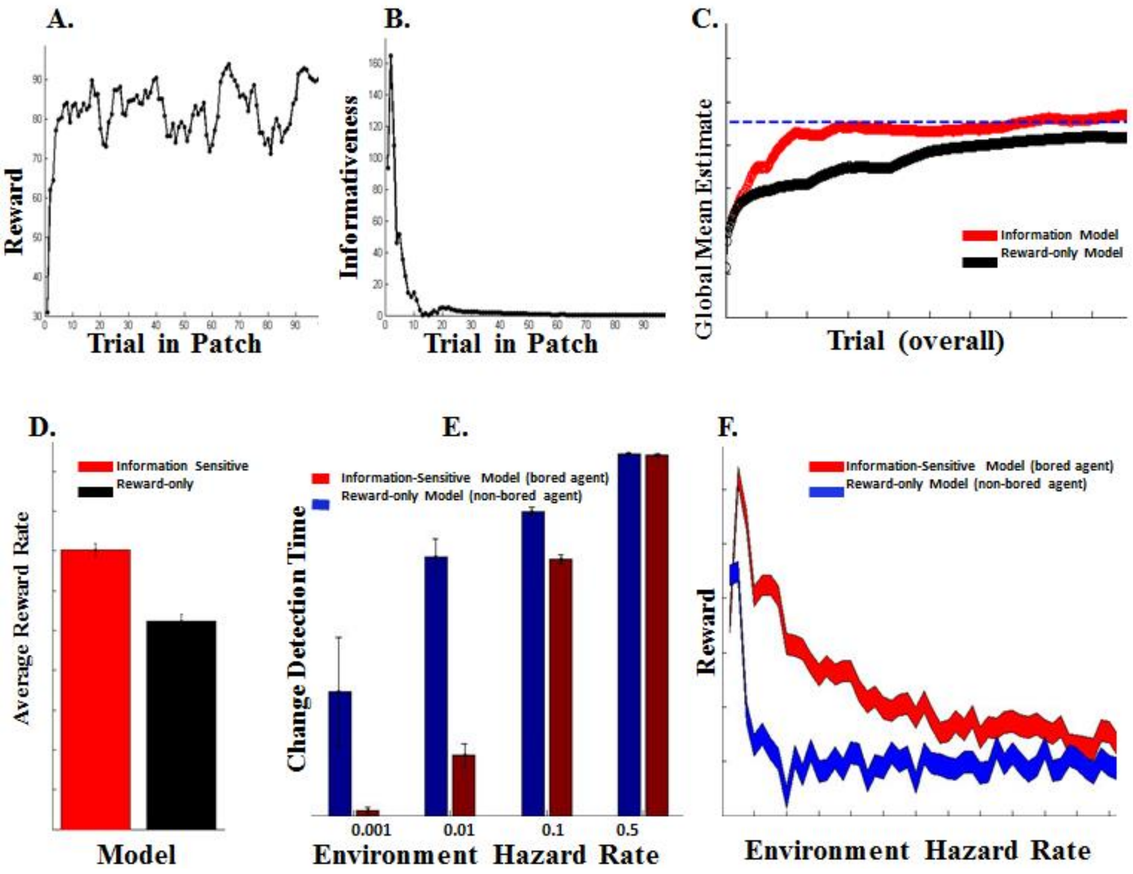


Figure 16: Information-sensitive model tracks reward-relevant information content of current action, relative to global available information content. Poor information content bias it toward switching away (exploration). A: Reward in a patch increased with time spent in that patch. B. Informativeness of each new data point decreased with time spent in patch. C, D: Simulations showed that quitting patches earlier caused model to learn global mean faster (C, red line) and earn higher reward rate (D) than a model that exhaustively exploited increasing patch reward. Dotted blue line in C represents global mean. The simulation included 10 runs, of 100 patches with a maximum of 100 trials within a patch. Error bars: SEM. E, F: Model is also capable of faster change detection if hazard rate is non-zero.

When looking at model predictions for the optimal time to switch away from a patch, as a function of both mean estimated variance of the global mean and, and as a function of time in game (though the two measures are somewhat correlated), the model predicted longer dwell times later in the game, when the uncertainty about the global mean had been significantly reduced.

Discussion

The model proposed in this section represents a first normative account for some of the phenomenology associated with boredom in a cognitive framework. Specifically, it provided a potential explanation for why boredom levels (defined as task disengagement and the increased desire to switch away) could be influenced by the quality of the information content of the current task (as suggested in Studies 4.1 and 4.2), as well as by a relative computation of the benefits of the local versus the global environments (Study 4.3). It also offered an account of why exploration could arise as an adaptive response to decreased informativeness of the current action (as shown in Studies 4.2 and 4.3).

This model assumed that when learning to choose between different reward alternatives, agents should take into account the value of information gained from staying with the current options, relative to the average value of switching away to explore other options. The tradeoff in this framework stemmed from the comparison of how much reward could be generated locally in the patch, compared to how much future reward could be generated by getting more information about other patches. Exploiting a patch until it was terminated would result in rich local reward – however, given the finite-time assumption of the model, choosing to maximize locally could actually result in a lower global reward due to failing to learn the global mean. Depending on the model parameters, the optimal strategy would instead dictate earlier quitting times for early patches (i.e. for the patches that are most informative to learning the global mean), and progressively later quitting times as the variance estimate around the global mean decreases.

This formulation closely parallels the marginal value theorem (MVT; Charnov, 1976) developed in the context of optimal foraging theory. However, our model considers information instead of reward. Under our set of assumptions, information can in fact gain value, if looked at in terms of its usefulness for gaining future reward (an idea also discussed at length in chapter 3). Model results here showed that quitting a current high-reward but low-information patch can in fact still lead to higher overall reward than staying in the uninformative patch; this, however, is highly dependent on the environment structure (such as, the number of available time-steps, as well as the difficulty of the learning problem) and on how well the agent has learned the environment. Primarily, these findings apply to the pre-asymptotic portion of the learning process, when gaining information from exploring patches can contribute to forming better representations of the environment and ultimately to better strategies for gaining reward. (However, given the complexity and breadth of learning processes in the real world, there are sufficiently many scenarios in which humans must deal with pre-asymptotic learning for prolonged periods of time, that the relevance of this theoretical model for real-world scenarios does not suffer.)

The current model explained boredom and exploration from an information-sampling perspective. This is consistent with previous theories, and it qualitatively captures the general direction of results from our first three studies investigating the link between boredom and exploration. However, none of the previous studies presented in this section followed the assumptions of this basic model, and therefore the model cannot make predictions for those participants' behavior. The last section of this chapter discusses an experimental design with all the assumptions from this model, and shows that participants' data qualitatively matches model predictions.

Experiment 4.5 Boredom as an Adaptive Mechanism for Exploration

The following experiment was designed to test model predictions from the model outlined in section 4.4. The main prediction from the model, given the task structure presented to the subjects, was that they would quit high-reward, low-information patches early when the value of gaining information from new patches was higher (i.e., near the beginning of the task, when they had not learned much about the global environment structure), but spend longer and longer in patches as the usefulness of new information decreased.

Methods

Participants

20 participants were recruited from the Princeton community. They gave informed consent, and were compensated for their time at a rate of $12/hour, plus a bonus of up to $5 for better performance. Experiment design and materials were approved by the Princeton IRB.

Procedure

Participants played a game in which they controlled a virtual archer that made his way through enemy territory toward a castle (fig 17, below). The underlying structure of this task was similar to the number-prediction tasks in Studies 4.1 and 4.2, but it reflected the properties of the environment prescribed by the model in section 4.4. The archer's ultimate goal was to defeat an "evil overlord", and it would learn to do so by first facing several waves of the overlord's "minions" on the way to the castle. The way to defeat the overlord (and the minions) was to anticipate where on the screen they would appear, and fire an arrow at the right spot. If the arrow was well-aimed, it would hit the minion, the minion would disappear, and the participant earned one point reward. (This mirrored the number-prediction game: participants could move the archer up and down on the screen as though they moved a slider, and once they settled on a position, they pressed the space key to face the minion for that trial. If their prediction – i.e., the position where they chose to shoot an arrow – was accurate, they were rewarded.) A hit and miss counter was available on the bottom left of the screen.

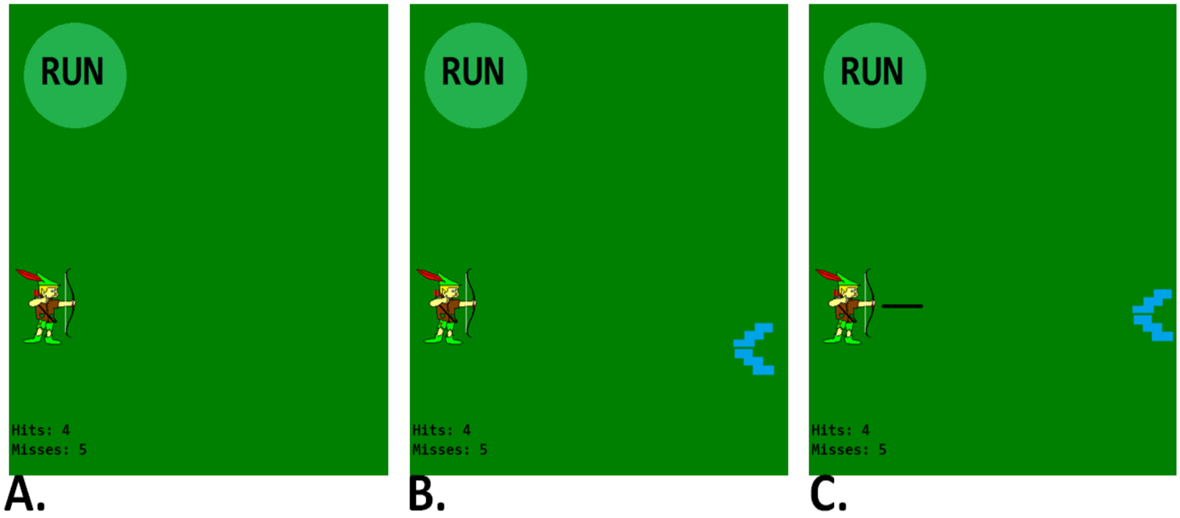


Figure 17: Archer game design. Example of one trial. A: Player adjusts position of archer, and presses space key when ready. B. After space key press, minion comes on screen. C. Archer fires a straight arrow from its current position. If the arrow hits the minion, the player earns one point reward.

The archer had to confront seven waves of minions before facing the overlord; each wave consisted of a maximum of thirty-five minions, which would come out one by one, from the right of the screen (fig 17A). Participants could adjust the archer's firing position on each trial, to better anticipate the minions upcoming location. At the end of seven waves, the archer would have to confront the overlord, and it would have only one shot to either defeat it (i.e., aim the arrow accurately enough to hit it), or be defeated by it (i.e. miss). A reward of 30 points was also available to the participants for defeating the overlord.

Before encountering each minion in a wave, participants had two options. First, they could choose to stay and confront that minion – in which case they would adjust the archer's firing position to their best prediction of where the minion would come from. However, they also had the option of "running away", by pressing the large "RUN" button at the top left of the screen. If they chose to run, the wave of minions would end, and participants would see a screen that announced a new wave (with a new distribution of locations). They would then have the same two options for each minion in the new wave.

Each wave of minions contained information necessary for learning about the overlord, in that the average location of each wave was drawn from a global distribution whose mean was the location of the overlord. However, not all the minions in a wave were equally informative – the informativeness of each data point decreased, as shown in figure 16B – and not all waves were equally informative, as the later ones reduced variance less than the earlier ones. Importantly, participants were told that they had only one hundred and fifty arrows to use on the minions – this operationalized the finite number of steps in our model – so they would have to decide how to use those arrows in a way that would give them the best shot of defeating the overlord, when it came. The model prediction would be that they would use fewer arrows on the earlier waves of minions, to make sure they get enough relevant information about average locations – and that once they had learned more, they could use arrows more liberally to earn points.

Results

Participants learned the task, as evidenced both by their increasing accuracy in targeting the minions, within a wave (figure 18A) and by the increasingly accurate first location estimate – i.e., change of hitting the first minion – in later waves compared earlier waves (significant linear trend, $F(1,6) = 9.42$, $p = 0.02$, figure 18B).

No participants attempted to defeat all minions in a wave. However, participants' average number of minions attempted within a wave increased in later waves compared to earlier waves (figure 18C). This equates to earlier quitting times earlier in the game. Average likelihood to "run" (i.e. quit the current wave and move on to the next one) increased, for all participants, as a function of the number of minions they had confronted in the current wave (i.e the number of time steps they had spent there, fig. 18D). Likelihood to run also increased as a number of minions actually defeated in the current wave (figure 18D, black line).

Participants' likelihood to run was negatively correlated with the informativeness of each new minion for learning about the overlord (figure 18E, where informativeness is calculated according to equation (6) in section 4.4 of this chapter). However, it was positively correlated with the amount left to learn about the overlord (figure 18F, where "amount left to learn" was calculated in terms of the future predicted reductions in uncertainty about the location of the overlord that each new minion wave could bring, see section 4.4 for details).

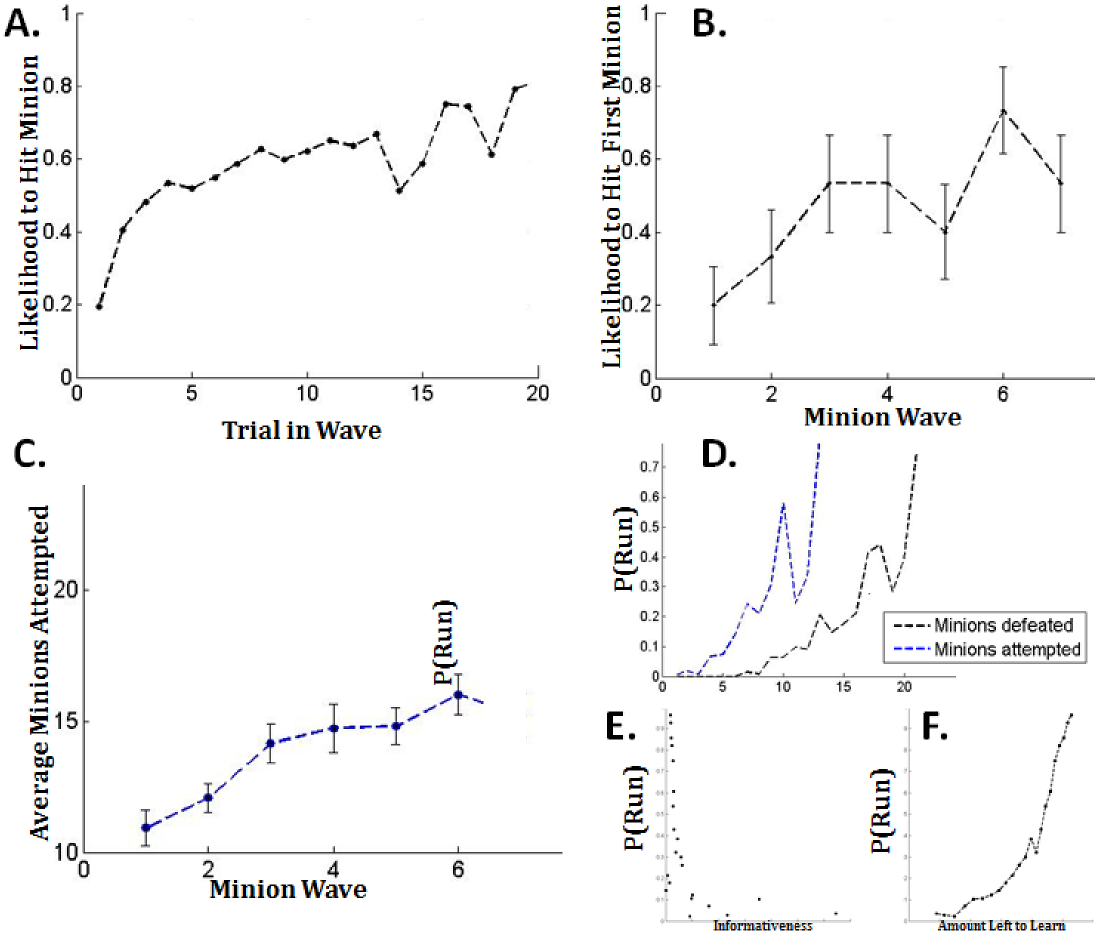


Figure 18: archer task results. Participants showed learning both within-game (A) and between games (B). C. Participants spent longer in a game, as quantified by average number of attempted minions, in later waves compared to earlier waves. D. Participants were more likely to choose to "Run" (i.e. end the current game and get a new minion wave) if they had attempted more minions, but also if they had defeated more minions. The likelihood of choosing to Run was inversely correlated with the informativeness of each new minion (E), but positively correlated with the amount left to learn about the overlord (F).

Discussion

Results from the archer task show that, in an environment that follows the assumption of our information-sampling model from section 4.4, participants show behavior qualitatively consistent with model predictions. Specifically, all participants chose to explore more (i.e. “Run”) earlier in the game than later (fig. 18C), and their probability of exploration was correlated with the information content of a wave, as well as the general information state of the environment (fig. 18 E,F). From a “naive” perspective, this finding is counterintuitive: participants chose to spend slightly *more* time on a given task (i.e., wave of minions) as the experiment progressed; that is, they appear to have become *less* bored as overall time-on-task increased, a reversal of the usual observation. However, this was predicted by the theory, insofar as the relative worth of information vs. immediate reward shifted over the course of the task, with the value of information (presumed to drive boredom) diminishing over time.

That raises the question, then, of why participants in Experiments 4.1 and 4.2 showed an almost opposite trend, getting more bored as time passed. An exact comparison is not possible given the differences in the task structures (Experiment 4.1 did not allow participants to switch away, while in Experiment 4.2 there was no overarching global mean to make quitting a local patch early a tool for better future generalization), and the archer task also took less time to complete than the number prediction task (almost half as long), which likely accounts for at least some difference in overall boredom perception (Danckert & Allman, 2005; London & Monello, 1974; Watt, 1991), though experiment 4.4 did not explicitly assess boredom ratings. However, participants′ consistent tendency to switch away from a boring context not only when switching away is normative (as in experiment 4.4) but also when it is not (Experiment 4.2) suggests the possibility of a global, built-in prior that people might have over the average value of information in different tasks. This is consistent with theories of intrinsic motivation stating that the drive to explore arises from an innate need to interact efficiently with the environment (Deci & Ryan, 1985; White, 1959), as well as with the notion of "flow" and the optimal arousal theory of motivation, according to which organisms seek to balance an internal need for optimal levels of stimulation (Fowler, 1954; Carrol, Zuckerman & Voegel, 1982).

In line with optimal performance, participants' accuracy in later games improved on the first trial of a game (before they get any actual data points from the current minion wave), indicating that they generalized the knowledge about the structure of previous games to make better predictions in the current one. This behavior is consistent with previous findings that humans can indeed learn about both local and global structures simultaneously (Diuk et al. 2013), and it is consistent with our proposed model. Further investigation is needed to establish whether individual participants who explored more on early waves were significantly more accurate in later waves than participants who explored less – which would make an even stronger claim regarding the adaptive role of exploration in this type of environment.

Developing a normative account of the observed phenomenology associated with boredom has been an important goal of the work presented in this chapter. Our theoretical model explained why the failure to extract further information from the current task could lead to boredom, and why in turn the behavioral consequence of the increased boredom would be increased exploration (i.e., switching away). Furthermore, it explained our findings of how people switched away from highly extrinsically-rewarding tasks if they gained no information (Study 4.2). Building on ideas from optimal foraging and reinforcement learning, the model in section 4.4 proposed that it is optimal to track not just reward, but also the amount of information derived from an action, with the goal of simultaneously learning accurate representations of both the global and local environments. This last experiment was designed to test to what extent people's behavior follows our model predictions, and preliminary results showed that participants at least qualitatively match the information-sampling strategies proposed by the model.

The link between exploratory behavior and boredom has been suggested many times in both human and animal literature (Fowler 1954; Vodanovich & Kass 1991; Cohen, McClure &Yu 2007). Our model represents a first attempt to provide a normative account for how boredom might emerge as a consequence of insufficient information-sampling opportunities, and how exploration – even as it moves us away from high-reward options – might constitute an adaptive strategy that ensures we continue to learn useful information. Our experimental findings, together with this theoretical framework, make an important contribution toward the future study of boredom and exploration, and raise important considerations regarding experimental design and the notions of reward, opportunity cost and information.

Chapter 5: General Discussion

Our daily lives regularly confront us with decision problems in which we have numerous alternatives to choose from (such as picking which restaurant to have dinner at on a Friday night, or which movie to watch), and, often, incomplete information about most of the available options. It is therefore essential that our decision-making processes entail an information-sampling component, to allow us a good understanding of the relative value of each option compared to others we could choose, as well as allow us to detect and adaptively respond to potential changes in the values of our decision options. The work presented here has discussed a comprehensive framework for how information sampling strategies can significantly impact the types of learning humans do, determine the efficiency of people's learning and exploration in different contexts, and even relate to behavioral consequences of affective phenomena such as boredom.

Chapter 2 showed that manipulating the way in which information is presented to participants can either speed up or impair their learning of the statistical structure of their environment: the availability of the sampled information, as well as the order in which specific information was experienced, played a significant role in how well participants learned our complex probabilistic task. These results are consistent with previous findings regarding the impact of sequential information accessing in uncertainty reduction (Jacoby et al. 1994; Nelson et al. 2010; Markant & Gureckis 2013), as well as with previous studies that suggest that humans and animals can use near-optimal information-acquisition strategies (Wilde 1980; Dall et al 2005) and adaptively adjust their information sampling strategies based on different environmental demands (Payne et al 1988; van Aahmen et al. 2003).

In Chapter 3, I discussed exploratory behavior as a mechanism for (potentially optimal) information acquisition, and showed how information acquisition strategies can change depending on environment parameters such a decision horizon, risk, and ambiguity. Experimentally, previous results on directed exploration (exploration aimed at gaining information) have been mixed, with some studies finding evidence for this strategy (Meyer & Shi, 1995; Frank et al., 2009; Lee, et al., 2011; Zhang & Yu, 2013) and others failing to do so (Daw et al., 2006; PayzanLeNestour & Bossaerts, 2011). We believe that one reason for these mixed results is the subtle confound between reward and information that arises in sequential choice tasks and makes directed exploration both hard to observe and difficult to confirm. In both experiments 3.1 and 3.2, we removed this confound on the first free-choice trial by manipulating reward and information before subjects made a free choice. This allowed us to unambiguously identify directed exploration on that trial. Due to its structure, the wheel of fortune task in experiment 3.2 could also be used to examine later trials in which the relative information between the two options has remained unchanged (i.e., if the wheel spin did not reveal any of the 'covered' slices).

Chapter 4 addressed exploration from an alternative angle, as the tendency to switch away that stems from finding the current task too boring. Under this framework, the resulting exploratory behavior can in fact be adaptive, if boredom is interpreted as a failure to adequately satisfy the participants' information acquisition needs. In line with recent directions in the study of boredom, the work in chapter 4 treated the phenomenon as cognitive state-related, rather than an affective trait (as also suggested by Speier, Vessich & Valacich 2003; Patyn et al. 2008; Eastwood et al. 2012), and conducted in-depth analyses of the various task-related factors that would lead to the subjective experience of boredom in our participants. Results found that the ability to change future prediction errors by learning about the structure of the task, as well as the perceived opportunity cost of doing the current task (relative to other sources of stimulation available in the environment) significantly impacted our participants' boredom ratings. These results are consistent with the idea that humans need constant access to a certain amount of information in order to maintain a satisfactory level of adaptive behavior (Kuhltam 1991, Zakay 2014), and that that information comes in the form of optimal levels of variability in the environment (Kidd, Piantadosi & Aslin 2012; Garner 2014). Furthermore, they are consistent with recent findings that the subjective experience of 'momentary happiness' results from a combination of obtained reward and prediction error relating to that reward (Rutledge et al. 2014), and further strengthen the notion that boredom could emerge from a computation that combines extrinsic value (as reward) and intrinsic value (information, as related to prediction error).

Building on that, I proposed a quantitative model that described the value of information sampling in terms of future reward, and showed how, if the information available in the current task was insufficient, it would in fact be more rewarding in the long-term to switch away even from a locally rewarding task. The idea that boredom can lead to useful exploration has been previously suggested in the literature (Vodanovich & Kass 1990; Cohen, McClure & Yu 2007), and theoretical work in machine learning has shown that artificial agents in simulated environments can learn faster and learn more complex actions if they are capable of boredom (Schmidhuber 1997; Simsek & Barto 2005), but those findings had not previously been tested in human participants. Furthermore, while previous studies have suggested the need to develop a cognitive model of boredom (Hill & Perkins 1988; Perkins & Hill 1989; Eastwood et al. 2012), to our knowledge, the model in section 4.4 is the first instance of providing an explicit, normative account for exploration as an adaptive response to boredom induced by insufficient information-acquisition opportunities.

Finally, it should be noted that, in the entire body of work shown in this dissertation, I discuss information acquisition as a purely experience-based phenomenon: all participants had to actually explore an available option in order to learn about its reward structure. While this is often the case in the real world, direct experience is also not the only way in which information can be gained. It is also possible to learn about available alternatives without directly sampling them, for instance, from description (such as reading newspaper reviews of different restaurants or movies), or from witnessing someone else's actions and noticing what rewards they obtain. There is ample evidence that obtaining information by description rather than experience, produces significantly different behavioral patterns (Hertwig et al. 2004, Newell & Rakow 2007, Ludvig & Spetch 2011), and that learning by example or instruction also produces different results than learning by direct trial-and-error (Love 2002; Murata et al. 2002). It is therefore entirely possible that different means of information acquisition would lead to different learning patterns (for instance, children learning tool-use by example behave differently than those learning by trying it themselves - Brown & Kane 1988, Zhao & Liu 2007), prescribe different optimal strategies (for instance, the value of exploration changes significantly when it is possible to learn about non-current options without actually sampling them – Smith et al. 2009) and even differently impact people's subjective experiences such as boredom. Examining information-acquisition from description, instruction or example is outside the scope of the present work – but, as the real world likely presents us with information through a combination of these different methods, further research into the function and efficiency of exploration in different informational contexts would be desirable.

References

Abdellaoui, M., Bleichrodt, H., Paraschiv, C. (2007). Loss aversion under prospect theory: A parameter-free measurement. *Management Science*, 53 (10), 1659-1674.

Adams, R.B. Jr., Gordon, H.L., Baird, A.A., Ambady, N. & Kleck, R. E. (2003). Effects of gaze on amygdala sensitivity to anger and fear faces. *Science,* 300 (5625), 1536.

Aleven, V., Sewall, J., McLaren, B. M., & Koedinger, K. R. (2006, July). Rapid authoring of intelligent tutors for real-world and experimental use. In *Advanced Learning Technologies, 2006. Sixth International Conference on* (pp. 847-851). IEEE.

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial intelligence*, *42*(1), 7-49.

Aston-Jones G, Rajkowski J, Kubiak P, Alexinsky T. (1994). Locus coeruleus neurons in the monkey are selectively activated by attended stimuli in a vigilance task. *J. Neurosci.* 14:4467–80

Aston-Jones, G. & Cohen, J.D. (2005). An integrative theory of locus coeruleus - norepinephrine function: adaptive gain and optimal performance. *Ann Rev. Neurosci* 28: 403- 50

Atkinson, A., Doney, A., & Tobias, R. (2007). Optimum Experimental Designs, with SAS. Atkinson RJCA, Hand DJ, Pierce DA, Schervish MJ, Titterington DM, editors.

Auer, P. & Ortner, R. (2010) UCB revisited: improved regret bounds for the stochastic multi- armed bandit problem

Auer, P., Cesa-Bianchi, N. & Fischer, P. (2002) Finite-time analysis of the multi-armed bandit problem. *Machine Learning* 47, 235 - 256

Bach, D.R., Hulme, O., Penny, W.D. & Dolan, R.J. (2011) The known unknowns: neural representation of second-order uncertainty, and ambiguity. *J Neurosci*. 31, 4811 - 4820

Bach, D.R., Seymour B. & Dolan, R.J. (2009) Neural activity associated with the passive prediction of ambiguity and risk for aversive events. *J. Neurosci* 29, 648 - 656

Banks, J. S., & Sundaram, R. K. (1994). Switching costs and the Gittins index. *Econometrica: Journal of the Economical Society, 62*, 687–694.

Barto, A. G., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. *Proceedings of the 3rd International Conference on Development and Learning (ICDL 2004)* 112–19

Baum, W. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior , 32*, 269-281

Beck, J., Woolf, B. P., & Beal, C. R. (2000). ADVISOR: A machine learning architecture for intelligent tutor construction. *AAAI/IAAI*, *2000*, 552-557.

Behrens. T.E.J., Woolrich, M.W., Walton, M.E. * Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience* 10: 1214 - 1221

Bellman, R. (1957). *Dynamic programming*. Princeton: Princeton University Press.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205). Cambridge, MA: MIT Press

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, *97*(1), 245-271.

Bossaerts, P., & Plott, C. (2004) Basic principles of asset pricing theory: evidence from large-scale experimental financial markets. *Rev. of Finance*, 8: 135–169

Brown, S., Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58, 49-67.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171-189.

Caraco, T. (1980). On foraging time allocation in a stochastic environment. *Ecology , 61*, 119-128.

Carrol, E. N., Zuckerman, M., & Vogel, W. H. (1982). A test of the optimal level of arousal theory of sensation seeking. *Journal of Personality and Social Psychology*, *42*(3), 572.

Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., & Zhu, X. (2009). Human active learning. In: *Advances in neural information processing systems* (pp. 241-248).

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences*, *106*(21), 8719-8724.

Coenen, A., Rehder, B., & Gureckis, T. Decisions to intervene on causal systems are adaptively selected.

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

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of artificial intelligence research*.

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.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, *121*(3), 364.

Damrad-Frye, R., & Laird, J. D. (1989). The experience of boredom: The role of the self-perception of attention. *Journal of Personality and Social Psychology*,*57*(2), 315.

Danckert, J. A., & Allman, A. A. A. (2005). Time flies when you're having fun: Temporal estimation and the experience of boredom. *Brain and cognition*, *59*(3), 236-245.

Daw, N. D., O'Doherty, J. P., Seymour, B., Dayan, P. & Dolan, R. J. (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.

Dias, R., Robbins, T. W., & Roberts, A. C. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. *Nature, 380*(6569), 69-72.

Doucet, A., de Freitas, N., & Gordon, N. (2001). Sequential Monte Carlo methods in practice. New York: Springer.

Doya, K. (Ed.). (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The Unengaged mind defining boredom in terms of attention. *Perspectives on Psychological Science*, *7*(5), 482-495.

Eckstein, M. P., Abbey, C. K., Pham, B. T., & Shimozaki, S. S. (2004). Perceptual learning through optimization of attentional weighting: Human versus optimal Bayesian learner. *Journal of Vision*, *4*(12), 3.

Elliott, R., Newman, J.L., Longe, O.A., Deakin, J.F. (2003) Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study. *J Neurosci* 23: 303–307

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

Ferster, C.B. & Skinner, B.F. (1957) *Schedules of Reinforcement*. New York, Appleton-Century-Crofts.

Frank, M.J., Doll, B.B., Oas-Terpstra, J. & Moreno, F. (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience* 12: 1062 - 1068

Gaffan, D., Murray, E.A., Fabre-Thorpe, M. (1993). Interaction of the Amygdala with the Frontal Lobe in Reward Memory. *European Journal of Neuroscience*, 5 (7), 968-975.

Gallistel, C.R. (1990). *The Organization of Learning*. Cambridge, MA: Bradford Books/MIT Press.

Gallistel, C.R. et al. (2007). Is matching innate? *Journal of the Experimental Analysis of Behavior*. 87: 161–199

Geana, A., Wilson, R. C., & Cohen, J. D. (2013, November). *Time, Risk and Ambiguity in Human Exploration: A Wheel of Fortune Task.* Poster presented at the first meeting of the Reinforcement Learning and Decision-Making conference, Princeton, NJ

Geana, A., Wilson, R. C., & Cohen, J. D. (2014, November). *Decision Horizon, Risk and Ambiguity in Human Exploration: A Wheel of Fortune Task.* Poster presented at the annual meeting of the Society for Neuroscience, Washington, DC.

Geana, A., Wilson, R. C., & Cohen, J. D. (2014, March). *Decision Horizon Risk and Ambiguity in Human Exploration: A Wheel of Fortune Task.* Workshop talk. In *Information Sampling and Optimization* workshop (Nassar, Wilson). Presented at the annual meeting of the Computational Systems Neuroscience (Cosyne), Snowbird, UT.

Gelly, S., & Wang, Y. (2006, December). Exploration exploitation in go: UCT for Monte-Carlo go. In *NIPS: Neural Information Processing Systems Conference On-line trading of Exploration and Exploitation Workshop*.

Ghirardato, P., Maccheroni, F. & Marinacci, M. (2004) Differentiating ambiguity and ambiguity attitude. *J. Econ. Theory*, 118: 133–173

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society*, Series B, 41, 148-177

Gittins, J. C. (1989). Multi-band Bandit Allocation Indices..

Gittins, J. C. & Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in statistics* (ed. J. Gans), pp. 241–266. Amsterdam, The

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*(2), 299-308.

Gureckis, T., & Markant, D. (2009). Active learning strategies in a spatial concept learning game. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 3145-3150).

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: a cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464-481.

Harris, M. B. (2000). Correlates and Characteristics of Boredom Proneness and Boredom1. *Journal of Applied Social Psychology, 30*(3), 576-598.

Herrnstein, R. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior , 4*, 267.

Herrnstein, R., & Heyman, G. (1979). Is matching compatible with reinforcement maximization on concurrent variable interval variable? *Journal of the Experimental Analysis of Behavior , 31*, 209-223.

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior*, 24, 107-116.

Heyman, G., & Herrnstein, R. (1986). More on concurrent interval-ratio schedules: A replication and review. *Journal of the Experimental Analysis of Behavior , 46*, 331– 351.

Hill, A. B., & Perkins, R. E. (1985). Towards a model of boredom. *British Journal of Psychology*, *76*(2), 235-240.

Hills, T. T., & Hertwig, R. (2010). Information Search in Decisions From Experience Do Our Patterns of Sampling Foreshadow Our Decisions?.*Psychological Science*.

Hogarth, R.M. & Einhorn H.J. (1990) Venture theory - a model of decision weights. *Management Science,* 36: 780 - 803

Holt, C.A., & Laury, S.K. (2002) Risk aversion and incentive effects. *The American Economic Review,* 92 (5), 1644-1656.

Hsee, C. K., Yang, A. X., & Wang, L. (2010). Idleness aversion and the need for justifiable busyness. *Psychological Science*, *21*(7), 926-930.

Hsu, M., Bhatt, M., Adolphs R., Tranel, D., Camerer, C.F. (2005) Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680–1683

Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T. & Platt, M. L. (2006) Neural signatures of economics preferences for risk and ambiguity. *Neuron* 49, 765 – 775

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: evidence for the adaptive gain theory. *Journal of cognitive neuroscience*, *23*(7), 1587-1596.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2011). Don't stop 'til you get enough: adaptive information sampling in a visuomotor estimation task. In *Proc. of the 33rd Annual Conf. of the Cognitive Science Society (eds L. Carlson, C. Hölscher & T. Shipley). Austin, TX: Cognitive Science Society*.

Kaelbling, L. P., Littman, M. L. & Moore, A. W. (1996) Reinforcement learning: a survey. *Journal of Artificial Intelligence Researcg*. 4, 237–285.

Kass, S. J., Vodanovich, S. J., Stanny, C. J., & Taylor, T. M. (2001). Watching the clock: boredom and vigilance performance. *Perceptual and motor skills*.

Knock, T. R., Ballard, K. J., Robin, D. A., & Schmidt, R. A. (2000). Influence of order of stimulus presentation on speech motor learning: A principled approach to treatment for apraxia of speech. *Aphasiology*, *14*(5-6), 653-668.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2011). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in psychology,2*.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244-247.

Krebs, J. R., Kacelnik, A. & Taylor, P. (1978) Tests of optimal sampling by foraging great tits. *Nature* 275, 27–31Lai & Robbins 1985

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210-226.

Kwok, N. M., Fang, G., & Zhou, W. (2005, August). Evolutionary particle filter: re-sampling from the genetic algorithm perspective. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on* (pp. 2935-2940). IEEE.

Levy, I., Snell, J., Nelson, A.J., Rustichini, A. & Glimcher, P.W. (2009) Neural representation of subjective value under risk and ambiguity. *JN Physiol*. 103, 1036 - 1047

Logue, A. W., Forzano, L. B., & Tobin, H. (1992). Independence of reinforcer amount and delay: The generalized matching law and self-control in humans.*Learning and Motivation*, *23*(3), 326-342.

London, H., & Monello, L. (1974). Cognitive manipulation of boredom.

Mahoney, K.T., Buboltz, W., Levin, I.P., Doverspike, D. & Syvantek, D.J. (2011) Individual differences in a within-subject risky-choice frame. *Personality and Individual Differences*. 51: 248 - 257

Margoliash, D. (1986). Preference for autogenous song by auditory neurons in a song system nucleus of the white-crowned sparrow. *The Journal of neuroscience*, *6*(6), 1643-1661.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Montague, P.R., Dayan, P., Person, C. & Sejnowski, T.J. (1995) Bee foraging in uncertain environments using predictive hebbian learning. *Nature* 377, 725–728

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, *30*(37), 12366-12378.

Navarro, J., & Osiurak, F. (2015). When Do We Use Automatic Tools Rather Than Doing a Task Manually? Influence of Automatic Tool Speed. *The American Journal of Psychology*, *128*(1), 77-88.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological science*, *21*(7), 960-969.

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior*, *10*(1), 5-24.

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

O'Doherty, J., Critchley, H., Deichmann, R., Dolan, R.J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *J Neurosci* 23: 7931–7939

Otto, A.R., Markman, A.B., Gurekis, T.M. & Love, B.C. (2010) Regulatory fit and systematic exploration in a dynamic decision-making environment. *J Exp Psych LMC* 36: 797 - 804

Pattyn, N., Neyt, X., Henderickx, D., & Soetens, E. (2008). Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue?. *Physiology & Behavior*, *93*(1), 369-378.

Payne, J. W., Bettman, J. R., & Luce, M. F. (1996). When time is money: Decision behavior under opportunity-cost time pressure. *Organizational behavior and human decision processes*, *66*(2), 131-152.

Perkins, R. E., & Hill, A. B. (1985). Cognitive and affective aspects of boredom. *British Journal of Psychology*, *76*(2), 221-234.

Pindyck, R.S. (1978) The optimal exploration and production of nonrenewable resources. *Journal of Political Economy,* 86 (5), 841 – 861.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological review*, *109*(3), 472.

Pratt, J. W. (1964). Risk Aversion in the Small and in the Large. *Econometrica*, 32,122-136.

Prechelt, L. (1998). Early stopping-but when?. In *Neural Networks: Tricks of the trade* (pp. 55-69). Springer Berlin Heidelberg.

Preuschoff, K., Bossaerts, P, & Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron,* 51, 381-390.

Rafferty, A. N., Zaharia, M., & Griffiths, T. L. (2012). Optimally designing games for cognitive science research. In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 893-898).

Reed, P., Schachtman, T.R. & Hall, G. (1987) Overshadowing and potentiation of instrumental responding in rats as a function of the schedule of reinforcement. *Learning and Motivation*. 19: 13 - 30

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In*Proceedings of the twenty-seventh annual conference of the cognitive science society. Mahwah, NJ: Erlbaum*.

Ritter, F. E. (2007). *In order to learn: How the sequence of topics influences learning*. Oxford University Press, USA.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of Australian Mathematical Society,* 55, 527-535

Rust, R. T., Simester, D., Brodie, R. J., & Nilikant, V. (1995). Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combinations of criteria. *Management Science*, *41*(2), 322-333.

Sarle, W. S. (1995). Stopped training and other remedies for overfitting. In*Proc. of the 27th symposium on the interface of computing science and statistics* (pp. 352-360).

Sebastiani, P., & Wynn, H. P. (2000). Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(1), 145-157.

Servan-Schreiber D., Printz H., and Cohen J.D. (1990). A network model of catecholamine effects: gain, signal-to-noise ratio and behavior *Science*, 249, 892-895

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394-3411.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1-42.

Şimşek, O., & Barto, A. (2006). An intrinsic reward mechanism for efficient exploration. *Proceedings of the 23rd international conference*, 833-840.

Smallwood, J., Fishman, D. J., & Schooler, J. W. (2007). Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review*, *14*(2), 230-236

Steyvers, M., Lee, M.D., & Wagenmakers, E.J. (2009) A Bayesian Analysis of Human Decision-Making on Bandit Problems. *Journal of Mathematical Psychology*, 53, 168-179

Sugrue, L.P., Corrado, G.S. & Newsome, W.T. (2004) Matching behavior and the representation of value in the parietal cortex. *Science*. 304: 1782 - 1787

Sutton, R.S., & Barto, A.G. (1998) Introduction to Reinforcement Learning. MIT Press, Cambridge, MA.

Sutton, R. S. (1984). Temporal credit assignment in reinforcement learning.

Tokic, M. & Palm, G. (2011) Value-difference based exploration: adaptive control between ε-greedy and softmax. *Advances in Artificial Intelligence* 7006, 335 - 346

Tokic, M. (2010) Adaptive ε-greedy exploration in reinforcement learning based on value differences. *Advances in Artificial Intelligence* 6359, 203 - 210

Trommershäuser, J., Gepshtein, S., Maloney, L. T., Landy, M. S., & Banks, M. S. (2005). Optimal compensation for changes in task-relevant movement variability. *The Journal of Neuroscience*, *25*(31), 7169-7178.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003). Statistical decision theory and the selection of rapid, goal-directed movements. *JOSA A*,*20*(7), 1419-1433.

Tversky, A., & Kahneman, D. (1986) Rational Choice and the Framing of Decisions. *The Journal of Business*, 59 (4), S251-S278.

Vodanovich, S. J., & Kass, S. J. (1990). A factor analytic study of the boredom proneness scale. *Journal of Personality Assessment*, *55*(1-2), 115-123.

Von Neumann, J., & Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton University Press; Princeton, NJ.

Wallace, J. C., Vodanovich, S. J., & Restino, B. M. (2003). Predicting cognitive failures from boredom proneness and daytime sleepiness scores: An investigation within military and undergraduate samples. *Personality and Individual Differences*, *34*(4), 635-644.

Watt, J. D. (1991). Effect of boredom proneness on time perception.*Psychological Reports*, *69*(1), 323-327.

Watt, J. D., & Hargis, M. B. (2010). Boredom proneness: Its relationship with subjective underemployment, perceived organizational support, and job performance. *Journal of business and psychology*, *25*(1), 163-174.

Weissinger, E., Caldwell, L. L., & Bandalos, D. L. (1992). Relation between intrinsic motivation and boredom in leisure time. *Leisure Sciences*, *14*(4), 317-325.

Whalen, P.J. (1998) Fear, vigilance and ambiguity: initial neuroimaging studies of the human amygdala. *Curr Dir Psychol Sci* 7:177–188

Whittle, P. (1980). Multi-Armed bandit tasks and the Gittins Index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42 (2), 143-149.

Whittle, P. (1988). Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, *25*A, 287–298.

Wilson, R., & Niv, Y. (2012). Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*. 5:189

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Wilson, R. C., & Cohen, J. D. (2014). Humans tradeoff information seeking and randomness in explore-exploit decisions *Society for Neuroscience Abstracts*

Wolf, C., & Pohlman, L. (1983). The Recovery of Risk Preferences from Actual Choices. *Econometrica*, 51 (3), 843 – 850.

Wu, S. W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making compared with an equivalent motor task. *Proceedings of the National Academy of Sciences*, *106*(15), 6088-6093.

Yi, M.S.K., Steyvers M. & Lee, M. D. (2009) Modeling human performance in restless bandits with particle filters. *The Journal of Problem Solving*, 2: 81 - 101

Yu, A. & Dayan, P. (2005) Uncertainty, neuromodulation and attention. *Neuron* 46, 681–692.

Zhang, S. & Lee, M. D. (2010) Human and Optimal Exploration and Exploitation in Bandit Problems. *Journal of Mathematical Psychology* 54 (6), 499-508

Appendix

Experiment 4.2 - Boredom Proneness Scale (Sundberg & Farmer 1986)

The statements can be answered using a true-false response (the original format used) or with a 7-point format from "1" (highly disagree) to "7" (highly agree) used in recent research.

_____ 1. It is easy for me to concentrate on my activities.

_____ 2. Frequently when I am working I find myself worryingabout other things.

_____ 3. Time always seems to be passing slowly.

_____ 4. I often find myself at "loose ends", not knowing what to do.

_____ 5. I am often trapped in situations where I have to do meaningless things.

_____ 6. Having to look at someone's home movies or travel slides bores me tremendously.

_____ 7. I have projects in mind all the time, things to do.

_____ 8. I find it easy to entertain myself.

_____ 9. Many things I have to do are repetitive and monotonous.

_____ 10. It takes more stimulation to get me going than most people.

_____ 11. I get a kick out of most things I do.

_____ 12. I am seldom excited about my work.

_____ 13. In any situation I can usually find something to do or see to keep me interested.

_____ 14. Much of the time I just sit around doing nothing.

_____ 15. I am good at waiting patiently.

_____ 16. I often find myself with nothing to do, time on my hands.

_____ 17. In situations where I have to wait, such as a line I get very restless.

_____ 18. I often wake up with a new idea.

_____ 19. It would be very hard for me to find a job that is exciting enough.

_____ 20. I would like more challenging things to do in life.

_____ 21. I feel that I am working below my abilities most of the time.

_____ 22. Many people would say that I am a creative or imaginative person.

_____ 23. I have so many interests, I don't have time to do everything.

_____ 24. Among my friends, I am the one who keeps doing something the longest.

_____ 25. Unless I am doing something exciting, even dangerous, I feel half-dead and dull.

_____ 26. It takes a lot of change and variety to keep me really happy.

_____ 27. It seems that the same things are on television or the movies all the time; it's getting old.

_____ 28. When I was young, I was often in monotonous and tiresome situations.

Experiment 4.3 – Full list of questions

These questions were given to subjects in the interval between two games, in part A of the task – i.e, while they played the bandit task. The first two questions were the questions of interest, and they were given to all subjects every nine games (a total of seven times throughout the bandit task). The other five questions represented distractors, and were mixed randomly in the other inter-game intervals.

Questions:

- Please rate how difficult you are finding this task, on a scale from 1 (not difficult at all) to 10 (extremely difficult)
- Please rate how interesting you are finding this task, on a scale from 1 (not interesting at all, very boring) to 10 (extremely interesting)
- How many points do you think you have obtained in this task so far? (Please enter a number in the box below)
- Please rate your estimated performance in this task, on a scale from 1 (I think I'm not performing well at all) to 10 (I think I'm performing extremely well)
- How long do you think you've been playing this task? (Please enter the number of minutes in the box below)
- What is the average number of points you think you obtained in a game, so far? (Please enter the number in the box below)
- Please rate how well you think you have learned this task, on a scale from 1 (I have not learned it at all) to 10 (I have learned it extremely well)