Causes and Consequences of Exploratory Choice

by

Alexander S. Rich

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Psychology
New York University
May 2018

Todd M. Gureckis

ABSTRACT

Life is full of dilemmas between exploring uncertain alternatives and exploiting alternatives known to be rewarding. These choices about when and what to explore determine not just the number of positive experiences people have over time, but also the data about their environment that people observe and what they consequently learn. Because of this, patterns of exploration, and particularly under-exploration, can lead to biased decisions and false beliefs. In three studies, this dissertation investigates the factors that determine how persistently people explore, and the biases that can result when they do not explore enough. The first study shows that people can judge the future value of the information that could be gained by exploring, and will modulate their degree of exploration based on the long-term costs of potential false beliefs. The second study investigates whether people's exploration might be inhibited by a motivational overweighting of present rewards, but finds no evidence of this effect. Finally, the third study provides a broader perspective of how exploratory choices and learning processes interact to produce false beliefs, a phenomenon we term a *learning trap*, and demonstrates a novel learning trap experimentally. This connection of exploratory choice to data-driven learning biases can be situated within the broader literature on sampling biases in psychology, and also has important implications for the understanding of recently uncovered biases in machine learning.

Contents

List of Figures

LIST OF FIGURES

INTRODUCTION

In his book Graphic Discovery, Howard Wainer describes an analysis of people's average age of death based on their birth year, using data collected from the Princeton, New Jersey, Cemetery (Wainer, 2005). His surprising finding was that life span held fairly stable until 1920, then dropped precipitously as the birth year drew closer to the present day. Why did people born in the 1980s die so much younger than those born in the 1910s? The answer, of course, is that the cemetery only provides data on the life spans of people who died before the present day, which includes most people born in the 1920s but only those who died young born in the 1980s. The moral of the story is that where data comes from can play an outsize role in the conclusions that can be drawn from it.

The idea that the provenance of data matters is central to statistics. Data produced through random sampling is the statistician's ideal (Diez, Barr, & Cetinkaya-Rundel, 2015). In random sampling each element of the environment, or sample, has equal probability of being observed. This creates data that tends, if one collects enough of it, to be representative of the true statistics of the world. To estimate the average life span of Princeton residents born in 1925 through random sampling, for example, one would randomly select a subset of the people born in that year, and then find the average life span of the sampled individuals.

In non-random sampling some mechanism other than randomness determines which elements of the environment are observed and which are not. In Wainer's case, that mechanism was to collect data at the cemetery, thus including in his sample all individuals who were buried before the date of the visit while excluding those who were not. Depending on its form, non-random sampling can cause parts of the environment to be over- or under-represented, or to be missing entirely. Whenever non-random sampling occurs, there is a risk of incorrect, incomplete, or otherwise biased learning, because what is true of the included samples might not be true of the environment as a whole.

The concept of sampling, while native to statistics, can be a powerful perspective through which to understand biases in everyday human behavior. As people interact with their environment, they constantly sample and learn from all kinds of data—data about what ice cream flavors they like, the personalities of their friends and acquaintances, or how much air they can blow into a balloon before it pops. These samples are often non-random, sometimes because of external factors of the environment and sometimes through people's own actions.

Over several decades, researchers in psychology and organizational learning have documented and unpacked the ways in which non-random sampling can lead to biases in everyone from oncologists to college admissions committees (K Fiedler, 2000; R. M. Hogarth, Lejarraga, & Soyer, 2015; Swets, Dawes, & Monahan, 2000). One particularly pervasive class of sampling biases, and the focus of the body of this dissertation, develops through the simple act of seeking out rewarding experiences (like eating blueberry ice cream) while avoiding negative ones (like eating durian ice cream). Reward-seeking behavior leads people to collect many samples of items *believed* to be good, and few of those believed to be bad, which can lead to beliefs that are persistently out of sync with what is actually good and bad. To form more accurate beliefs, people must take a break from exploiting what they believe to be the best options in order to explore more broadly, often by taking a more random and representative sample of the available options. This explore–exploit tradeoff is one of the central problems of the academic field of reinforcement learning (Sutton & Barto, 1998; Mehlhorn et al., 2015).

As a prelude to the later empirical chapters, this introduction synthesizes the literature on how non-random sampling affects people's beliefs and decisions, paying special attention to sampling biases that relate to reward-seeking behavior and the explore-exploit tradeoff. The upshot of this body of work is that even when the way we draw conclusions from our observations is reasonable, or even optimal, the qualities of the data itself can doom us to misunderstanding the world.

When research on sampling biases began, the effects of these biases on human welfare and society were mainly manifest through the beliefs and decisions of people. But in recent years machine learning systems have become increasingly ubiquitous and have been tasked with increasingly consequential decisions. An explosion of new research has shown that they too can instantiate sampling biases, in automated and large-scale ways, and often in the the exact same domains where human biases have long been observed (Campolo, Sanfilippo, Whittaker, & Crawford, 2017; O'Neil, 2017). Thus, before turning to the empirical portion of the dissertation, the final section of the introduction focuses on recent work on bias in data science and machine learning, and seeks to build connections between insights from psychology and this young field.

0.1 Biases from non-random sampling

Psychology has a long history of understanding human behavior as shaped not just by the capabilities of the mind but by the properties of the environment (Simon, 1990). This approach has lead researchers to understand many judgment and decision making biases as consequences of how the structure of the environment leads some samples of data to be observed while others are hidden (K Fiedler, 2000; R. M. Hogarth et al., 2015).

One of the most common ways in which the environment is unevenly revealed is through conditional or selective sampling, in which an observer only encounters information for items that satisfy a certain condition. Howard Wainer, for instance, only observed life spans for individuals on the condition that they had already been buried in the cemetery. In many cases, data that has not been produced by conditional sampling is unavailable. For example, Denrell (2003) pointed out that people can observe the practices of existing companies, but not of those that have gone out of business. Because of this, they are likely to reach incorrect conclusions about the management practices that lead to success. In particular, taking big risks, even if it does not reliably pay off, tends to appear effective among surviving companies because the risk-takers who failed were driven out of business. Similarly, underperforming mutual funds tend to disappear, leaving behind only those that have done well (whether by skillful management or by luck) and producing an overly rosy picture of fund performance (Elton, Gruber, & Blake, 1996). Conditional sampling also occurs in the advertisement of mutual funds, where companies tend to display only their better-performing funds. Investors, including financial professionals, tend not to adjust their expectations to take this selection into account (Koehler & Mercer, 2009).

In other situations, conditional sampling occurs through a decision maker's own actions. Einhorn and Robin M. Hogarth (1978) posited that in such situations, people may become overconfident in their ability to take effective action due to a lack of feedback about unselected samples. If there are many qualified applicants for every space at a college, for instance, an admissions committee might observe that admitted freshmen do well and conclude that its rules for admissions are excellent. But the committee never gets to observe the papers written, jobs landed, or prizes won by the students who were rejected. Their confidence might be shaken if they were able to observe an alternative freshman class and see that they performed nearly as well, or even better, than those who were actually admitted.

Information about unselected samples may also be theoretically available but still remain unexamined. For example, a doctor may look at the symptoms and lifestyle of patients whom she has diagnosed with a disease to find the disease's cause, without collecting the necessary data to test how common these properties are in the healthy population (Robyn M Dawes, 1993). Surprisingly simple rules can often be formulated using the full set of information that outperform the judgments of expert clinicians, who have years of experience but have relied on conditional samples (R M Dawes, Faust, & Meehl, 1989).

People may use and be biased by conditional sampling even when all information is directly available. In a set of experiments, Klaus Fiedler, Brinkmann, Betsch, and Wild

(2000) asked participants to estimate the probability that a patient had breast cancer given that she had a positive mammogram. This kind of judgments can be difficult for clinicians and laypeople alike because of the tendency to consider only the probability of a positive test result given that the disease is present, without considering the number of people without the disease who produce false positives (Swets et al., 2000). But in these experiments, participants were given access to a set of representative index cards that each had a single patient's mammogram result on one side and their true breast cancer status on the other, which could have allowed participants to collect a random sample and draw correct conclusions.

Even in this situation, sample bias sometimes remained. When index cards were organized in drawers by mammogram result, participants could easily collect a sample of patients with positive mammograms and ascertain what proportion had breast cancer. In contrast, when index cards were organized by cancer status participants tended to sample an equal number from each group, over-representing patients with cancer and creating an erroneous estimate. While participants could have produced an less biased sample by selecting fewer index cards of patients with cancer, this would have produced its own set of learning problems due to the resulting small sample size (Klaus Fiedler, 2008). The only robust solution would be to select a biased sample and correct the resulting estimates after the fact, a calculation that seems difficult for people to achieve. These findings further point to the fact that sampling biases are integrated deeply into the way people come to understand the world, and that data-driven, rule-based decision making, not personal experience, may be the best guide to medical and other clinical judgments (Swets et al., 2000).

Conditional sampling is one of the most common and widely studied forms of non-random sampling, but it is far from the only one. People are also exposed to censored samples, in which outcomes for all examples are observed, but the outcomes themselves are cut off at some value (R. M. Hogarth et al., 2015; Feiler, Tong, & Larrick, 2012). For example, a manager might observe exactly how much work employees completed when they fail to complete their assigned tasks, but not how much more work they were capable of when they do complete their assignments. Similarly, a shopkeeper will know exactly how much demand there was for her products on days when they don't sell out, but not when they do. While observers of censored samples should adjust their estimates to take into account the unobservable values, Feiler et al. (2012) found that they do not do so sufficiently, and rely too heavily on the distribution of observable values. This means that managers and shopkeeper may systematically underestimate the capabilities of their workers and the demand for their products.

Even groups of samples that are simply of unequal sizes can act to produce decision biases. If a teacher starts out believing every student has an intermediate value of intelligence, and incrementally updates those beliefs each time a student answers a question, then the teacher can end up with more extreme intelligence beliefs for students who raise their hands often and answer many questions than for those who raise their hands rarely. While this kind of updating is actually a sound learning strategy consistent with Bayesian statistics, it can lead to harmful outcomes like illusory correlations in which rare minority groups are falsely believed to have worse (or better) attributes than common majority groups (Klaus Fiedler, Walther, Freytag, & Plessner, 2002; Klaus Fiedler & Unkelbach, 2014). (Other researchers have provided an alternative account of illusory correlations based on conditional sampling (Denrell & Le Mens, 2011).) People may also change the number of samples they collect about each item on the fly, stopping information-gathering when they have enough information to make a decision, and this adaptive sample size has the potential to warp later, additional judgments as well (Coenen & Todd M Gureckis, 2016).

0.2 Biases from the explore-exploit tradeoff

Some sampling biases can be best thought about through the perspective of reinforcement learning. Reinforcement learning, or learning to maximize reward through decisions over time, provides a broad computational lens through which to understand the behavior of humans and other intelligent agents (Sutton & Barto, 1998). Within this perspective, one of the main challenges is the explore–exploit tradeoff (March, 1991; Cohen, McClure, & Angela, 2007; Mehlhorn et al., 2015). To maximize reward, an agent should choose the actions it believes are most rewarding based on past experience; to learn which actions are most rewarding, an agent should choose the actions that are uncertain (for a suitable definition of uncertainty). Since these two goals are often in direct conflict, determining how to balance exploration with exploitation becomes a thorny issue, and finding an optimal balance is often intractible (Guez, Silver, & Dayan, 2013).

There is a deep link between the explore–exploit tradeoff and non-random sampling. While the goal of a reinforcement-learning agent is eventually to adopt exploitative behavior to maximize reward, doing so only produces data from the actions believed to be most rewarding. By exploring, the agent sacrifices short-term rewards in order collect data from otherwise neglected actions to form a sample that better represents the full environment and facilitate learning. This connection to random sampling is made particularly clear by the mechanisms that simple reinforcement-learning algorithms use to produce exploratory behavior. In *ϵ-greedy* action selection a random action is taken instead of the exploitative action with probability ϵ, while in *softmax* action selection actions are chosen probabilistically in proportion to their relative expected values (Sutton & Barto, 1998). In both cases, exploration serves to make the sample of actions more random, and less conditional on current beliefs.

While much of the reinforcement learning literature focuses on how agents can maximize reward, a smaller body of work has emerged that examines how the interaction of reinforcement learning agents with their environment leads to specific biases and false beliefs. Though these biases share similarities with the biases from conditional sampling described in the preceding section, they are differentiated by their dynamic nature. While the overconfidence described by Einhorn and Robin M. Hogarth (1978) can develop in a static context in which the selection criterion is already set, the biases intrinsic to reinforcement learning arise through a temporally extended interaction between an agent's pattern of decisions and belief revisions and the structure of the environment.

The prototypical bias caused by reward-seeking behavior is the *hot stove effect*, described by Denrell and March (2001). The hot stove effect is the tendency for a decision maker to learn overly negative estimates of items with varying outcomes and avoid them, even if their true average value is positive. This occurs because of an asymmetry of the actions taken and information received when the agent has incorrect positive beliefs versus incorrect negative beliefs about an item. If the initial observations suggest an item is positive, the agent will continue sampling it in order to maximize reward and will find out if its early learning was wrong. But if the initial observations suggest the item is negative, the agent will begin to avoid it, and its beliefs may go uncorrected indefinitely. Thus, for example, because people have the (partial) ability to avoid other people whom they don't like, negative first impressions probably tend to last longer than positive first impressions. The hot stove effect also implies that people will tend to be risk averse, because risk-less items, with outcomes that never vary, can be learned about with a single sample and will not be subject to this negativity bias.

The hot stove effect has been used to explain risk aversion and other biases in areas ranging from social stereotypes and nepotism (Denrell, 2005; C. Liu, Eubanks, & Chater, 2015) to the foraging behavior of bees (Niv, Joel, Meilijson, & Ruppin, 2002). Unlike some biases that can be resolved through proper statistical assessment of available samples (e.g., Klaus Fiedler, Brinkmann, et al., 2000; Feiler et al., 2012; Coenen & Todd M Gureckis, 2016), the hot stove effect occurs even for optimal decision makers (Le Mens & Denrell, 2011). But while it cannot be prevented, it can be lessened through persistent exploration. In Chapters 1 and 2 of this dissertation, we'll investigate some of the factors that lead people to explore more or less. In particular, these studies focus on how people's consideration of the future affects their willingness to forego immediate exploitation in favor of the longer-term rewards of exploration.

The hot stove effect is the best-known bias caused by the explore–exploit tradeoff, but other related biases occur in environments with more complex structure (e.g., Russell H Fazio, Eiser, & Shook, 2004). Chapter 3 of this dissertation aims to provide a broader understanding of how these biases, including the hot stove effect, develop. It highlights a common pathway through which early exploration creates a false belief which the decision maker then attempts to exploit, preventing further learning and belief correction in what we term a *learning trap*. For example, a child who tastes rocky road ice cream and dislikes it might form the belief that she dislikes all ice creams with nuts and avoid them, never trying pralines and cream to discover that she truly dislikes only ice creams that mix nuts with chocolate. Because learning traps occur as a consequence of what people initially learn from their experiences, understanding them requires understanding learning mechanisms, and particularly the inductive biases that control how people generalize from one experience to another (Shepard, 1987; Mitchell, 1980).

0.3 Sampling biases and learning traps as guideposts for machine learning

The studies in this dissertation, along with most of the sampling-biases literature, are focused on human decision makers. This makes sense, because historically the decision makers collecting samples to take consequential actions have been, for the most part, human. Recently, this has begun to change, as machine learning systems take a larger role in commercial, healthcare, educational, and governmental decisions. In some ways, what we have learned about human biases makes the shift to algorithmic decision making appealing. Even very simple algorithmic rules can often perform better than expert humans, as was found decades ago by psychologists and confirmed recently by machine learning researchers (Swets et al., 2000; Jung, Concannon, Shroff, Goel, & Goldstein, 2017). But as machine learning enters into sensitive and societally important areas, there has been a growing body of evidence that automated systems do not always prevent biases, but can perpetuate and often amplify them (Campolo et al., 2017; O'Neil, 2017). There is an unmet need to connect what is known about the sampling biases of human decision-makers with the novel study of bias in machine learning systems. Doing so could provide a useful road map to machine learning and data science researchers, as well as demonstrate the severity and importance of sampling biases when they lead to biased decisions at scale.

A major source of problems in machine learning systems is that the data sets on which the systems are trained often differ from those on which they should be making decisions. This mismatch reflects R. M. Hogarth et al. (2015)'s concept of *wicked* learning environments, in which a mismatch between the samples a person learns from and the samples they make decisions on leads to biases. One way in which a training set can fail to reflect the true environment is if it contains observations of the actions of people who are themselves biased. For instance, when a hospital in the United Kingdom created an algorithm to sort medical school applicants, the algorithm was trained on data from past decisions. Because the past decisions had discriminated against women and minorities, the algorithm did as well (Barocas & Selbst, 2016). Similarly, a system may become biased if the designers exhibited non-random sampling in the training data they selected. Commercial facial recognition software tends to make more errors identifying the gender of darker-skinned than lighter-skinned individuals, possibly because most training sets over-represent lighter-skinned people and men and under-represent darker-skinned people and women (Buolamwini & Gebru, 2018). This type of issue was put into stark relief by a recent incident in which Google's image recognition system classified African-American women as gorillas (Dougherty, 2015).

In general, the data sets from which algorithms learn are likely to exclude those who are on the margins of society, even if unintentionally (Lerman, 2013). For example, Boston recently deployed a smartphone app called StreetBump to identify potholes using accelerometer data from drivers. While this could be an effective approach, it can also shift city resources towards areas where most residents own smartphones and away from older, poorer neighborhoods where they do not (Crawford, 2013). And as algorithms are applied to things like application screening and workplace analytics, they will have to deal with the same data biases, including conditional sampling and censored sampling, that people have dealt with and that are intrinsic to these domains. While there are statistical techniques to correct for such data biases (A Gelman, Carlin, Stern, & Rubin, 2014), the predictive models used for many applied machine learning problems, such as neural networks and random forests, do not naturally incorporate these techniques (Breiman, 2001).

Machine learning systems also often iteratively select their own training samples through their actions, which are based on their current "beliefs," creating the potential for the types of feedback loops that cause learning traps in people. In situations from hiring decisions to bail decisions, there is an assymetry between the information gained when the system chooses one action (hire a candidate; release a defendent) versus the other (reject a candidate; hold a defendent until the trial) (Jabbari, Joseph, Kearns, Morgenstern, & Roth, 2016; Kleinberg et al., 2017; Jung et al., 2017). The consequences of conditional feedback are beginning to be felt in fields such as predictive policing. Because predictive policing algorithms send officers to locations where crime has occurred in the past (Mohler et al., 2015), they can lead to a feedback loop in which more and more police resources are sent to a few "hot spots" while other areas appear to be crime-free because low-level infractions there are not observed (Lum & Isaac, 2016; Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2017). In essense, these algorithms undervalue exploration, repeatedly exploiting known high-crime areas to produce more tickets and arrests, with the result that certain communities, often those with poor and minority populations, are overpoliced.

Researchers have begun to propose solutions to problems of bias in machine learning. Some possible solutions are technical. For instance, recent work in *fair reinforcement learning* (Joseph, Kearns, Morgenstern, & Roth, 2016; Y. Liu, Radanovic, Dimitrakakis, Mandel, & Parkes, 2017) has sought algorithms that are guaranteed never to prefer an inferior action to a superior action or treat two equivalent actions differently, effectively making learning traps impossible. While promising, this approach also presents new challenges, and there are environments in which a fairness requirement makes learning almost impossible. Other academics have proposed more societal and regulatory solutions, acknowledging the fact that in many cases the underlying cause of biased algorithms is biased humans (Campolo et al., 2017).

Unfortunately, the literature on sampling biases in humans provides no ready-made fixes to problems in machine learning and data science. But it does provide several things of value. It provides a rich record of domains where sampling biases are likely to occur, from hiring decisions to medical diagnosis (Swets et al., 2000), that can act as warning signs when algorithms enter these areas. It provides an enumeration and classification of the kinds of sampling biases that can occur (R. M. Hogarth et al., 2015; K Fiedler, 2000), so that new biases can be understood in the context of known ones. And it provides a philosophical foundation that identifies when sampling biases can be completely resolved, and when they can only be reduced (Le Mens & Denrell, 2011). Thus, even as decision making in many domains shifts away from humans and towards machines, studies of human biases can continue to play a role in improving societally important decisions.

0.4 Overview of included studies

The body of this dissertation consists of three studies, each of which pertains to exploratory choice and its relation to learning biases. The first two concern the factors that influence how persistently people are willing to explore in order to gain accurate knowledge about the environment and avoid biases. The final study introduces the concept of learning traps and provides evidence of a novel learning trap.

0.4.1 Exploratory choice reflects the future value of information

This study examines whether people adjust their level of exploration based on the future value of information. As alluded to above, exploitative behavior tends to yield the greatest immediate reward, while exploratory behavior yields information that can be valuable for future actions. This means that the degree to which an agent explores should depend on its expectations about how many times similar actions will be available in the future. For instance, it's more worthwhile to try a new, just-opened restaurant on the first night of a trip to a foreign city, when there will be more chances to eat there if it is good, than on the last night, when one risks a poor meal with no potential longer-term payoff. However, prior work from several fields provided conflicting evidence about whether people actually adapt their exploration in this manner (Ching, Erdem, & Keane, 2013; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Wilson, Geana, White, Ludvig, & Cohen, 2014).

In a set of experiments, we showed that when people are faced with an uncertain item, they do scale the persistence of their exploration to the number of future encounters they expect with that item. We also demonstrated that when people are unsure about the exact number of future encounters, they are able to use subtler cues about the frequency of encounters in a similar manner. Interestingly, people behaved differently in conditions where they knew they would be told the possible outcomes after each encounter regardless of their action, removing the problem of non-random sampling. In these settings, people explored less over all and did not change their behavior based on cues about the future, confirming that this behavior acts as an information-collection strategy. In sum, the study provides evidence that human choice strategies serve to minimize the risk of biases like the hot stove effect, and that people will persist longer in exploration when the costs of such biases would be high.

0.4.2 Does present bias inhibit exploratory choice?

While the previous study showed that people are able to consider the future when deciding how much to explore, in real world settings future choices are spread out over time in a way rarely studied in the lab. A wealth of research has documented that people have a present bias that leads them to undervalue the future and over-weight immediate rewards, in everything from dieting to saving for retirement (Green & Myerson, 2004; Frederick, Loewenstein, & O'Donoghue, 2002; Kirby & Herrnstein, 1995). This tendency might drive people to over-exploit while neglecting the long-term rewards of exploration, possibly explaining under-exploration in a variety of settings (Levinthal & March, 1993; Gopher, Weil, & Siegel, 1989; Huys & Dayan, 2009). In this study, we attempted to test whether present bias affects exploratory choice by constructing a choice task that used immediately consumable rewards (in this case, time watching a fun video) to produce present bias (D. Navarick, 1998; Brown, Chua, & Camerer, 2009). We conducted a pre-registered experiment testing whether delaying these rewards to break the present bias would increase exploration. Interestingly, we found no evidence that present bias affected exploratory choice, suggesting that although exploratory choices unfold over time there might be separate mechanisms that guide exploratory actions and other intertemporal choices. However, we also found that our experimental stimuli, though used in prior studies, might not have effectively induced present bias, indicating that alternative approaches to this question may be needed.

0.4.3 The limits of learning: Exploration, generalization, and the development of learning traps

The final study begins with a review and synthesis of past studies that investigated phenomena that can be classified as learning traps. The learning traps identified in the past have involved generalization from past encounters with an item to future encounters (e.g., the hot stove effect, Denrell & March, 2001), or from one item to a similar item (Russell H Fazio et al., 2004). In the empirical section of our study, we investigated a novel learning trap, the *attentional* learning trap, which emerges as a side effect of selective attention during category learning (Kruschke, 1992). In essense, the learning trap can cause a decision maker to over-rely on one feature while ignoring other, equally relevant features, such as a recruiter learning to hire applicants solely based on what college they attended while ignoring prior experience or other important attributes. We used a simulated agent to unpack the learning characteristics of the agent and sampling characteristics of the environment that combine to cause the learning trap, and showed experimentally that people fall into the trap. We also studied several interventions predicted to lessen the learning trap, and found that they helped the simulated agent achieve improved performance. However, none of these interventions were able to help human participants, demonstrating the learning trap's persistence over a wide range of learning conditions.

1

EXPLORATORY CHOICE REFLECTS THE FUTURE VALUE OF INFORMATION [1]

People are often faced with decisions among uncertain alternatives (Mehlhorn et al., 2015; Sutton & Barto, 1998). To behave effectively, they must balance exploration of relatively unknown options with exploitation of those currently believed to be most rewarding.

An optimal decision maker achieves this balance by considering the future when deciding whether to explore. Rather than evaluate only the immediate reward from choosing an option, he or she also considers the degree to which learning about the option is expected to improve future choices, known as the value of information (Bellman, 1957; Howard, 1966). This means that expectations about future encounters with an option— how soon they will occur, how often they will occur, or if they will occur at all—are relevant to the immediate decision to explore something new. For example, the value of learning about a new local cafe is lower for someone who will move away in two weeks than for someone who will move in two years, and is higher for someone who drinks coffee frequently than for someone who drinks it rarely.

Work in economics and marketing has argued that people make exploratory choices in a manner well-described by forward-looking models that match or approach optimal

[1] Published as: Rich, A.S. & Gureckis, T.M. (2017). Exploratory choice Reflects the Future Value of Information. *Decision*.

behavior (Aguirregabiria & Mira, 2010; Ching et al., 2013; Erdem & Keane, 1996; Kreps & Porteus, 1978). While this class of models has been used to explain consumer choice (Erdem & Keane, 1996) and behavior ranging from medical decision making to college enrollment (Chintagunta, Goettler, & Kim, 2012; Stange, 2012), it is difficult to determine from field data whether people are truly forward-looking in their exploration (Ching et al., 2013). Experimental methods from behavioral economics may more clearly identify the signature of future-sensitive exploratory behavior, but such studies have produced mixed evidence, particularly in cases where the number of future choices is not precisely known (Meyer & Shi, 1995; Banks, Olson, & Porter, 1997; Lee, Zhang, Munro, & Steyvers, 2011; Wilson et al., 2014).

In the current study, we report on a series of large, online experiments using a set of novel approach-avoid decision-making tasks. We find that people are sensitive to multiple forms of expectations about the future and use these expectations to guide exploratory choice. In Experiment Sequence 1 we clarify the way in which the future number of encounters with a prospect affects exploration, a relationship that has been identified in past work (Meyer & Shi, 1995; Lee et al., 2011; Wilson et al., 2014). In Experiment Sequence 2 we extend our paradigm to include uncertainty about the number of future encounters, a situation common to daily life but in which past studies have failed to uncover future-sensitive exploration (Banks et al., 1997). We show that when expectations about the future are expressed as the frequency of future encounters with a prospect, people effectively make use of this relative frequency information to guide their exploration.

1.1 Approach-avoid decision making

We focus on a class of decision-making dilemmas that we term “approach-avoid” decision making, in which a person must choose whether to approach and sample an uncertain prospect, or avoid the prospect in favor of a well-known default. Many common decisions reduce to approach-avoid dilemmas. For example, a person may choose to either try a new cafe or maintain a default routine, which (depending on the individual) may mean going without coffee, making coffee at home, or going to a well-known cafe. Similarly, a consumer may choose between familiar and unfamiliar brands, and a doctor may select between standard and novel treatments. Critically, we focus on situations where this decision is not “one-off”, but where instead the decision-maker may expect to make the same or similar decisions in the future.

This kind of scenario, in which an agent makes a series of choices between one alternative with an uncertain reward distribution and one with a known reward distribution, is known in the statistics literature as a one-armed bandit problem (Berry & Fristedt, 1979). In the formal definition of the problem, an agent has a distribution of beliefs F over the reward distribution of the uncertain alternative (e.g., approaching and trying the new cafe), and we assume without loss of generality that the known alternative (e.g., avoiding the cafe and skipping coffee today) yields a mean reward of zero. The agent’s goal is to maximize summed expected reward over a sequence of choices weighed by a *discount sequence* $A = (\alpha_1, \alpha_2, \dots)$, a sequence of non-negative numbers that determine the importance of rewards received from each choice.

The discount sequence represents the agent's expectations about and valuation of future choices. Two types of discount sequences are of particular interest. In the first, $\alpha_m = 1$ for $m \leq n$ and $\alpha_m = 0$ for $m > n$. This is known as a finite horizon (Sutton & Barto, 1998), and corresponds to cases where a person makes a known number of choices, n, and cares equally about the outcomes from each. In the second, $\alpha_m = d^{m-1}$ and d is a *discount rate* that is non-negative and less than 1. This is known as an infinite horizon (Sutton & Barto, 1998) and corresponds to cases where a person is unsure of the number of future choices. The progressively decreasing weight of future choices reflects uncertainty over whether a given future choice and its resultant rewards will occur. (If the decision maker intrinsically prefers earlier rewards to later rewards (Frederick et al., 2002), this time preference can also be incorporated into the weight of future choices.)

Within both finite and infinite horizons, we can compare the *length* of two horizons. One finite horizon is longer than another when the number of future choices is higher, which occurs when n is larger. One infinite horizon is longer than another when the *expected* number of future choices is higher, which occurs when d is larger. In both cases, for a decision maker who is sensitive to future choices, the weight of future rewards relative to immediate reward will increase as the horizon lengthens.

1.1.1 Effect of horizon on optimal exploration

A general relationship holds between expectations about the future and optimal choice: as the horizon grows longer, the value of approaching relative to avoiding increases (or remains the same in the limiting case; see Proof of the non-decreasing relative value of approaching in Appendix A.). More precisely, let $V_{ap}(F, A)$ and $V_{av}(F, A)$ be the expected value of first choosing the uncertain (approach) or certain (avoid) alternative, respectively, and subsequently following an optimal strategy, and let discount sequence A^+ be longer than A. Then for any belief distribution F,

$$V_{ap}(F, A^+) - V_{av}(F, A^+) \geq V_{ap}(F, A) - V_{av}(F, A)$$

The horizon has no effect on the immediate expected reward of either action, so this change in relative values reflects solely the increasing value of approaching as an exploratory, information-seeking action. When there is only a single choice remaining or the discount rate is zero, information about the uncertain alternative cannot be used for consequential future choices and thus has no value. As the horizon lengthens, collecting information to improve future choices becomes increasingly valuable. Because gaining information is *contingent* on approaching, the relative value of approaching increases.

To illustrate this relationship and its consequences, consider the example of a finite-horizon problem where the known alternative has a constant payoff of 0, and the uncertain alternative produces payoffs of 1 and -1. Suppose the uncertain alternative is expected to produce the higher payoff on either $1/3$ or $2/3$ of trials, with these two possibilities equally likely *a priori*.

Figure 1.1 shows the behavior of an optimal agent engaging in problems of this type with horizons ranging from one to 32 choices. When there is a single choice, the agent is indifferent between approaching and avoiding because the expected value of both options is zero. When information can be used to inform at least one future choice, the agent initially approaches, and persists longer in approaching when the horizon is longer. This persistence represents a trade-off that harms the agent in some cases and helps it in others, because it occurs both when the prospect's true expected value is positive and when it is negative. But while approaching a mostly-negative prospect yields information that corrects the agent's beliefs and causes only short term costs, avoiding a mostly-positive prospect yields no belief-correcting information and inflicts long-term costs (Denrell & March, 2001). This means that as the horizon lengthens, the relative cost of avoiding a good prospect grows relative to that of approaching a bad one.

The role of the value of information in determining optimal behavior is made particularly clear by comparing the standard one-armed bandit problem, with choice-contingent information, to a problem with *full* information where feedback about the foregone payoff is provided upon avoiding. Optimal behavior in this situation is plotted as a dotted gray line in Figure 1.1. The expected immediate reward of approaching is the same as in the original formulation, but approaching no longer has additional value because information is provided regardless of choice. As a result, horizon loses its influence on choice and the optimal policy becomes a myopic reward-maximizing policy that tracks the expected immediate return from each arm.

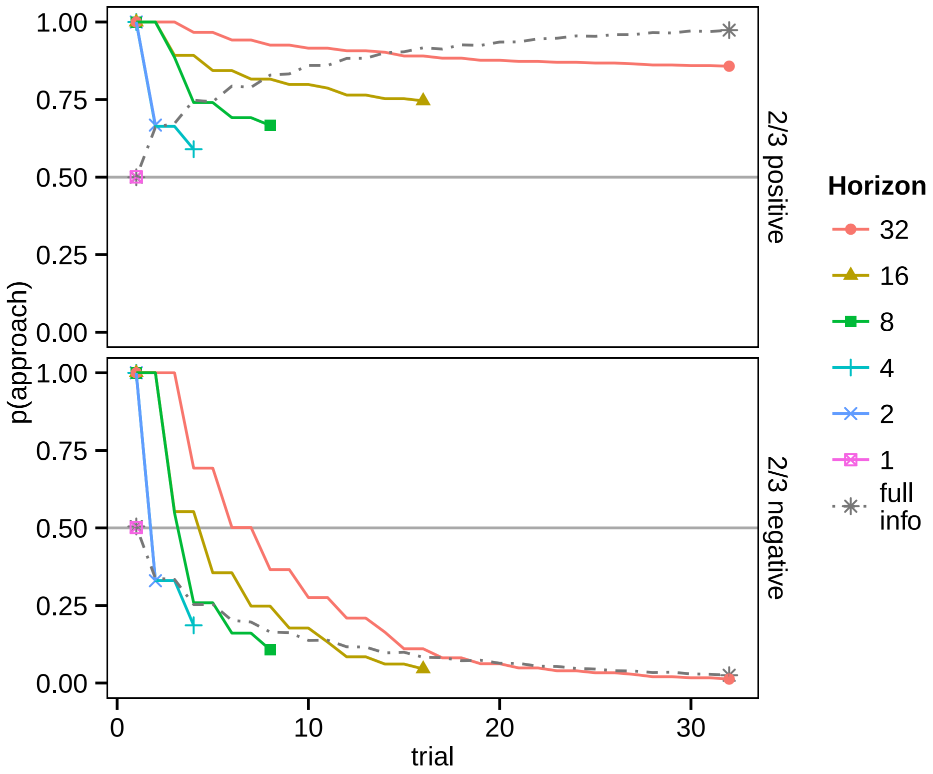


Figure 1.1: Simulated behavior of an optimal agent when encountering a prospect over a finite horizon. The top panel shows behavior when the prospect is truly 2/3 positive, and the bottom panel shows behavior when it is truly 2/3 negative. With choice-contingent information, the agent approaches on early trials to collect information, and approaches more persistently in longer horizons. With full information, the agent's behavior simply tracks the prospect's immediate expected reward given the observed outcomes, regardless of horizon. The model was simulated over 10000 task iterations.

1.2 Past Experimental Tests of Forward-looking Exploration

As the rational analysis in the previous section makes clear, the expectation of future encounters should cause a bias towards approaching an uncertain prospect, and this bias should increase as the horizon lengthens. Past experiments have yielded mixed results as to how much human exploratory choice reflects these two patterns.

First, some work has suggested that there is no bias towards choosing uncertain options, and that people are uncertainty-insensitive (Daw et al., 2006) or even uncertainty-averse (Payzan-LeNestour & Bossaerts, 2011). However, studies using more sophisticated models (Speekenbrink & Konstantinidis, 2014) or more constrained tasks (Knox, Otto, Stone, & Love, 2012) report that people do in fact tend to explore uncertain options. In addition, experiments in which participants can make a series of purely information-seeking actions before making a single consequential choice show that people are willing to sacrifice both time and money to reduce uncertainty (Hertwig, Barron, Weber, & Erev, 2004; Juni, Gureckis, & Maloney, 2016).

These studies establish that uncertainty can drive exploratory choice. A smaller number of studies have more directly examined whether exploration is forward-looking by testing the effects of horizon length, particularly with finite horizons. While these studies have generally found that exploration is uncertainty-sensitive, they vary in their reported effect of horizon. Meyer and Shi (1995) found that people chose the uncertain option more in a one-armed bandit task of 20 trials than in one of 5 trials, and Lee et al. (2011) found evidence of horizon-sensitivity among some participants in 8-trial and 16-trial four-armed bandit tasks. Most recently, Wilson et al. (2014) found that people explored more when six trials remained than on the final trial of a bandit task, but that exploration did not increase further when eleven trials remained. They proposed that people may have a heuristic to explore more when future encounters are expected than when they are not.

To our knowledge, Banks et al. (1997) conducted the only experiment examining the effect of infinite horizons with differing discount rates on exploration. To manipulate discount rate, participants in two conditions of a one-armed bandit task were informed that the probability of the task continuing after each trial was .8 or .9, respectively. While participants tended to choose the uncertain arm, there was no difference in this tendency between the two conditions. It might be that the small difference in stopping probabilities along with a low power design contributed to this reported null effect.

In summary, there is evidence that people show some form of sensitivity to the future value of information when the horizon is finite. However, there is no existing evidence that people are sensitive to the value of information in an infinite horizon setting, despite finite horizons being relatively uncommon in everyday life.

In the following experiments, we first test whether approach behavior parametrically increases with the length of a finite horizon, as should be the case for a forward-looking decision-maker. These studies are designed to confirm the existing literature and test the generality of those findings. We then provide a novel test of forward-looking exploration in an infinite horizon. This represents an important contribution to the literature, especially given the relevance of infinite horizon tasks to everyday choice behavior. In both cases we also compare behavior in the standard, contingent-information task with behavior in a non-contingent, full-information task where foregone payoffs are revealed. These control conditions verify that horizon-dependent behavior is in fact due to information seeking rather than being an idiosyncratic aspect of our experimental task. While several studies have described the effects of receiving foregone payoffs (Eldad Yechiam & Busemeyer, 2006; Grosskopf, Erev, & Yechiam, 2006), little work has documented the effect of *anticipating* foregone payoff information on exploratory choice.

1.3 Experiment Sequence 1—Finite Horizons

We conducted two experiments to investigate whether the length of a finite horizon affects approach-avoid decision-making, and whether this effect is linked to information-seeking. In Experiment 1a, we tested whether exploration increased parametrically with horizon when information is choice-contingent. In Experiment 1b, we replicated the results of Experiment 1a in a setting where participants were given more precise prior information about the environment, and also added a control condition in which participants were given full information regardless of choice, and in which no effect of horizon was predicted.

Participants completed a sequence of one-armed bandit problems in the form of a mushroom-foraging game. They visited patches that each contained a unique mushroom species and had different numbers of mushrooms available, creating different horizons. They encountered each mushroom in the patch in turn, and chose to either eat it to receive an uncertain payoff, or avoid it to receive a payoff of zero. This simple scenario mimics the formal analysis described above. Compared to past experiments, we used a relatively simple task, wider range of horizons, and greater number of participants.

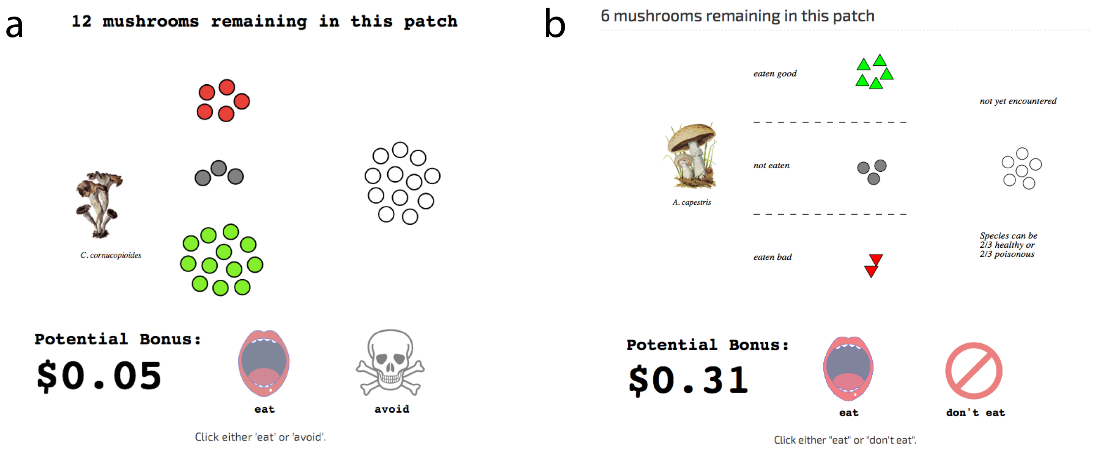
1.3.1 Method

Participants

Participants were recruited via Amazon Mechanical Turk using the psiTurk framework (Todd M Gureckis et al., 2015) and compensated with a monetary payment and performance-based bonus. Past work has shown that data collected using AMT is comparable to data collected in a lab setting (Crump, McDonnell, & Gureckis, 2013). Participants were tested on their comprehension of the experiment instructions, and data from participants who failed the comprehension test more than twice was excluded. Based on model simulations we predicted that a clear effect of horizon would emerge with 100 participants. For both experiments we conducted a preliminary qualitative analysis of the first 50 participants before completing participant recruitment. Final sample sizes were 143 (3 excluded) in Experiment 1a and 254 (22 excluded) across the two conditions of Experiment 1b. No variables or conditions were dropped from our analysis.

Design and procedure. Participants in both experiments played a game based around foraging for edible mushrooms. Mushrooms were represented by color illustrations. Figure 1.2 shows examples of the task in each experiment.

Experiment 1a. Participants sequentially encountered patches of mushrooms that contained 1, 2, 4, 8, 16, or 32 exemplars of a single species, and were informed that each species was unique to a single patch. The goal of the task was to eat healthy mush-



12 mushrooms remaining in this patch

Click either 'eat' or 'avoid'.

6 mushrooms remaining in this patch

Species can be 2/3 healthy or 2/3 poisonous

Click either "eat" or "don't eat".

Figure 1.2: Example of the Experiment 1a task (a), and of the Experiment 1b task (b) in the contingent-information condition.

rooms while avoiding poisonous ones. Participants were told that species varied in their proportion of healthy mushrooms, from “almost-always healthy” to “almost-always poisonous”. They encountered four patches of each length; one each of proportions $p(healthy) = \{.125, .375, .625, .875\}$. The patches were pseudo-randomly ordered.

In each patch participants first observed the set of available mushrooms represented as a group of empty circles. On each trial, participants chose whether to eat or avoid the next mushroom in the patch by clicking buttons labeled “eat” and “avoid”. Upon eating a mushroom, it turned green (if healthy) or red (if poisonous) and moved to a group of healthy or poisonous mushrooms. Upon avoiding a mushroom, it turned gray and moved to group of avoided mushrooms. The number of remaining mushrooms in the patch was denoted at the top of the screen at all times.

Participants started with a bonus of $.25. They earned $.02 for each healthy mushroom eaten and lost $.02 for each poisonous mushroom eaten. They did not gain or lose money for avoiding a mushroom. The bonus was cumulative over all patches, and its value was visible throughout the game.

Experiment 1b. As in Experiment 1a, participants encountered patches of mushrooms that contained 1, 2, 4, 8, 16, or 32 exemplars. Participants were given more precise information that each species was either of the "mostly-healthy" type, for which 2/3 of individuals were healthy, or the "mostly-poisonous" type, for which 2/3 of individuals were poisonous, and that these two types were equally common. These environmental statistics match those for which the optimal policy is shown in Figure 1.1. Participants encountered four patches of each length, with two of each type.

Participants were split into two conditions. In the contingent-information condition, the button to avoid (labeled "don't eat") was described simply as not eating the mushroom. In the full-information condition, participants were told that when they chose not to eat the mushroom, they would put it into a "mushroom-testing kit" and learn its value.

Eaten mushrooms were represented as upward-pointing green triangles if healthy and downward-pointing red triangles if poisonous. Not-eaten mushroom were represented as gray circles in the contingent-information condition. In the full-information condition, they were represented as grayish-green upward-pointing or grayish-red downward-pointing triangles, depending on their healthiness.

As in Experiment 1a, participants earned a cumulative bonus that started at $.25 and changed in increments of $.02 as they ate healthy or poisonous mushrooms.

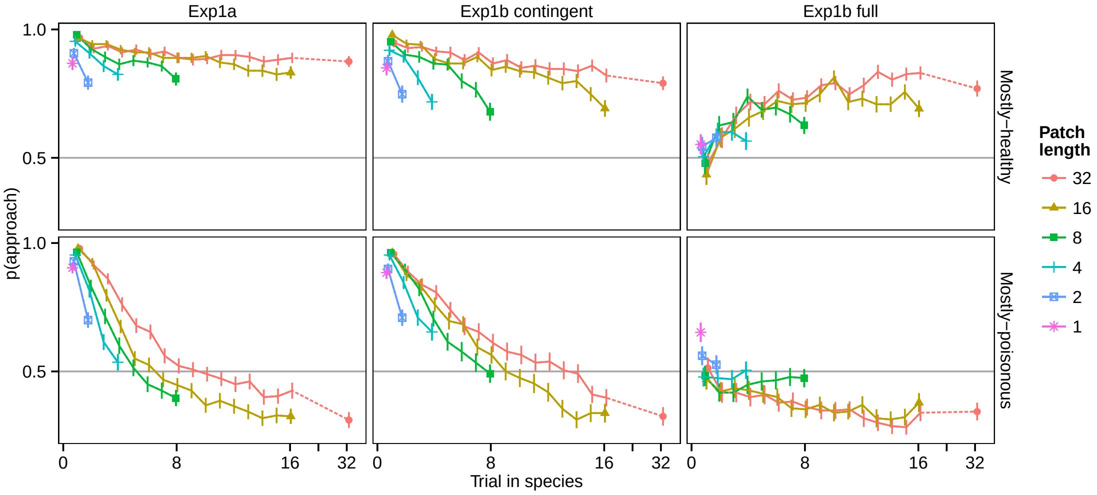


Figure 1.3: Probability of approaching a mushroom on each encounter within a patch for each finite-horizon experiment and condition. Top panels show participant behavior for patches where p(healthy) was .625 or .875 (Exp. 1a), or .67 (Exp. 1b), and bottom panels show behavior for patches where p(healthy) was .125 or .375 (Exp. 1a), or .33 (Exp. 1b). Error bars show standard error of the mean. In order to focus on comparisons among patch lengths, the x axis is compressed after trial 16 and the change in p(approach) from trial 16 to trial 32 is shown as a dotted line. Participants tended to approach early on and to have higher p(approach) for longer horizons in Experiment 1a and in the contingent-information condition of Experiment 1b, but not in the full-information condition of Experiment 1b.

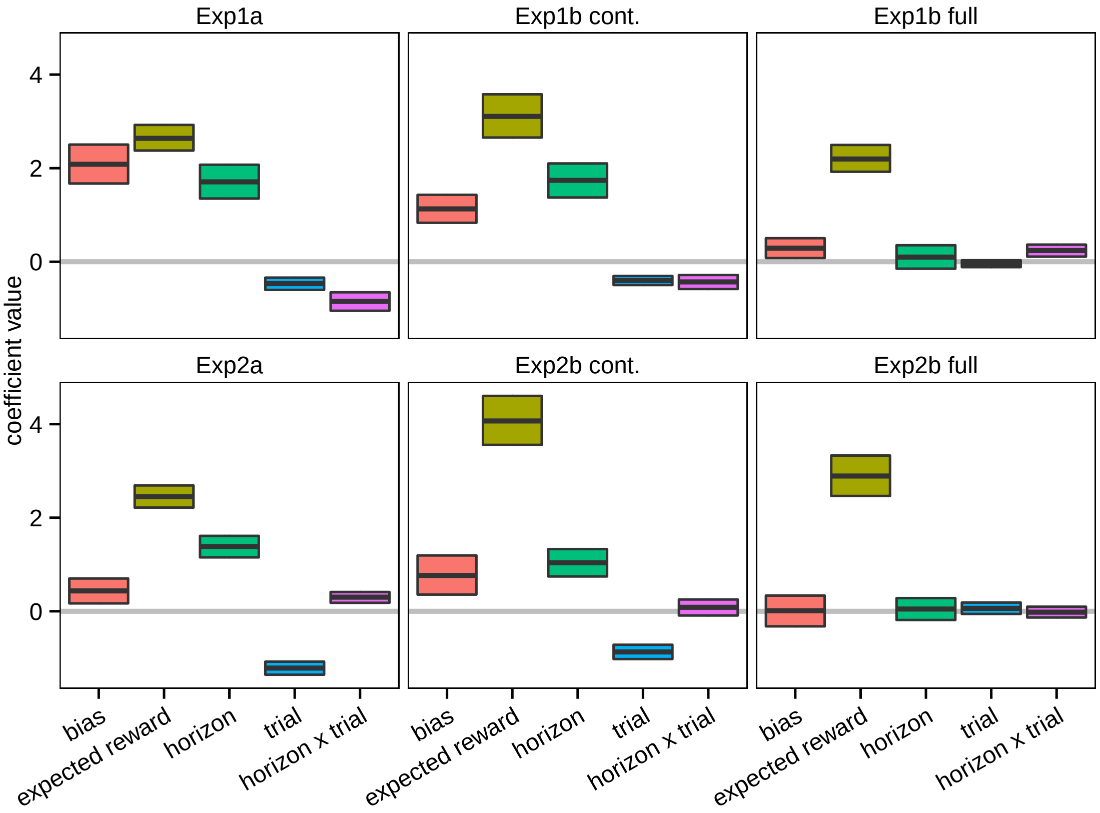


Figure 1.4: Posterior estimates of the population-level effects of five predictors on behavior, estimated using a hierarchical Bayesian logistic regression of participants' choices in all experiments. The top panels show results for finite-horizon experiments, and the bottom panels show results for infinite-horizon experiments. Thick lines show posterior means, and colored intervals show 95% posterior intervals. As predicted by a forward-looking model, participants exhibited a positive effect of horizon in all four contingent-information scenarios, but not in the two full-information scenarios.

1.3.2 Results

Participants' probability of approaching a mushroom on each trial within patches of each length is shown in Figure 1.3. In both Experiment 1a and the contingent-information condition of Experiment 1b, participants' behavior resembles that of the forward-looking, information-sensitive model (see Figure 1.1), with a high rate of exploration in early trials and more persistent exploration in larger patches. In the full-information condition of Experiment 1b, participants appeared to begin approaching at a rate near chance and then to modify their behavior based on observed outcomes, similar to the myopic reward-maximizing policy shown in Figure 1.1.

To quantify the effects of horizon and other variables on trial-by-trial behavior we used a hierarchical Bayesian logistic regression that allowed for individual differences among participants (see Description of data analysis in Appendix A). We included five predictors: a bias term (capturing overall tendencies to approach or avoid), immediate expected reward (i.e., expected payoff from approaching on the next trial), number of remaining trials in the patch (horizon length), trial number within the patch (which may have an independent effect if participants were uncertainty- or novelty-seeking), and the interaction between trial number and horizon.

We calculated expected reward by applying Bayes' rule using the participant's observed outcomes for a patch, with a uniform prior over p(healthy) in Experiment 1a and the prior given in the instructions in Experiment 1b. Horizon and trial-number were log transformed because the marginal effects of both factors are expected to be decreasing. We rescaled expected reward to equal -1 when p(healthy)= $1/3$ and 1 when p(healthy)= $2/3$, and rescaled log horizon to range from 0 (at horizon 1) to 1 (at horizon 32). Finally, we shifted log trial-number to have zero mean so that the horizon coefficient could be interpreted as an average effect across all trials.

The model posterior was estimated for each experiment and condition using the Stan modeling language (Stan Development Team, 2015). Posterior estimates of the population-level parameters are presented in Figure 1.4. Simulated data from the model posteriors confirmed a close match to the key features of the data (see Figure A.1 in Appendix A).

Participants were positively sensitive to expected reward across all three scenarios. The model posteriors confirm that behavior was similar in Experiment 1a and the contingent-information condition of Experiment 1b. In both scenarios, participants were highly exploratory early on; the 95% posterior predictive intervals for first-trial approach proportion were [.920, .937] and [.894, .916], respectively. Participants became less likely to approach over the course of a patch, as shown by a negative effect of trial number. This decrease in exploration likely results from the decreased uncertainty about and novelty of later mushrooms in a patch.

Critically, as shown in Figure 1.4, there was a positive population-level effect of horizon in both contingent-information scenarios, suggesting that people were not simply uncertainty-seeking but used a forward-looking strategy that tracked the value of information. This sensitivity to horizon also held broadly at the individual level. In Experiment 1a, the posterior mean effect of horizon was above zero for 81% of participants, and the 95% posterior interval for the effect of horizon was entirely above zero for 51% of participants. In the contingent-information condition of Experiment 1b, the posterior mean was above zero for 79% of participants, and the posterior interval was entirely above zero for 58% of participants. There was also a negative interaction between horizon and trial, such that the effect of horizon decreased in later trials. This may reflect that further into a patch participants were confident about the healthiness of the species, reducing the potential value of information.

Behavior in the full-information condition of Experiment 1b was markedly different from behavior in the contingent-information scenarios. Participants were roughly indifferent between approaching and avoiding on the first trial, with a 95% posterior predictive interval of [.500, .538]. There was no main effect of trial number or of horizon, and the 95% posterior interval for horizon was completely below those of the two contingent-information scenarios. We also observed that participants started approaching less in the last few trials of mostly-healthy patches and more in the last few trials of mostly-poisonous patches. This effect is not well-captured by our model and is contrary to our predictions, but is distinct from the information-related horizon effect in that it tends towards increased approach when the horizon is short and expected reward is negative. We hypothesize that this effect may be an instance of the gambler's fallacy, with participants expecting that a string of good outcomes would be balanced out by bad ones and vice versa (Tversky & Kahneman, 1971).

Recent work raises the question of whether the effect of horizon is parametric and smoothly increasing, or categorical with two distinct levels for situations with and without future encounters (Wilson et al., 2014). While our main analysis assumed that the effect of horizon was logarithmic, we tested this assumption with an expanded model in which each experienced horizon, from 1 to 32, had a unique population-level parameter that was allowed to vary freely (see Description of data analysis in Appendix A). We found that in both contingent-information scenarios these values increased at a roughly logarithmic rate, supporting our supposition that the effect of horizon increases parametrically but with a decreasing marginal effect (see Figure A.2 in Appendix A).

Finally, while participants did modulate their exploration based on the the length of the horizon, they also appeared to have a general bias towards information seeking. We tested this by comparing the final trials of the two conditions of Experiment 1b, where the true value of information was zero but participants could still choose to make an information-seeking choice in the contingent-information condition. A regression on these final trials reveals that, controlling for expected reward, participants in the contingent-information condition were more likely to approach, $z = 7.82, p < .001$. This bias could reflect a heuristic to collect information even when not clearly useful, for example in case the species was encountered again (although participants were informed it would not be) or the information could improve prior knowledge about future species (although the prior over $p(healthy)$ was given).

1.4 Infinite horizons and prospect frequency

In Experiments 1a and 1b, as in most previous studies of task horizon and exploration (Lee et al., 2011; Meyer & Shi, 1995; Wilson et al., 2014), participants knew the exact number of times they would encounter a prospect in the future. This kind of finite horizon, though, is not representative of many of the decisions faced in everyday life. Often, people do not know how many times they will encounter a prospect. For example, there is no precise limit on the number of times someone might have the opportunity to visit a local cafe. This type of situation is more naturally formalized as an infinite-horizon problem.

Banks et al. (1997) attempted to induce infinite horizons with different discount rates in a one-armed bandit task by informing some participants that the task had a .9 probability of continuing after each trial, and others that the task had a .8 probability of continuing. They found that this manipulation did not affect participants' exploration, though this null result may have been due to low power. In addition, while a probabilistic experiment length is an effective way to create uncertainty about the horizon (Camerer & Weigelt, 1996), explicit information about ending probabilities may feel artificial to participants and be difficult to integrate into decision-making strategies.

A more natural and common way in which an individuals face differing effective discount rates is through the differing frequencies of prospects within a wider environment. Some types of products are purchased more often than others, and likewise some social situations are encountered more than others. Intuitively, if the length of the environment is uncertain (e.g., length of time living in a city, length of life), and encounters with one prospect are less frequent than encounters with another, then a person should expect more future encounters with the frequent prospect and value information about that prospect more highly.

We can formalize this idea by considering an agent facing a *contextual* one-armed bandit problem. The agent has a single infinite-horizon discount sequence A with discount rate d, but rather than always facing the same uncertain alternative, on each trial it encounters one of k independent uncertain alternatives (i.e., the context for that trial).

Each alternative has its own starting belief distribution F_i, and each is encountered with a known frequency f_i, such that $\sum_{i=1}^{k} f_k = 1$. For example, the agent may have to decide whether to buy an espresso each morning after observing which of k baristas (with potentially varying skill) are working at the cafe. We assume the agent knows how frequently each barista works at the cafe, though this could also be learned from experience.

Since the alternatives are independent, this situation can be reduced to k independent one-armed bandit problems by considering only the sequence of encounters with the k^{th} alternative. However, the timing, and thus the discounted weight, of future encounters with each bandit will depend on its frequency, causing future encounters with rare prospects to tend to receive lower weight. While the exact discount sequence for a given prospect is unknown in advance, these uncertain sequences can be replaced with their expected values (Berry & Fristedt, 1985), resulting in a set of k independent one-armed bandits with effective discount rates d_k.

Figure 1.5 shows the behavior of an optimal agent in an environment with an infinite horizon, for prospects that have differing frequencies. To explore optimally, the agent approaches more when the prospect is frequent, just as it does when the finite horizon is longer (see Figure 1.1).

1.5 Experiment Sequence 2—Infinite Horizons

In Experiment Sequence 2, we tested whether participants display a sensitivity to prospect frequency. While there is a large literature surrounding the effects of rare and common

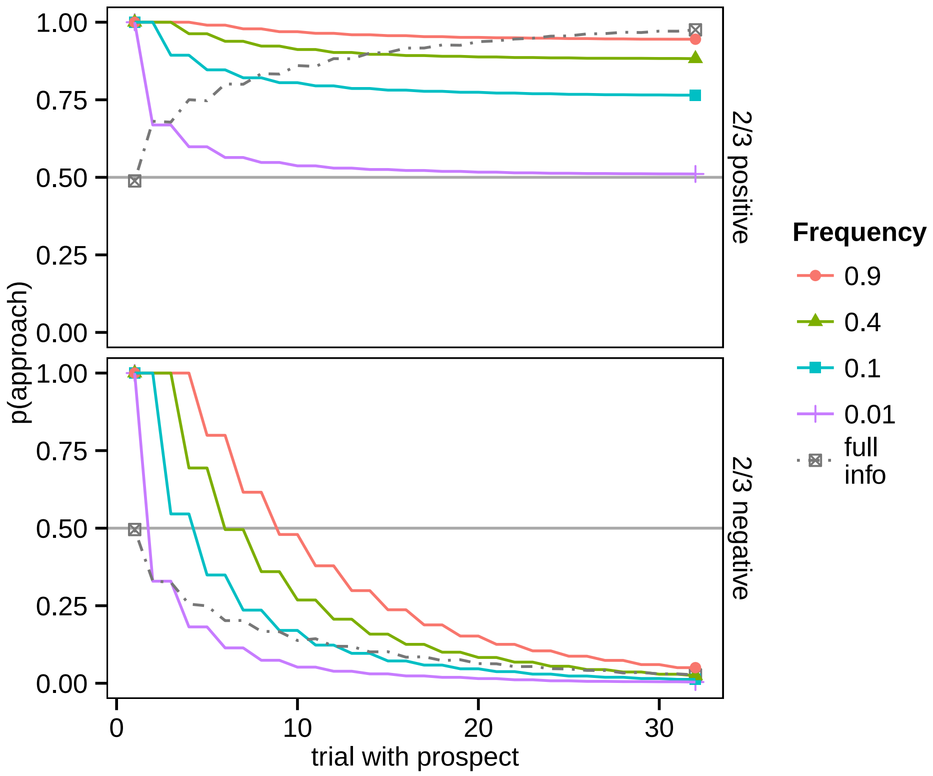


Figure 1.5: Simulated behavior of an optimal agent with a base discount rate of .99 when encountering prospects with differing frequencies of occurrence over an infinite horizon. Each prospect is known to either yield a payoff of +1 with probability 2/3 or a payoff of -1 with probability 2/3, and the opposite payoff otherwise. The top panel shows behavior when the prospect is truly 2/3 positive, and the bottom panel shows behavior when it is truly 2/3 negative. With choice-contingent information, the agent approaches to collect information, and approaches more persistently when the prospect is more frequent. With full information, the agent's behavior simply tracks the prospect's expected value given the observed outcomes, regardless of frequency. The model was simulated over 10000 task iterations.

outcomes on decision strategies (e.g., Hertwig et al., 2004), to our knowledge there are few studies that examine the effect of rare and common prospects—that is, of rarely and commonly faced decision contexts. In Experiment 2a, participants performed a contingent-information task in which the horizon was unknown. In Experiment 2b, we replicated the results of Experiment 2a in a similar task with more precise information about the possible values of prospects and the distribution of possible horizons, and also added a full-information control condition.

Participants played a mushroom-foraging game that instantiated the kind of contextual one-armed bandit problem described above. Participants played through a series of *habitats* that each had four unique mushroom species, any one of which could appear on a given trial. They were trained on the relative frequency of the species at the beginning of the task, and then made a series of approach-avoid decisions. In Experiment 2a, participants were not told the length of the habitats. In Experiment 2b, we incorporated a probabilistic habitat-ending mechanism so that participants had explicit prior information about the uncertainty over the horizon.

1.5.1 Method

Participants

Participants were recruited via Amazon Mechanical Turk using the psiTurk framework and compensated with a monetary payment and performance-based bonus. Data from participants who failed the pre-experiment comprehension test more than twice was excluded. For both experiments we conducted a qualitative analysis of the first 50 par-

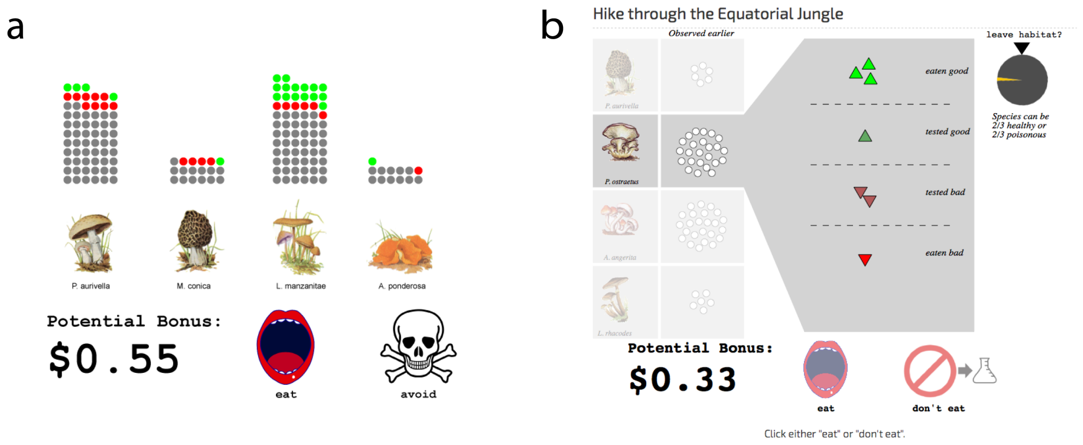


Figure 1.6: Example of the Experiment 2a task (a), and of the Experiment 2b task (b) in the full-information condition.

ticipants before completing data collection. Sample sizes were 152 (3 excluded) in Experiment 2a, and 159 (5 excluded) across the two conditions of Experiment 2b. No variables or conditions were dropped from our analysis.

Design and procedure

Participants in both experiments played a game based around foraging for healthy mushrooms. Mushrooms were represented by color illustrations. The experiments were divided into sub-tasks called "habitats" (e.g., "New England Forest", "Amazonian Rainforest"). Each habitat contained four unique mushroom species, two of which occurred with frequency 4/10 and two of which occurred with frequency 1/10. Within each habitat, the game was broken into two phases. In the first phase, participants observed a large, representative sample of the mushrooms in the habitat. Mushrooms encountered in this sample were depicted as circles that appeared without participant input. Once the entire sample had been shown, the species were highlighted one at a time and participants submitted a “field report” by answering questions of the form “If you saw 10 mushrooms on your hike back through the [Habitat Name], how many would you expect to be from the species [Species Name]?” This ensured that participants noticed and encoded the relative frequency of each species. In the second phase, participants played a game similar to those in Experiment Sequence 1, but where the species encountered on any trial was randomly interleaved with other species based on the underlying frequencies. The net result was that the number of trials between successive encounters tended to be greater for infrequent species than for frequent species. Figure 1.6 shows examples of the decision-making phase of each experiment.

Experiment 2a. Participants completed two habitats. Each habitat had one rare and one common species that were healthy with proportion .7 and one rare and one common species that were healthy with proportion .3. Participants were not informed of the exact possible $p(healthy)$ values. In each trial of the decision-making phase, one of the four species was highlighted (with frequencies matching those learned in the observation phase) and participants chose whether to eat or avoid a mushroom from that species. Healthy and poisonous eaten mushrooms were represented as green and red dots on a “histogram” of observations, while avoided mushrooms were represented as gray dots along with the samples from the observation phase. The decision-making phase of each habitat lasted 120 trials, but participants were not informed of when the habitat would end. Participants started with a bonus of $.50, and gained and lost money in increments of $.05. Potential bonuses were earned separately for the two habitats, and one of the two bonuses was randomly awarded at the conclusion of the experiment.

Experiment 2b. Participants completed four habitats. The task parameters and game interface were designed to be similar to that in Experiment 1b. Participants were informed that mushroom species could be either $2/3$ healthy or $2/3$ poisonous, and that these types were equally common. Mostly-healthy and mostly-poisonous species were pseudo-randomly distributed across habitats, but each habitat contained at least one healthy species to maintain engagement (this was not revealed to participants in the instructions). The presentation of mushrooms was similar to Experiment 1b, and mushrooms from the observation phase were displayed separately from those from the decision-making phase. Participants were randomly assigned to either a contingent-information or a full-information condition. Upon making a decision a new circle or triangle appeared and was added to the appropriate group for that species as explained for Experiment 1b.

Rather than giving participants no information about habitat length, as in Experiment 2a, participants were informed that after each trial of the decision-making task there was a constant small probability of leaving the habitat and continuing to the next one. This probability was illustrated by a spinner with a small yellow wedge covering 1.5% of its area. The spinner was spun after each trial, and the habitat ended when the black arrow landed on the wedge. The habitat lengths were in fact predetermined to be 40, 60, 80, and 100 trials, randomly ordered. The bonus started at \$.25 and changed in increments of \$.02, and was cumulative over the four habitats.

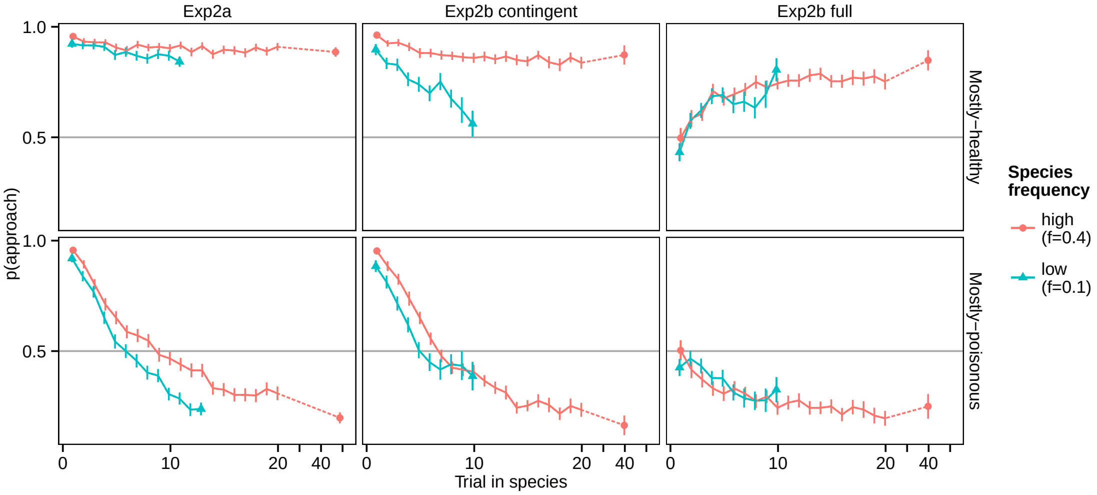


Figure 1.7: Probability of approaching a mushroom on each encounter within a patch for each infinite-horizon experiment and condition. Top panels show participant behavior for species where p(healthy) was .7 (Exp. 2a) or .67 (Exp. 2b), and bottom panels show behavior for patches where p(healthy) was .3 (Exp. 2a) or .33 (Exp. 2b). Error bars show standard error of the mean; error bars are wider on later trials in Experiment 2b because some habitats ended earlier than others. In order to focus on comparisons between patch frequencies, the x axis is compressed after trial 20 and the change in p(approach) from trial 20 to the final trial with a species is shown as a dotted line. Participants tended to approach early on and to have higher p(approach) for longer horizons in Experiment 2a and in the contingent-information condition of Experiment 2b, but not in the full-information condition of Experiment 2b.

1.5.2 Results

When asked in the "field report" how many times out of ten each species would be encountered in the future, participants in Experiment 2a reported 4.7 for frequent species and 1.6 for rare species, while participants in Experiment 2b reported 4.8 for frequent species and 1.6 for rare species, on average. The target values based on the empirical frequencies were 4.0 and 1.0. Furthermore, participants gave the exact correct response 62% of the time, suggesting frequency information was encoded accurately, and critically that the difference between frequent and rare prospects was encoded.

Results of the decision-making phase are shown in Figure 1.7. Visual inspection shows a high initial exploration rate and greater approach probability for high-frequency species in Experiment 2a and the contingent-information condition of Experiment 2b, similar to the choices of the optimal policy (Figure 1.5). In the full-information condition of Experiment 2b, people appeared to behave in a myopic reward-maximizing manner.

We analyzed the choice data using the hierarchical Bayesian logistic regression model introduced in Experiment Sequence 1, and predictors were calculated and rescaled as described above. Horizon was coded as 1 for high-frequency species and 0 for low-frequency species, and trial number represented the encounter number within the species encountered on that trial (rather than the trial number within the habitat). The posterior estimates of the population-level coefficients are presented in Figure 1.4, and posterior simulations from the model again confirmed a good qualitative match to the data (see Figure A.3 in Appendix A).

Participants were positively sensitive to expected reward across all three scenarios. In the two contingent-information scenarios, participants were highly exploratory early on, with 95% posterior predictive intervals for first-trial approach proportion of [.935, .960] and [.905, .936], and became less exploratory in later encounters.

Participants were also more exploratory in encounters with high-frequency species, supporting the hypothesis that they engaged in a forward-looking evaluation of the value of information. This effect of frequency was exhibited by a large portion of the participant population. In Experiment 2a, the posterior mean of the horizon parameter was greater than zero for 96% of participants, and the 95% posterior interval was entirely above zero for 64% of participants. In the contingent-information condition of Experiment 2b, the posterior mean was above zero for 86% of participants and the posterior interval was entirely above zero for 47% of participants. Participants did not exhibit the negative interaction between horizon and trial observed in Experiment Sequence 1, and in fact there was a small positive interaction in Experiment 2a. Further work is needed to investigate how task horizon interacts with past encounters and uncertainty.

In the full-information scenario, expected reward appeared to be the primary driver of behavior. The 95% posterior predictive interval for first-trial approach was [.450, .509], and there was no effect of horizon or encounter number. Furthermore, the 95% posterior interval for the effect of horizon was fully below those for the two contingent-information scenarios.

These findings appear to show a clear effect of frequency on approach behavior when information is choice-contingent. However, one alternative possibility is that participants simply were exploratory in the early trials of each habitat (rather than the early trials of each species), and became less exploratory later on, regardless of what species was encountered on a given trial. Since the n^{th} encounter with a frequent species will tend to come earlier in the habitat than the n^{th} encounter with an infrequent species, this could cause participants to falsely appear frequency-sensitive. We tested this explanation by fitting a logistic regression model to the first encounter with each mushroom species, using species frequency and the trial in the habitat on which the species was encountered to predict choice. There was a large effect of frequency on p(approach) in both contingent-information scenarios, z=2.58,p=.01 and z=3.96,p<.001, but no effect of trial in habitat, z=.71,p>.250 and z=0.06,p>.250. Therefore, trial in habitat appears insufficient to explain our full pattern of results.

1.6 Discussion

Exploratory decisions are a constant feature of life in an uncertain and changing world, and people's choices about what to explore often determine their later behavior, preferences, and beliefs. While work in applied fields has posited that people's decisions reflect the future value of exploring various prospects (Ching et al., 2013), experimental work has left unclear whether this is actually the case, particularly in naturalistic environments with an uncertain or "infinite" horizon. Across four experiments, we found that people are indeed sensitive to the future value of information when making exploratory decisions. Participants increased their exploration as finite task horizon increased, and were sensitive not just to definite, discrete horizons but also to differences in prospect frequency when the horizon was uncertain.

These results show that people can use exploratory strategies that consider the future, but the mechanisms underlying their behavior remain unclear. Optimal information seeking requires complex dynamic programming calculations over possible future outcomes that seem implausible for humans to compute. People may perform simpler computations, such simulating a few possible chains of future outcomes (Vul, Goodman, Griffiths, & Tenenbaum, 2014), or considering possible outcomes only one or two choices ahead (Zhang & Yu, 2013). Alternatively, they may use simple heuristics (e.g., Seale & Rapoport, 2000) or exploration bonuses (e.g., Daw et al., 2006) that are sensitive to the horizon. Finally, people may mix more and less complex exploration strategies, and the degree of forward-looking thinking may vary across individuals and situations as has been found in other decisions that involve future rewards (Frederick et al., 2002). While distinguishing these potential mechanisms of forward-looking exploration will be difficult, many interesting avenues of research towards this goal are open. These include studying exploratory choice under cognitive load, investigating the developmental trajectory of forward-looking exploration, and testing the degree to which the exploration-enhancing effect of a long horizon is altered by the context of recently experienced horizons.

Regardless of the mechanisms driving their choices, our finding that people use frequency as a cue to the value of information may offer a better understanding of exploratory behavior in many domains. In our experiments, people were more likely to eat uncertain mushrooms from a species they expected to encounter frequently. If this pattern generalizes widely, then consumers may be more likely to purchase a new brand when shopping for commonly-purchases goods (Ching et al., 2013), and doctors may be more likely to prescribe a new drug for commonly-treated diseases (Chintagunta et al., 2012). Generally, frequency-dependent exploration should allow individuals to make better exploratory decisions compared to non-forward-looking agents (Denrell & March, 2001).

Interestingly, however, this individually rational behavior might amplify societal biases. Research on fads and social influence has described how self-reinforcing cycles of popularity can develop in consumer and cultural settings (Bikhchandani, Hirshleifer, & Welch, 1992; Denrell & Le Mens, 2007). Frequency-dependent exploration may strengthen these cycles, as items that are popular and thus common become more valuable to learn about than those that are unpopular and rare.

Similarly, work on social attitude formation has suggested that one cause of negative attitudes towards outgroups is that outgroup members are more easily avoided than ingroup members, allowing false beliefs about them to persist (Gordon W Allport, 1979; Denrell, 2005). To the extent that outgroup members are rare in daily life, we predict this tendency to be exacerbated by frequency-dependent exploration. Fortunately, interventions that increase contact with an outgroup may erode prejudice (Shook & Russell H. Fazio, 2008), possibly in part by increasing the future rewards of interacting and learning. To speculate, it is possible that even an intervention that simply increased a person's belief that outgroup members are a frequent part of their social environment might increase later exploratory interactions. Thus, while forward-looking exploration may cause biases, it might also be leveraged as a tool to reverse them.

2

DOES PRESENT BIAS INHIBIT EXPLORATORY CHOICE?

Decision makers acting in uncertain environments frequently face the dilemma between choosing options known to be rewarding (i.e., exploitation) and choosing options that are unknown or uncertain (i.e., exploration). A child buying ice cream, for example, must select between getting a cone of her favorite flavor and trying something new that might become a favorite but could also be disappointing. Researchers in reinforcement learning have created a large body of knowledge about how people and animals balance this "explore–exploit dilemma" (Mehlhorn et al., 2015) as well as how the problem should be approached computationally (Sutton & Barto, 1998).

A key aspect of exploratory decision making is that it is spread over time. With only a single decision, exploration makes little sense. If a decision making knew with certainty they only had one remaining chance to buy ice cream in their life, they should buy their favorite flavor, because that is the flavor they expect to enjoy the most. Exploring new flavors has the possibility of introducing you to a new favorite, but it is only when there will be many more chances to choose that new flavor in the future that the risk of a disappointment starts to look worthwhile.

While research on exploratory choice acknowledges that the value of exploration depends on its payoffs in the future (Wilson et al., 2014; Rich & Todd M. Gureckis, 2017), it has not addressed how this might interact with the manner in which people value future rewards. When considering rewards spread over time, people tend to be present biased, overweighting immediate rewards in comparison to delayed rewards (Myerson & Green, 1995; Frederick et al., 2002). In scenarios with explore–exploit tradeoffs, this could lead to over-exploitation and under-exploration. While lab experiments tend to be conducted in short sessions with non-consumable rewards, preventing present bias from being a major factor, this preference for immediate reward could be a major factor leading to under-exploration in more temporally extended, real world settings.

In this paper, we discuss the potential connection between exploratory choice and intertemporal choice. We report on a set of experiments using directly consumable rewards in an exploratory choice task to test for effects of present bias on exploration. While we did not find an effect of present bias on exploratory choice, a follow-up experiment revealed that our consumable rewards did not in fact produce a reliable present bias, despite evidence that they did so in earier studies (Solnick, Kannenberg, Eckerman, & Waller, 1980; D. Navarick, 1998). Nonetheless, we hope that this work can serve as a useful first step towards unifying our understanding of exploratory and intertemporal decision making.

2.1 Exploration inside and outside the lab

Many researchers have examined patterns of exploration, both in naturalistic settings and in the lab. Interestingly, differing findings have emerged as to the nature and severity of biases in exploratory choice.

2.1.1 Exploration outside the lab

Exploration has been studied outside the lab in a wide range of contexts. While these domains vary greatly in their superficial characteristics, a bias towards under-exploration has often been observed.

Learned helplessness, a phenomenon applicable to many behaviors and domains, has been described as an example of insufficient exploration. In learned helplessness, an organism experiences the absence of control over the environment, learns that the environment is uncontrollable, and thus ceases to take actions that might allow it to discover that it can in fact exert control. While the initial discovery of learned helplessness occurred in the lab (Maier & Seligman, 1976), it has since been proposed to underly forms of depression (L Y Abramson, Seligman, & Teasdale, 1978; Lyn Y Abramson, Metalsky, & Alloy, 1989) as well as problems ranging from difficulties in school (Diener & Dweck, 1978) to poverty (Evans, Gonnella, Marcynyszyn, Gentile, & Salpekar, 2005). While the cognitive appraisal of experienced events affects the development of learned helplessness (L Y Abramson et al., 1978), patterns of exploration clearly play a role as well (Huys & Dayan, 2009; Teodorescu & Erev, 2014a). In the case of depression, interventions aimed at increasing the exploration of activities that might be rewarding have been found to be as effective as those with a more cognitive orientation (Jacobson et al., 1996).

Under-exploration also seems to occur in the development of complex skills, such as flying a plane or playing a sport (Gopher et al., 1989). In these settings, an "emphasis change" training method that encourages people to continually explore the performance space leads to greater performance gains than unguided practice or more complex training methods. Without this intervention, people often enter a “local maximum” in which exploration decreases and performance plateaus (E Yechiam, Erev, & Gopher, 2001).

In many other areas under-exploration is less clearly established, but is suspected to play a role in maladaptive behavior. A recent analysis of supermarket shopping behavior showed that people engage in long runs of exploitative behavior that were inconsistent with optimal behavior (Riefer, Prior, Blair, Pavey, & Love, 2017). Insufficient exploratory interaction with outgroups may be one cause of stereotypes and prejudice (Denrell, 2005), and interventions that increase inter-group contact reduce stereotypes (Shook & Russell H. Fazio, 2008). The crowding out of exploration by exploitation is a concern in organizational behavior as well (March, 1991; Levinthal & March, 1993), prompting research into organizational structures that may preserve exploration (Fang, Lee, & Schilling, 2010).

2.1.2 Exploration inside the lab

Lab studies of exploratory choice have allowed researchers to fully control the reward structure of the environment and precisely measure behavior, as well as compare behavior to optimal choice and other formal models. These studies have yielded a number of insights into the factors leading to more or less exploration, including aspiration levels (Wulff, Hills, & Hertwig, 2015), uncertainty (Speekenbrink & Konstantinidis, 2015), and the future value of information (Rich & Todd M. Gureckis, 2017; Wilson et al., 2014)

Interestingly, under-exploration has not emerged as a clear pattern in lab experiments. Instead, results are mixed with people sometimes under-exploring, sometimes over-exploring, and sometimes exploring close to an optimal amount. To take two illustrative examples, (Zwick, Rapoport, Lo, & Muthukrishnan, 2003) found that in a sequential search task people under-searched when there were no information costs but over-searched when there were information costs, and (Teodorescu & Erev, 2014b) found that people explored unknown alternatives too often or not often enough depending on whether rare outcomes were positive or negative. Similar results have been obtained within and across a variety of other studies and paradigms (Tversky & Edwards, 1966; Hertwig et al., 2004; Navarro, Newell, & Schulze, 2016; Juni et al., 2016; Sang, Todd, & Goldstone, 2011).

These experimental studies raise the question of why under-exploration appears more widespread outside the lab, but not in the lab. One possibility is that both forms of deviation from optimality are in fact prevalent, though perhaps in different settings, and that the seemingly general bias toward under-exploration is illusory. An alternative is that there are some important aspects of real-world decisions—or peoples' cognitive and motivational states when making those decisions—that makes differentiate them from decisions in the lab. One clear possibility is that in real world exploration, choices and outcomes are spread out over time in a manner that is rarely found in the lab, and that people's bias towards immediate rewards might therefore account for a portion of people's tendency to under-explore.

2.2 Temporal discounting

Temporal discounting refers to the underweighting of temporally distant rewards relative to close ones, and is a ubiquitous phenomenon across decision-making agents including people, animals, and organizations. Temporal discounting is rational if it occurs at an exponential rate δ, where the value V of a reward r at time t is

$$V(r,t) = re^{-t\delta}$$

In exponential discounting, each additional unit of waiting time decreases the value of a reward by an equal proportion (Samuelson, 1937; Frederick et al., 2002). This means that the relative values of an early and a late rewards are the same no matter what time point they are considered from, or equivalently that their relative values are unaffected by adding an additional waiting time to both.

An extensive literature documents that people and animals violate exponential discounting. Specifically, in the short term rewards are discounted at a steep rate with each additional unit of waiting time, while in the long term rewards are discounted at a shallow rate. This sort of non-exponential discounting leads to a present bias, in which in the short term people excessively over-weight immediate over future rewards. For example, people will often prefer a larger, later monetary reward to a smaller, sooner reward when both rewards are in the future, but will switch their preference when the time until both rewards is reduced so that the sooner reward is immediate or nearly immediate (Kirby & Herrnstein, 1995). With monetary rewards, the delay or speed-up required to observe preference reversals is usually several days. With non-monetary rewards, such as the cessation of an annoying noise (Solnick et al., 1980), watching a video when bored (D. Navarick, 1998), or drinking soda when thirsty (Brown et al., 2009), a bias towards immediate rewards has been observed on the scale of minutes or seconds.

There is debate about how to formally describe non-exponential discounting. Many studies have found that humans and animals appear to discount future rewards at a hyperbolic rate, allowing the value of a future reward to be written as

$$V(r,t) = \frac{r}{1+kt}$$

for appropriate constant k (Myerson & Green, 1995). This formulation often fits data well, but is difficult to deal with analytically. An alternative is to posit that discounting after the present proceeds exponentially, but that there is a one-time drop in value when the reward goes from being immediate to being in the future (Laibson, 1997). In this model, known as the beta–delta or quasi-hyperbolic model, the value of a future reward is

$$V(r,t) = \begin{cases} r & \text{if } t = 0 \\ \beta r e^{-t\delta} & \text{if } t > 0 \end{cases}$$

where δ is the rate of exponential discounting and β is the degree of present bias. This model suffers from ambiguity in when exactly the "present" ends and the future begins (e.g., should the value of reward received in 30 seconds be discounted by β, or should it be considered immediate?). However, it captures in a simple and tractible way many of the qualities of human intertemporal choice, and for this reason we will adopt it for our additional analyses below.

2.3 Exploration and temporal discounting

The rewards from exploratory choice are inherently distributed over time. In expectation, an exploitative action yields the greatest reward in the present, because it is the action *currently believed* to yield the highest reward. An exploratory action is expected to yield less immediate reward, but it can compensate for this by providing useful information. This information can allow the decision-maker to make better choices in the future, leading to higher rewards later on.

Thus, temporal discounting plays a central role in determining the balance between exploration and exploitation. Rational, exponential discounting ensures that a decision-making agent explores neither too little nor too much given its degree of interest in the future. Some degree of discounting is generally good, because at some point the distant gains from continued exploration are not worth their immediate costs (Le Mens & Denrell, 2011). But as past theoretical work has highlighted, discounting that is too steep or that exhibits a present bias can lead to chronic over-exploitation and under-exploration (March, 1991; Levinthal & March, 1993).

To understand how patterns of discounting affect exploration, consider a simple scenario in which an agent must make a sequence of choices between two actions. Action A is to choose a sure-bet option that always provides a payoff of 2. Action B is to choose from a large set of uncertain options. For each uncertain option, there is a 25$\%$ chance

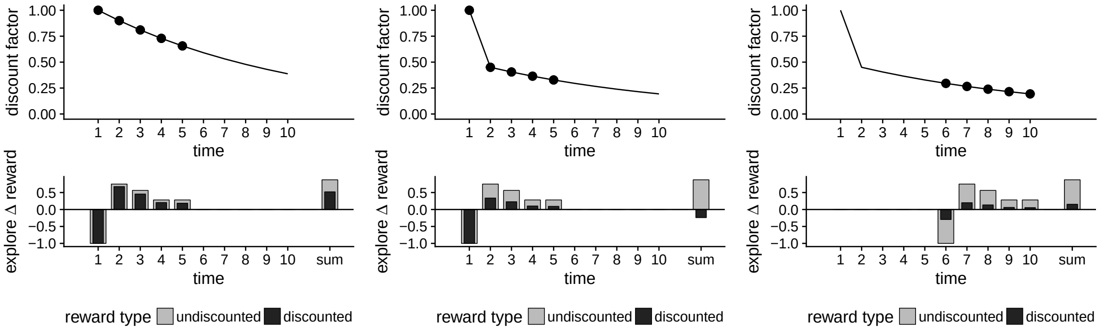


Figure 2.1: Effects of exploration over time for different discount curves in a simple exploratory choice task (see text for more details). The top row of panels show the degree of discounting at each time step. The bottom row of panels show the expected change of undiscounted (gray) and discounted (black) reward at each time step from exploring at the *first* action. The left panels shows exponential discounting, the center panels show quasi-hyperbolic discounting, and the right panels show quasi-hyperbolic discounting with a front-end delay. Exploration appears worthwhile to an agent with exponential discounting or quasi-hyperbolic discounting with a delay, but not to an agent with quasi-hyperbolic discounting and no delay.

that it produces a consistent payoff of 4, and a 75% chance that it produces a payoff of 0. Once a high-payoff uncertain option is found, it can be selected on every subsequent choice.

This scenario presents an explore–exploit dilemma because as long as a high-payoff option has not been found, the best immediate action is A, with an expected payoff of 2, rather than B, with an expected payoff of $.25 \cdot 4 = 1$. Long term payoffs, in contrast, are increased by exploring the options available through action B, because the agent may find a high-payoff option that can be exploited on all future choices.

Whether the agent decides to forgo the immediate gains of exploiting A in order to explore B will depend on how much it values the future. Figure 2.1 shows the effects of various patterns of discounting on the expected rewards over a sequence of five choices. The left column shows the case of exponential discounting with $\delta = .9$. The top graph shows the exponential discounting curve, with dots indicating the time and weight of each of the five choices. The bottom graph shows the change in expected reward at each choice that is caused by selecting action B rather than A at the *first* choice. (This analysis assumes that all subsequent choices are made optimally in terms of undiscounted rewards.) For mild exponential discounting, we see that at time 1, choosing B over A causes a steep decrease in expected reward, because it trades an expected payoff of 2 for an expected payoff of 1. At times 2–5, however, the expected payoff goes up; choosing B at time 1 can only increase payoffs at later times, by revealing an high-payoff option. At the far right of the graph, we see that the summed discounted change in reward, in black, is positive, and thus that the agent will choose to explore. The undiscounted reward, in gray, is larger, but doesn't differ in sign from the reward after mild exponential discounting.

The center column shows the case of beta–delta, or psuedo-hyperbolic, discounting, with $\delta = .9$ and $\beta = .5$. As the top graph shows, rewards from later time points are weighted much less than in exponential discounting. Because of this, the expected gain in future reward for choosing B becomes smaller, while the immediate expected loss remains the same. The summed discounted change in reward becomes negative, and the agent adopts a completely exploitative policy of choosing A instead of initially exploring the uncertain action B.

To preview our experimental manipulation, the right column shows a case of beta–delta discounting considered from a temporal distance. Now, the first choice is at time

6, while the last is at time 5. Suppose the agent was given the opportunity to commit to a first action from time 1. As shown in the bottom graph, the sequence of rewards viewed from this distance is highly discounted, but is no longer heavily biased towards the first outcome. Instead, the expected discounted reward sequence is now a scaled version of the exponentially discounted rewards, since beta–delta discounding is identical to exponential discounting after the present. The summed expected rewards for exploration are greater than those for exploitation, so the agent will choose the exploratory action.

2.3.1 Capturing present bias in exploratory choice

As alluded to earlier, several approaches have been used to study present bias in the lab. Many studies use monetary rewards, and offer participants various one-off choices between different quantities of money at different delays to determine their discounting curve (Myerson & Green, 1995). However, exploratory choice is inherently not "one-off." Choices can only be considered exploratory or exploitative if they are embedded within an ordered sequence of choices, where the knowledge gained from one choice can be used to inform the next. Thus, to study present bias during exploratory choice, an experiment must include a sequence of choices and outcomes, with enough time between them for discounting of the future to be non-negligible. With monetary rewards, this means the choices in an experiment would have to be spread out over weeks or months. This leads us to consider non-monetary, directly consumable rewards.

While people tend to discount money relatively slowly, they often discount primary rewards significantly for delays of minutes or seconds. This can be measured in a number of ways. In some cases, an explicit choice between a larger later and a smaller sooner reward is offered. Mcclure, Ericson, Laibson, Loewenstein, and Cohen (2007), for example, found that thirsty participants showed present bias when asked to choose between larger and smaller juice rewards separated by a few minutes. In other cases, the choice between a smaller-sooner and larger-later reward is offered repeatedly, but without explicit description, and participants are allowed to build a preference through experience. Using this approach, researchers have found present biases on the scale of seconds for playing a video game, watching a movie, or relief from an annoying noise (D. Navarick, 1998; Millar & D. J. Navarick, 1984; Solnick et al., 1980).

In the above studies, each choice is "one-off," creating rewards but not affecting future choices. Brown et al. (2009) provided evidence of present bias in a task in which immediate consumption affected consumption from future choices. They created a life-cycle savings game in which participants gained income and decided how much to spend over 30 periods spaced a minute apart. They arrived to the experiment thirsty, and were allowed to consume their spent income in the form of soda. In the immediate-reward condition, participants made choices at each period and then immediately consumed their soda reward. In the delayed-reward condition, the experimenters imposed a 10 minute delay between choices and reward consumption; thus, after a choice was made, the soda earned from that choice was consumed 10 periods later. Participants in the delayed-reward condition were able to consume more total soda on average, suggesting that the temporal delay decreased their present bias and allowed them to choose in a manner leading to greater long-term reward.

In the following two experiments, we use an intervention similar to that of Brown et al. (2009) to test for effects of present bias on exploratory choice. As indicated in Figure 2.1, if an exploratory choice task is paired with immediate consumption we predict present bias to lead to underexploration. However, if a temporal delay is introduced between decisions and outcomes, the present bias will be decreased, leading to greater exploration.

We used videos as a positive outcome that could produce present bias (D. Navarick, 1998), and a boring slider task (Gill & Prowse, 2012), along with, in Experiment 2, annoying noises (Solnick et al., 1980), as negative outcomes. It is worth noting that we also piloted an experiment using a video game as a positive outcome (Millar & D. J. Navarick, 1984), but found that participants did not find the video game sufficiently enjoyable. Experiment 1 represents a first attempt to test for effects of present bias on exploratory choice, and Experiment 2 is a larger, preregistered study that improves on Experiment 1 in several ways. After finding no evidence of present bias producing an effect in Experiments 1 or 2, in Experiment 3 we test directly, using a simpler design, whether our outcome stimuli in fact produce a consistent preference towards immediate rewards.

2.4 Experiment 1

2.4.1 Methods

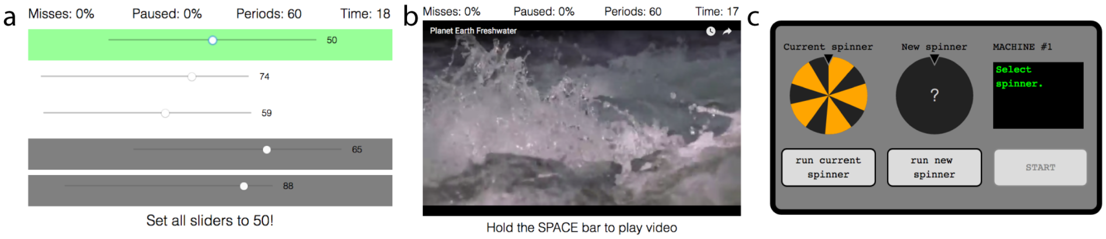
Participants

Forty participants completed the experiment, which was conducted over Amazon Mechanical Turk (AMT) using the psiTurk framework (Todd M Gureckis et al., 2015). The participants had a mean age of 37.7 (SD=10.6). Twenty eight self-reported female, twelve male. Participants were paid $5.00 for their participation, with a performance-based bonus of up to $3.00. The experiment and all following experiments were approved by the Institutional Review Board at New York University. All participants received the full $3.00 bonus. Participants were pseudo-randomly counterbalanced across two conditions.

Design and procedure

Consumption tasks. Participants were informed that their job was to perform a monotonous slider task that would be split into 30-second "work periods," but that they would be able to make choices throughout the experiment that would give them a chance to watch a YouTube video instead. The number of remaining work periods in the experiment was shown at the top of the screen, as was the number of seconds left in the current work period.

The slider task was based on a task previously used by (Gill & Prowse, 2012). In each period of the slider task, five horizontal sliders appeared on the screen (Figure



Misses: 0% Paused: 0% Periods: 60 Time: 18

Set all sliders to 50!

Misses: 0% Paused: 0% Periods: 60 Time: 17

Hold the SPACE bar to play video

Select spinner.

Figure 2.2: Examples of the Experiment 1 tasks. (a): The slider task. Participants had to move all sliders to "50" in 30 seconds. (b): The video-watching task. Participants had to hold the space bar to watch their chosen video. (c): The decision-making task. Participants had to choose to run the machine with the current spinner or try a new spinner. If their chosen spinner landed on a gold wedge, they performed the video-watching task instead of the slider task.

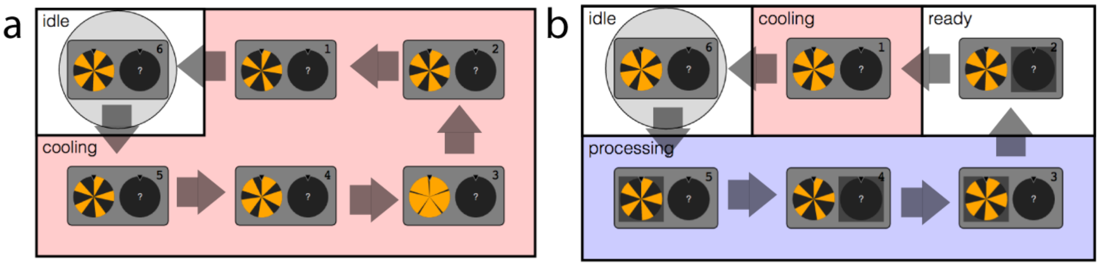


Figure 2.3: The machine display seen by participants. The display allowed participants track the value of each machine and the next time each machine would be ready to make a choice or produce an outcome. Gray arrows have been added to depict the counterclockwise movement of machines around the display after each work period. (a): the display seen by participants in the immediate condition of Experiment 1. (b): the display seen by participants in the delayed condition of Experiment 1.

2.2a). Each started at a random setting between 0 and 100, with the slider's value shown to its right, and with a random horizontal offset so that the sliders were not aligned. The participant's task was to use the mouse to move each slider to "50" before the work period ended. When a participant released the mouse at the correct setting, the slider turned green to show it had been completed. To ensure that the task took close to the allotted 30 seconds, at the beginning of the task only the top slider was enabled, and the other four were grayed out. Additional sliders were enabled at five-second intervals, such that all five sliders were available after 20 seconds.

Before beginning the experiment, participants chose one of four videos available on YouTube: an episode of "Planet Earth", and episode of "The Great British Bakeoff", and episode of "Mythbusters", or an "Ellen Degeneres comedy special". The video was embedded in the experiment with all user controls (such as skipping ahead) disabled (Figure 2.2b). When given access to the video, participants had to keep the browser window open and hold down the space bar for the video to play. This allowed us to ensure that participants maintained engagement with the content.

Participants completed a total of 70 work periods. For the first 10 work periods, participants simply clicked a button to begin the slider task. After these initial periods, participants gained access to six machines that could potentially complete the slider task for the participant, allowing the participant to watch their chosen video instead. However, the machines did not always function, and participants had to make a decision about how to set the machine before each use.

Decision-making task and conditions. Following the initial 10 periods, participants were shown a machine before each work period and, as shown in Figure 2.2c, had to select between two circular spinners with arrows at the top: the "current spinner" (exploit) and the "new spinner" (explore). The current spinner was split into five black and five gold wedges. If a participant chose the current spinner, it spun and, if the arrow landed on gold, the machine worked and the participant could watch the video. Initially, the current spinner's gold wedges were randomly set for each machine to cover between 1/3 and 2/3 of the spinner.

The new spinner initially showed a question mark. If a participant chose the new spinner, then a new spinner was created and appeared on the machine. The new spinner's gold wedge could cover anywhere from 0% to 100% of the spinner. The new spinner then spun and, if the arrow landed on gold, the machine worked.

Critically, if the new spinner had a greater gold area than the current spinner, the new spinner was "saved" and the current spinner was updated to the new spinner. This created an explore–exploit tradeoff in which choosing a new spinner carried immediate risk, but could carry long-term benefits by improving the current spinner from its initial value.

The experiment's two conditions differed in what occured after the participant spun the spinner. In the immediate condition, the machine ran immediately after the choice was made and affected the next work period, as shown in Figure 2.3a. It then "cooled off" for the following five periods, as choices were made with the other five machines. In the delayed condition, each machine was presented to the participant four work periods before it was scheduled to run, and the participant made a choice at that time. The machine then had to "process" for four work periods, thus delaying the outcome by over 2 minutes (Figure 2.3b). The participant then returned to the machine to observe its outcome and either perform the slider task or watch the video. The machine then cooled off for a single period before being ready for another choice.

Finally, in order to induce exploration throughout the entire experiment, the six machines would occasionally "reset" after they ran. When this occurred, the current spinner would be set to a new random value between 1/3 and 2/3 gold. Participants were informed that this would occur randomly on 1/6 of trials. In fact, the procedure was designed so exactly one of the six machines would reset on each cycle through the machines, and no machine would be reset on two consecutive uses.

Training, incentives, and post-experiment questions. Before beginning the full experiment, participants completed two practice phases. First, they performed several trials of practice using the machines, with the actual work periods removed. Then, they performed two work periods practicing the slider task and one work period practicing the video task. During the machine choices, participants had access to an "info" button at the bottom of the screen that provided reminders about the dynamics of the task.

Participants were given a performance-based bonus of $3.00 for completing the consumption tasks accurately. If they missed fewer than 10% of sliders throughout the experiment and left the video paused less than 20% of the time, they were not penalized. However, if they missed more sliders or left the video paused for longer, they lost 10 cents from their bonus for each additional percentage of sliders missed or time with the video paused. The running percentage of sliders missed and video pause time was displayed at the top of the screen throughout the experiment.

Following the experiment, participants were asked to rate their enjoyment of the slider task and of the video-watching task on a 1 to 7 scale.

2.4.2 Results

As a basic check of our consumption tasks, we confirmed that participants rated the video as more enjoyable on average (5.65 out of 7) than the slider task (3.13 out of 7), $t(39) = 9.26, p < .001$.

To analyze participants' trial-by-trial decision-making, we conducted a hierarchical Bayesian logistic regression using the Stan modeling language (Stan Development Team, 2015). This approach allowed us to estimate population-level effects of the current-spinner value and of condition, while also allowing for individual differences. The regression model included an intercept term as well as terms for the value of the current spinner, the participant's condition, and a condition by value of current spinner interaction. We included predictors for the value of the current spinner, the participant's condition, and the interaction between condition and current spinner value. Condition was coded as -1 for the immediate condition and 1 for the delayed condition; current spinner value was rescaled to have zero mean and unit variance across participants. We assumed that individuals could vary in their overall tendency to explore (i.e., intercept) as well as their responsiveness to current spinner value (slope). Participants' individual-level parameters were assumed to be drawn from a t distribution with $df = 5$, making our population level estimates robust to potential outliers. The priors on the

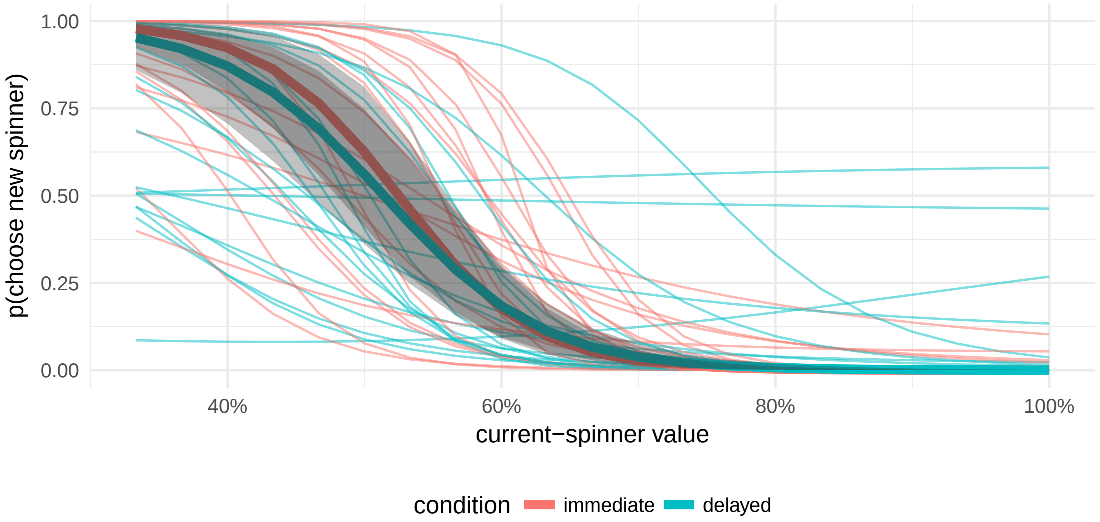


Figure 2.4: Model-based estimates of participants' probability of choosing a new spinner for different values of the current spinner in Experiment 1. Thick lines and shaded regions indicate the mean and 95% posterior interval for the population-level parameters, while the thin lines indicate the mean posterior parameters for each of the 40 individual participants. Participants in the delayed-outcome condition were no more likely to explore at a given current-spinner value than those in the immediate-outcome condition.

the population-level predictor coefficients, and on the standard deviation of the t distributions from which individual-level parameters were drawn, were (truncated) normal distributions with a mean of 0 and a standard deviation of 5.

The model posterior was estimated using the Stan modeling language (Carpenter et al., 2017). We ran four chains of Hamiltonian Monte Carlo sampling, with 1000 samples per chain, the first half of which were discarded as burn-in. We confirmed convergence using the $\hat{R}$ convergence criterion (A Gelman et al., 2014). In the results below, we report 95% credible intervals (CIs) on model parameters of interest. An overview of the model posterior is displayed in Figure 2.4.

Participants were less likely to choose a new spinner when the current spinner has a high value, $CI = [-4.13, -2.52]$. However, in this experiment we found no evidence of an effect of condition, $CI = [-.58, .73]$. This means that participants were no more likely to explore a new spinner when there was a temporal delay imposed between their choices and the received outcomes. However, there may have been a small interaction between current spinner value and condition, such that participants in the delayed condition were less sensitive to the value of the spinner when making their choices $CI = [-.27, 1.42]$. This might indicate that the delayed condition was confusing to some participants, as a few individuals (as seen in Figure 2.4) changed their behavior very little across current-spinner values.

2.5 Experiment 2

In Experiment 1, we found no evidence of the delay in rewards leading to an increase in exploratory choice. However, there were several potential flaws in the experiment design which may have prevented present bias from occurring or its effects from being observed. In Experiment 2, we preregistered the design, collected a larger sample, and attempted to improve on Experiment 1 in several ways.

We conducted Experiment 2 in person, rather than using AMT. This ensured that participants had few distractions from the consumption tasks, potentially increasing their motivational effect. We also made the slider task more aversive and the video task more pleasant. To do so, we added an intermittent static noise during the slider tasks, and allowed people to switch among the four videos at will, without having to hold down the space bar to keep the video playing.

To simplify and improve the exploratory choice task, in Experiment 2 there was a single machine, rather than six. Rather than the machine "processing" for four trials in the delayed condition, outcomes were added to a "work queue" that delayed the consumption task by eight trials. This was both simpler and increased the delay length. The visual appearance of the exploratory choice task was also redesigned to make the statistics of the task more transparent.

Finally, we measured participants' impulsivity, a potentially important covariate, using the Barratt Impulsiveness Scale (Patton, Stanford, & Barratt, 1995). There is evidence that this scale correlates with present-focused behavior in repeated choice tasks (a. R. Otto, Markman, & Love, 2012), though other studies have not found a relationship (Brown et al., 2009).

2.5.1 Methods

The experiment was preregistered through the Open Science Framework. The preregistration can be found at: osf.io/3r9ke.

Participants

One hundred people from the general community took part in the study in person at New York University. The participants had a mean age of 23.9 (SD=6.1). Fifty eight self-reported female, forty one male. Participants received $10 for taking part in the study, which lasted approximately one hour, and received a performance-based bonus

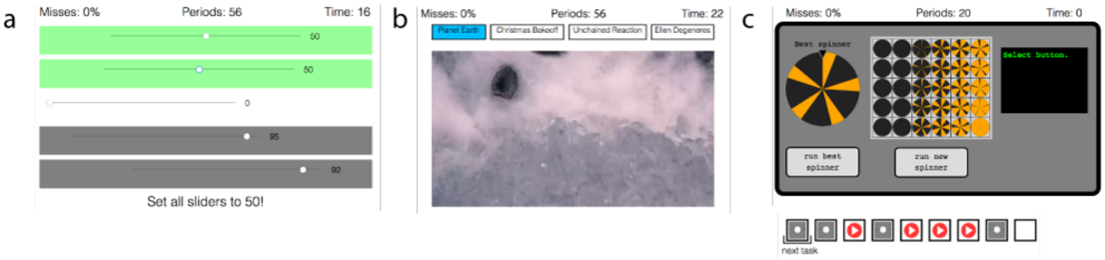


Figure 2.5: Examples of the Experiment 2 tasks, which resemble the Experiment 1 tasks. (a): the slider task. (b): the video task. (c): the decision-making task. After making a choice in the decision-making task, the produced outcome was added to the work queue, pictured at the bottom of (c).

of up to $5. All but one participant received a bonus of $5, with the remaining participant receiving $4.4. Participants who failed a post-instructions questionnaire more than twice were excluded from further analyses. Ten participants were excluded in this manner.

Design and procedure

Barratt Impulsiveness Scale. Prior to reading the experiment instructions, participants completed the 30-item Barratt Impulsiveness scale (Patton et al., 1995) on the computer.

Consumption tasks. Participants were informed that there were two types of tasks, a "slider task" and a "video task," that they would complete during 30-second "work periods." The number of remaining work periods in the experiment was shown at the top of the screen, as was the number of seconds left in the current work period.

The slider task was the same as the one described in Experiment 1, and is pictured in Figure 2.5a. To make the slider task more unpleasant, a short static noise was played through the computer speakers at a moderate volume (78db) at irregular intervals of approximately once every three seconds during the task.

As in Experiment 1, the video tasks consisted of simply watching one of four videos: an episode of “Planet Earth”, and episode of “The Great British Bakeoff”, and episode of “Unchained Reactions”, or an Ellen Degeneres comedy special. Participants watched the video through a player on the computer screen. Unlike in Experiment 1, they did not have to hold down a button to play the video. They were free to fast forward or rewind the video at will, and could also switch among the videos at any time by clicking one of four tabs above the player (see Figure 2.5b).

Choice task. Participants completed a total of 56 work periods. The first eight were automatically spent performing the slider task. For the remaining 48, participants made a choice prior to each work period that determined whether the work period would be devoted to the slider task or the video task. This choice task resembled the choice task used in Experiment 1.

Participants were shown a “machine” that could create slider or video tasks (Figure 2.5c). The machine consisted of a black-and-gold “best” spinner and a panel of possible new spinners. Participants selected either “run best spinner” or “run new spinner”. If the participant selected “run best spinner”, the spinner would visually rotate on the screen. If it landed on gold, the machine created a video task; if it landed on black, the machine created a slider task.

If the participant selected “run new spinner”, the new spinners in the panel were covered up and randomly shuffled. The participant then clicked one of the gray squares, revealing the new spinner underneath. As was explained to the particpants, and was visually apparent, one third of the possible new spinners are completely black, while the remaining two thirds range from 5% to 100% gold, in even increments of 5%.

After revealing a new spinner, it was spun, producing a video or slider task in the same manner as the "best spinner". As in Experiment 1, if the new spinner selected had a higher proportion gold than the best spinner, it would replace the best spinner for future choices.

Participants were also informed that after every work period there was a one in six chance that the machine would reset itself. In fact, the experiment was designed so that there was exactly one reset in every set of 6 trials (i.e., trials 1–6, 7–12, etc.). When the machine reset, the "best spinner" was set to a new starting value. The starting values following resets (including the intial starting value) were {20%, 25%, … 55%, 60%}, randomly ordered.

Immediate and delayed conditions. Participants were pseudo-randomly assigned to one of two conditions. In the Immediate condition, participants completed the task produced by a choice in the work period immediately following the choice. In the Delayed condition, participants completed the task produced by a choice after eight intervening work periods had passed, which was about five minutes after making the choice. This means that participants in the Delayed condition began making choices during the initial eight slider task work periods, in order to have outcomes determined when they reached the ninth and later work periods.

To make this delay intuitive, participants were shown a work queue at the bottom of the screen that contained eight tasks (see the bottom of Figure 2.5c). In the Delayed condition, upon making a choice a new slider or video task icon was added to the right of the queue, and then the leftmost task on the queue was performed and removed. In the Immediate condition, participants were still shown the cue, but upon adding an icon to the right of the queue that outcome was performed immediately. This means that in the Immediate condition the queue acted simply as a history of the past eight outcomes.

Training, incentives, and post-task questions. As in Experiment 1, participants had opportunities to practice the decision making and consumption tasks prior to the main task, and rated their enjoyment of the slider and video tasks on a scale from 1 to 7.

To incentivize participants to attend to and perform the slider task, they were penalized if they missed more than 10% of the sliders. For each percentage over 10% of sliders that were not set to 50 over the course of the experiment, $.20 was deducted from a bonus that started at $5.00.

2.5.2 Results

Our primary hypothesis was that participants in the delayed-outcome condition would take more exploratory actions (that is, choose a new spinner more often) than those in the immediate-outcome condition. A secondary hypothesis was that this change would be moderated by participants' scores on the Barratt Impulsivity Scale. Specifically, we expected that highly impulsive participants would explore less and show a bigger difference in exploration between the delayed and immediate conditions.

We tested these predictions via hierarchical Bayesian logistic regression on participant choices. All aspects of this analyses were preregistered prior to data collection.

We included predictors for the value of the current spinner, the participant's BIS score, the participant's condition, and interactions between condition and current spinner value and condition and BIS score. Condition was coded as -1 for the immediate condition and 1 for the delayed condition; current spinner value and BIS score were rescaled to have zero mean and unit variance across participants. We assumed that individuals could vary in their overall tendency to explore (i.e., intercept) as well as their responsiveness to current spinner value (slope). Participants' individual-level parameters were assumed to be drawn from a t distribution with $df = 5$, making our population level estimates robust to potential outliers. The priors on the the population-level predictor coefficients, and on the standard deviation of the t distributions from which individual-level parameters were drawn, were (truncated) normal distributions with a mean of 0 and a standard deviation of 5.

The model posterior was estimated using the Stan modeling language (Carpenter et al., 2017). We ran four chains of Hamiltonian Monte Carlo sampling, with 1000 samples per chain, the first half of which were discarded as burn-in. We confirmed convergence using the $\hat{R}$ convergence criterion (A Gelman et al., 2014). In the results below, we report 95% credible intervals (CIs) on model parameters of interest. An overview of the model posterior is displayed in Figure 2.6

We found a strongly negative effect of current spinner value on participant's probability of selecting a new spinner, $CI = [-4.18, -3.08]$. This indicates that participants understood the general structure of the task, and explored (i.e., selected a new spinner) only when it was advantageous to do so. While the estimate was in the predicted direction, we found no clear effect of condition on the tendency to explore $CI = [-.11, .61]$.

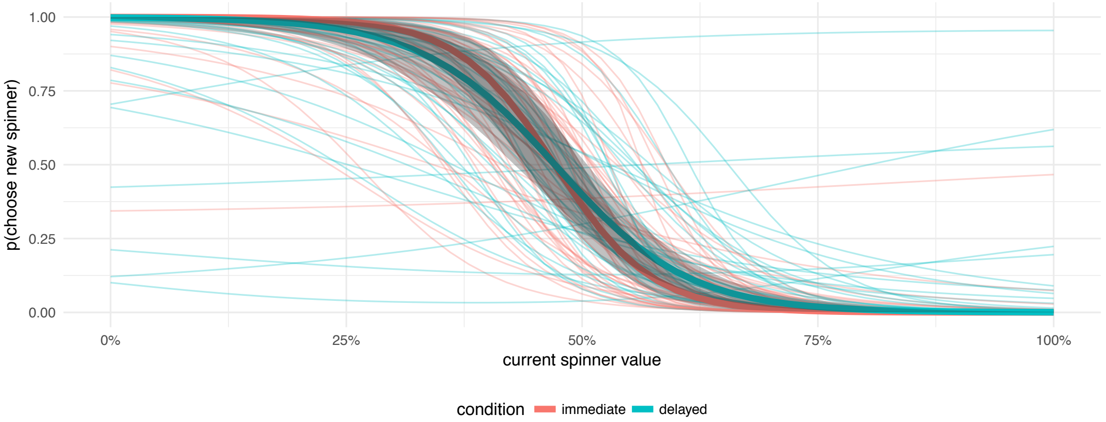


Figure 2.6: Model-based estimates of participants' probability of choosing a new spinner for different values of the current spinner in Experiment 2. Thick lines and shaded regions indicate the mean and 95% posterior interval for the population-level parameters, while the thin lines indicate the mean posterior parameters for each of the 100 individual participants. Participants in the delayed-outcome condition were no more likely to explore at a given current-spinner value than those in the immediate-outcome condition.

Additionally, there was no effect of BIS score on behavior $CI = [-.34, .36]$ and no interaction between BIS score and condition $CI = [-.21, .51]$. We did find a positive interaction between condition and current spinner value, $CI = [.07, 1.06]$. This means that while people in the delayed condition were not more or less likely to explore in general, they were more likely to explore for high current spinner values, and less likely to explore for low current spinner values. In other words, they were less sensitive to the current value of the spinner.

Our preregistered analyses provided no support for our hypotheses. As an additional, exploratory analysis, we re-ran the Bayesian model replacing participants' BIS scores with their ratings difference between the slider task and the video task in post-experiment questionnaire. Overall, participants rated the video task as more enjoyable (6.37 out of 7 on average) than the slider task, (3.43 out of 7), $t(89) = 16.3, p < .001$. Our intuition was that participants who rated the video task much higher than the slider task may have felt a greater motivational pull to immediately watch a video instead of move slider, and may thus have been more susceptible to the delay manipulation. However, we found no main effect of ratings difference on exploration $CI = [-.32, .16]$, and no interaction between ratings difference and condition $CI = [-.33, .38]$. All other effects remained qualitatively the same.

Finally, we examined whether the lower sensitivity to current spinner value in the delay condition might indicate that a group of participants in that condition were responding near-randomly, possibly due to confusion with the task, and if this could affect our other results. We found that that the individual-level effect of current spinner value did not differ significantly from zero for 14 of 44 participants in the delayed condition, and only 2 of 46 participants in the immediate condition. To determine whether these near-random participants influenced our results, we re-ran our preregistered regression, including only the 30 participants with the highest-magnitude slopes in each condition from the initial analysis. We did not find that this new selection criterion affected our results. In particular, the credible interval for the main effect of delay still included zero, $CI = [-.08, .72]$.

2.6 Experiment 3

In both Experiments 1 and 2, we found no evidence that delaying rewards affected the degree of exploratory behavior, and thus no evidence that exploratory choice is influenced by present bias. Experiment 2 attempted to fix several flaws of Experiment 1 by collecting data in person, making the consumption tasks more pleasant or more aversive, increasing the reward delay, and simplifying the exploratory choice task. This may indicate that there is truly no effect of present bias on exploratory choice, but it remains possible that this null effect is due to a weakness in our experiment design.

The most apparent potential weakness is that the consumption tasks did not induce very much present bias, or that discounting of these stimuli occurred on a scale much longer than the delay of a few minutes used in our experiments. Our use of these stimuli was based on several past studies. Access to videos has been shown to induce present bias with a delay of around a minute (D. Navarick, 1998), and cessation of annoying noises can induce present bias with a delay of around ten seconds (Solnick et al., 1980). However, these studies were were small and differed from the current setting in important ways. For example Solnick et al. (1980) had participants make choices about noise cessation while solving math problems, which prevented them from focusing fully on the choice task.

It may be that the consumption tasks and setting we used did not, in fact induce present bias, which would mean that inducing a delay would have no predicted effect. Therefore, in Experiment 3 we conducted a simple follow-up experiment using the two consumption tasks to test whether people have a present bias for watching the videos and avoiding the slider task and static noises, based on the design of past experiments which studied time preferences for videos or video games (Millar & D. J. Navarick, 1984; D. Navarick, 1998).

2.6.1 Methods

Participants

Twenty eight undergraduate students at New York University took part in the study for extra credit. The participants had a mean age of 19.0 (SD=1.0). Eighteen self-reported female, ten male. Participants received a performance-based bonus of up to $5. All participants received the full $5 bonus. Participants who failed a post-instructions questionnaire more than twice were excluded from further analyses. One participant was excluded in this manner.

Design and procedure

As in Experiment 2, the Barratt Impulsiveness Scale was administered prior to completing the main task.

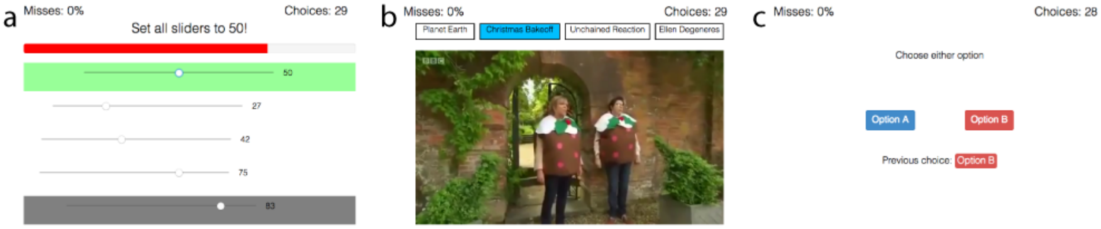


Figure 2.7: Examples of the Experiment 3 tasks. (a): an example of the slider task, in which the numeric timer has been replaced by a red timer bar. (b): an example of the video task. (c): an example of the decision-making task.

In the main task, shown in Figure 2.7c, participants were instructed that they would have to make a series of choices between two buttons. They were told that after selecting a button they would spend some amount of time performing a boring slider task and a fun video task, and that their choice could affect the amount of time spent on each task and the order of the tasks. They were also instructed that for their first two choices they would have to click first one button, then the other, to ensure that they had experienced both outcomes, and that occasionally the outcomes would change, at which point they would be instructed to try each of the two buttons again. On all other trials, they were told to select whichever button they preferred. Participants' previous choice was displayed at the bottom of the screen as a memory aid.

The slider and video tasks were very similar to the tasks used in Experiment 2 (Figure 2.7a/b). To prevent participants from explicitly measuring the amount of video and slider time following a choice, the timer showing how many seconds remaining in the consumption task was removed. For the slider task, it was replaced by horizontal red "progress bar" that steadily shrank over the course of the task, thereby indicating seconds remaining. For the video task, there was no indication of seconds remaining. In addition, instead of always lasting 30 seconds, the consumption tasks varied in length. For a slider task that lasted s seconds, there were $s/5 - 1$ sliders to complete.

After practicing the slider and video tasks, participants completed 30 trials of the choice task. This was divided into three groups of ten trials, each of which had a new pair of outcomes. The outcomes always lasted 90 seconds in total, and for each group there was always one button that produced the video task immediately, followed by the slider task, and one that produced the slider task immediately, followed by the videos. The reward amounts and reward orders of the three groups were as follows:

1. 30s videos/60s sliders vs. 60s sliders/30s videos
2. 35s videos/55s sliders vs. 65s sliders/25s videos
3. 25s videos/65s sliders vs. 55s sliders/35s videos

The ordering of the three pairs of outcomes was counterbalanced across participants, and the pairing of outcomes the left and right button was randomized. Absent discounting, participants should be indifferent between the two options in pair 1, and prefer the options with more video time in pairs 2 and 3. However, we predicted that while amount of video time would also matter, participants would display a bias towards selecting the option with the immediate video task.

As in Experiments 1 and 2, participants were asked to rate their enjoyment of the two consumption tasks following the experiment. They were also penalized for missing sliders using the same scoring method as Experiment 2.

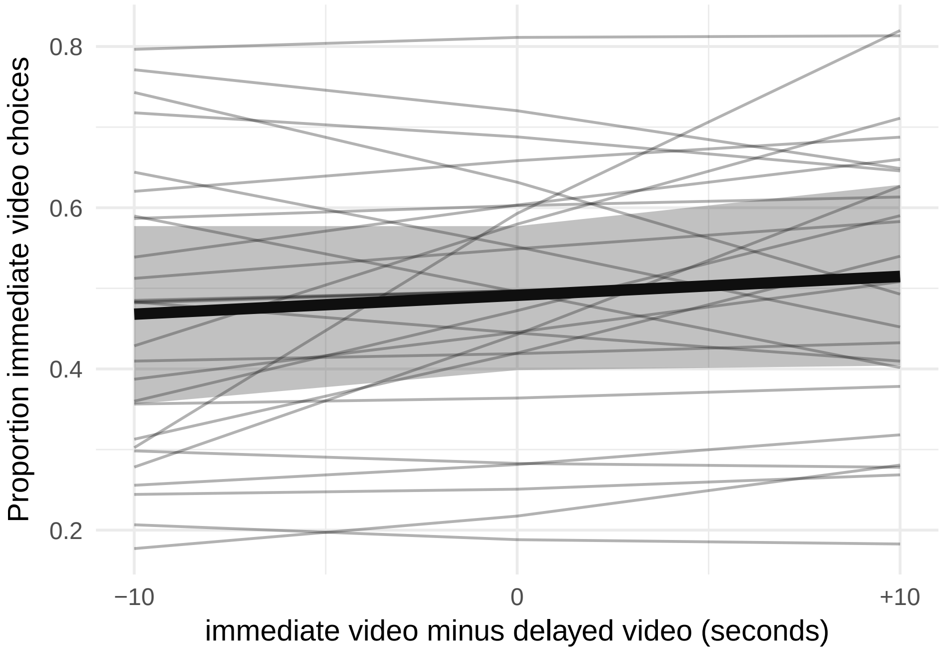


Figure 2.8: Model-based estimates of participants' probability of choosing the immediate-video option when that option provides 10 seconds less than, the same as, or 10 seconds more than the delayed-video option. The thick line and shaded region indicates the mean and 95% posterior interval for the population-level parameters, while the thin lines indicate the mean posterior parameters for each of the individual participants. Participants showed neither a preference for immediate video not a clear preference for greater video amounts.

2.6.2 Results

As in Experiments 1 and 2, participants self-reported liking the video task more than the slider task, $t(26)=5.70, $p < .001$. They rated the videos 5.0 out of 7, on average, and the sliders 2.8 out of 7, on average.

To test whether this preference was expressed in the participants choice behavior, and whether participants preferred watching the video immediately, we conducted a hierarchical Bayesian logistic regression on trial-by-trial choices. We included individual-level parameters for over-all preference for immediate reward and for the effect of the difference in video time between the immediate and delayed options. The option difference was coded as -1 for the 25 second immediate video, 0 for the 30 second immediate video, and 1 for the 35 second immediate video. As in earlier analyses, participants' parameters were assumed to be drawn from a t distribution with $df = 5$. The priors on the the population-level predictor coefficients, and on the standard deviation of the t distributions from which individual-level parameters were drawn, were (truncated) normal distributions with a mean of 0 and a standard deviation of 5. We estimated the model using the Stan modeling language using the same procedure as Experiments 1 and 2 (Carpenter et al., 2017).

The model results are plotted in Figure 2.8. We found that participants were not very sensitive to our experimental manipulations. While some individual participants appeared to prefer either the immediate or the delayed option, the population as a whole showed no average preference, $CI = [-.41, .31]$. Participants also showed almost no sensitivity to which option produced more video time, $CI = [-.19, .37]$, even though choosing the superior option in the two non-equal-time groups of trials would have allowed them to watch 10 seconds more video and perform two fewer sliders on each trial. It may be that this manipulation was too subtle for many participants to become aware of it.

2.7 Discussion

In this study, we set out to test whether people's well-documented bias towards immediate rewards affects exploratory choice in a manner similar to other intertemporal choices, thus producing a bias towards under-exploration (Frederick et al., 2002; Myerson & Green, 1995). We tested this using a paradigm in which we added a temporal delay to outcomes in an exploratory choice task, following the design of many prior intertemporal choice tasks (Kirby & Herrnstein, 1995; Solnick et al., 1980; Brown et al., 2009). In two experiments we found that adding a temporal delay did not affect exploration, suggesting that people treat exploratory choices differently from intertemporal choices. However, a followup study showed that the rewards used in our task did not consistently produce present bias, making it difficult to draw clear conclusions from our earlier results. In this section, we briefly discuss why our rewards may not have been motivationally effective, as well as delve into the similarities and differences between exploratory and intertemporal choice.

2.7.1 Revisiting immediately consumable rewards

Experiments 2 and 3 used three types of consumable rewards to try to induce a present bias: enjoyable videos, a boring slider task, and an annoying static noise. The time difference within which we expected to observe a present bias was about five minutes (Experiment 2) or one minute (Experiment 3). While the slider task has not been used in the past to produce present bias, both videos and static noises have. D. Navarick (1998) found that 40% of participant showed a consistent strong bias towards watching an immediate shorter video, even when a longer video could be obtained by waiting 30 seconds. Solnick et al. (1980) found that a 90 second cessation of noise was preferred over a 120 second sessation with a 60 second wait, but found that this preference flipped when a front-end delay of only 30 seconds was added. And in a similar vein, Millar and D. J. Navarick (1984) found that people strongly preferred 20 seconds playing a video game followed by a 40 second wait to a 40 second wait followed by 20 seconds playing a video game, but that this preference shrank when a 60 second front-end delay was added.

Assuming that these past results are indicative of a true underlying present bias in the average populations, it may by that our experiment differed from past experiments in ways that undermined present bias. For example, the experiments of D. Navarick (1998) and Millar and D. J. Navarick (1984) were conducted in a dark room, which might have cut down on external distractions, while ours were not. The experiments of Solnick et al. (1980) used louder noises than ours, and were conducted with a distractor task of solving math problems which might have caused people to make their choices with more impulsivity and less cognitive reflection. We hoped that our combination of multiple stimuli (videos, noises and sliders) would overcome any weaknesses in the implementation of any one, but this may not have been the case. Additionally, the setup of Experiment 3 could have lead some participants to purposely adopt negative discount rates, saving the positive experience for last, as has been found in some past experiments (Loewenstein, 1987).

It is also worth considering that these older experiments might not meet current statistical standards, and that the motivational effectiveness of these sorts of consumable rewards should be reconsidered. D. Navarick (1998), in fact, did not report present bias at the population level, and focused his analyses on individuals. Solnick et al. (1980) and Millar and D. J. Navarick (1984) did make claims of present bias over short time scales on a group level, but used between-participant designs with quite small groups of 10 to 15 people per condition. Their results in many cases appear strong qualitatively but the statistics calculated are not clearly reported. More recent tests of present bias with consumable rewards also reveal some statistical weaknesses; in their experiments with soda as a reward, Brown et al. (2009) first ran 44 participants and then increased their sample size to 55 after reviewing the results, and also found an effect of delaying rewards (by 10 minutes) only at the $p < .1$ significance level (both of which are clearly acknowledged in their paper).

Unfortunately, as discussed by Brown et al. (2009), collecting a large amount of decision-making data with immediately consumable rewards is highly time consuming, because each consumption event takes time and trials must be temporally distributed to induce temporal discounting. However, given our findings in the current study, and the limitations of past studies, we would strongly recommend that future researchers endeavor to replicate the finding that present bias can be induced for videos, noises, or any stimuli of interest before attempting to use those stimuli for novel research questions.

2.7.2 Relating Exploratory and Intertemporal Choice

While our results in Experiments 1 and 2 do not provide strong evidence against present bias influencing exploratory choice, it is still worth considering ways in which exploratory choice may, in fact, differ from standard intertemporal decision making. In some situations, exploratory decisions can look very much like other intertemporal decisions, and thus it would be highly surprising if the same qualities of temporal discounting were not involved. This would be expected particularly in situations where an extended bout of exploration is very likely to produce greater long term rewards. For example, going to college and selecting classes could be considered a series of exploratory choice among a variety of life paths. In this case, many students are likely highly confident that through this exploration they will find a life path more rewarding than those available without it, making college a more straightforward tradeoff between up-front costs and long-term benefits (Stange, 2012).

But while some exploratory choices are in practice very much like intertemporal choices, and all exploratory choices *in theory* have aspects of intertemporal choice, there are also important differences. Intertemporal choices, as they appear in the lab and (sometimes) in real life, present a clear choice between rewards now and later. In exploratory choice, the tradeoff between the present and future is implicit, with exploring leading to a decreased probability of high reward immediately and a greater probability of high reward in later choices. While past work shows that people are able to consider these reward tradeoffs in some situations (Wilson et al., 2014; Rich & Todd M. Gureckis, 2017), doing so might not be natural or easy in all situations. Even in standard intertemporal choice tasks, the effect of time delays on decision making seems to be fragile and sensitive to contextual effects (Ebert & Prelec, 2007; Lempert & Phelps, 2015).

Instead of treating exploratory choice like intertemporal choice, people may rely on other motivational and cognitive factors to balance exploration and exploitation. Curiosity acts as an innate drive towards information-seeking (Berlyne, 1966; Loewenstein, 1994; Kidd & Hayden, 2015), and some researchers speculate that curiousity may in fact have evolved to induce exploration (Singh, Barto, & Chentanez, 2004; Oudeyer & Kaplan, 2009). If exploration is inherently rewarding due to its potential to reveal information, then its rewards are moved from the future to the present and temporal advantage of exploitation is removed. Even in situations where people likely have low intrinsic curiousity about outcomes, people still appear to have strategies for choosing when to explore that are based on heuristics, exploration "bonuses," or added decision noise, rather than a full consideration of the future (Daw et al., 2006; Speekenbrink & Konstantinidis, 2015; Wilson et al., 2014). To the extent that people use these strategies, they may not change their behavior based on the timing of rewards, even if doing so would be in line with their preferences.

The considerations of this discussion suggest that continuing to study exploratory choice in the lab, in order to show that present bias simply does or does not exist, may not be most fruitful path for future research. A more rewarding route may be to further consider the range of exploratory choices in their true, naturalistic settings, such as daily shopping habits (Riefer et al., 2017), and seek to understand when exploratory choice is treated similarly to intertemporal choice,and when it is not. While field data do not allow the experimental control available in the lab, they would allow researchers to collect data with significant temporal spans and in situations where major motivational forces, from present bias to curiosity, are at play.

3

The limits of learning: Exploration, generalization, and the development of learning traps [1]

Conventional wisdom holds that learning from experience brings us closer to understanding the world around us. For instance, with enough practice we can learn even complex skills like playing an instrument or learning a new language. However, in many situations learning new things trades off against the goal of maximizing reward. Our desire to try every restaurant in a city has to be balanced against our desire to have a good meal tonight. This trade-off, known as the *exploration–exploitation* dilemma, has been extensively studied in computer science (Sutton & Barto, 1998), organizational behavior (March, 1991), and psychology (Mehlhorn et al., 2015). The central premise of this paper is that in these situations even adaptive learning can lead people (as well as other agents such as organizations, animals, and machines) to form robust false beliefs about the world. For instance, consider a child who is exploring different foods, and finds that she dislikes spinach and cucumbers. As a result of this experience, she

[1] Published as: Rich, A.S. & Gureckis, T.M. (*in press*). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*.

may falsely believe she dislikes all vegetables and avoid vegetables in the future, never learning that she really likes broccoli even as she continues to explore other foods.

Inspired by the organizational-learning theorists Levinthal and March (1993), we call the development of these types of false or incomplete beliefs "learning traps" (see also, Erev, 2014; Teodorescu & Erev, 2014b). We use this term because the failure of beliefs to move towards truth in these situations is not caused by ignorance, but instead by the interaction of exploratory choice with what is already believed. Learning traps represent a very general class of phenomena that reoccur in many different situations for many types of learners.

In this paper, we aim to provide a comprehensive summary of known learning traps and the factors that generate them. We begin by describing a variety of disparate findings that can be interpreted as reflecting learning traps, in fields from organizational theory (Denrell, 2007) to social psychology (Russell H Fazio et al., 2004) and domains from foraging behavior (Niv, Joel, et al., 2002) to stereotype formation (Denrell, 2005) and nepotism (C. Liu et al., 2015). Our synthesis shows that a crucial ingredient for the formation of learning traps is generalization from past experience that "traps" people into premature exploitative behaviors and limits further learning and belief revision. Within this review, we progress from traps caused by simple forms of generalization to traps caused by more complex forms of generalization, and provide a unifying framing in which learning traps represent intrinsic limits of learning and adaptation.

In the empirical portion of the paper we present a computational model and a series of simulations and studies with which we explored one particular learning trap. This learning trap can form when people selectively attend to a subset of relevant stimulus attributes while learning, which is interesting because selective attention is most often considered a facilitative mechanism for learning (e.g., improving speed of acquisition and performance by filtering out irrelevant information). As a test of the robustness of this attentional learning trap, we explored a variety of manipulations aimed at reducing the prevalence of this trap for both simulated and human learners. To foreshadow, our results confirm the robustness and prevalence of learning traps. We conclude by discussing additional routes to the prevention of learning traps and highlighting other unstudied settings in which they are likely to emerge.

3.1 Trapped by Stochasticity (i.e., Noise)

The simplest learning traps arise in situations where a person (or other agent) repeatedly encounters a prospect and must decide whether or not to interact with it and experience a consequential outcome (e.g., approach–avoidance). A prospect might, for example, be another person, with the outcome of interaction being a pleasant or unpleasant conversation. A reasonable assumption in these cases is that the prospect will be stable over time, and that past experiences with the prospect will generalize well to future experiences. When this assumption holds, then correct knowledge about the whether the prospect is good or bad can be gained from the first encounter. If instead there is stochasticity in the outcomes, then there are times when generalization from early encounters will be misleading, and learning traps can develop. (Note that for the purposes of this paper stochasticity means random, unsystematic noise over time or across instances. Stochasticity that causes systematic changes over time, as described for example by Behrens, Woolrich, Walton, and Rushworth (2007), could be the source of additional learning traps but will not be directly addressed.)

Perhaps the paradigmatic example of this type of learning trap is the “hot stove effect”, introduced by Denrell and March (2001). If a prospect is positive on average, but stochastic, early experiences may make the prospect appear negative. If a person learns from and generalizes from these early experiences, they will begin to avoid the prospect. Once this has occurred, learning halts because no more information is gained about the prospect, and the person is left with a false negative belief that is difficult to overcome. Avoiding the option appears to be the safest strategy, but it is exactly the strategy that prevents learning the true structure of the environment. Notably, this trap does not occur when a prospect is usually negative but mistakenly believed to be positive. In this case, the false belief causes continued interaction, which eventually allows the belief to be corrected.

The hot stove effect is intrinsic to experiential, reward-driven learning in a variable environment, and is exhibited even by an optimal agent (Le Mens & Denrell, 2011), though it can be mitigated through a variety of factors including more persistent exploration (Rich & Todd M. Gureckis, 2017). In addition, the effect can explain under-exploration and apparent risk aversion across a broad range of experience-based decision-making domains, ranging from the decisions of firms to the foraging behavior of bees (Denrell & March, 2001; Niv, Joel, et al., 2002). It also provides one explanation for why people tend to prefer novel prospects over previously-experienced ones (Le Mens, Kareev, & Avrahami, 2016), as well as for why people prefer ingroup-members (with whom they must interact) over outgroup-members (with whom they repeatedly choose whether or not to interact) (Denrell, 2005; C. Liu et al., 2015).

A second learning trap that develops from the expectation of stable outcomes is the underweighting of rare events. This learning trap was identified by researchers in the decisions from experience literature (Hertwig et al., 2004), using a sampling paradigm in which participants repeatedly tested several prospects, paying a cost of time and sometimes money for each sample (Juni et al., 2016), before selecting one for a consequential choice. As participants explore each alternative they learn the common outcomes well and become confident enough to make a final, exploitative choice, but their samples often fail to include rare outcomes. Thus, people's beliefs about prospect values when making their final decisions systematically underweight these rare outcomes. This underweighting of rare events may also occur in settings where all choices are consequential, both in the lab (Teodorescu & Erev, 2014b) and for real-world rare events ranging from driving accidents (Fuller, 1991) to nuclear disasters and scientific discoveries (Levinthal & March, 1993). Due to its interaction with the hot stove effect, the underweighing of rare events is likely more persistent when the rare event is positive, and the prospect may thus be mistakenly avoided. When the rare event is negative, the decision maker is likely to continue sampling the prospect and discover it in the long run, though the result of this may be disastrous. For example, failure to experience the rare event of discovery may drive someone out of science permanently, while failure to experience the rare event of a car accident may lead someone to continue driving without a seat belt until the event is experienced.

3.2 Trapped by Similarity

In environments with not just one prospect but several, new learning traps become possible as patterns of generalization become more complex. Rather than just determining how much to generalize from the past to the future, an agent must determine how much to generalize from experience with one prospect to experience with another.

One answer for how to generalize across prospects is based on their similarity. If two prospects are highly similar, for example sharing many perceptual features, then it may be safe to assume that the outcomes of interacting with them are similar as well (Tversky, 1977; Shepard, 1987; Tenenbaum & Griffiths, 2001). This notion of similarity is a core aspect of many theories of human category learning and categorization. Studies of natural and artificial categories show that a novel item is judged to be a typical category member to the extent that it is similar to other members of the category and dissimilar to members of different categories (Rosch & Mervis, 1975). In addition, many formal models of categorization determine the membership of an item through similarity-based comparison to past category members or to a more abstract category representations (Nosofsky, 1986; Love, Douglas L. Medin, & Gureckis, 2004).

The expectation that similar prospects will yield similar outcomes, like the expectation that prospects are stable over time, is a reasonable one. However, in value-based decision making with choice-contingent feedback, similarity-based generalization may cause the value of prospects near the boundary between good and bad outcomes to be permanently misestimated. This is particularly likely to be the case for positive prospects that are similar to negative ones. When a novel prospect is similar to other prospects with which an agent has had negative experiences, the decision maker must choose to either approach it to learn about its value, or to avoid it, exploiting what it has learned from past experience. If the prospect is avoided, this creates a learning trap in which the decision maker never learns the true value of positive prospects that fall "in the shadow" of similar negative prospects that were experienced earlier.

Russell H Fazio et al. (2004) confirmed the existence of this trap in a series of experiments investigating how people learn to approach or avoid in an environment where multiple distinct prospects varied along two continuous dimensions. (Prospects were "beans" that varied in shape and number of speckles.) They found a pronounced learning asymmetry, with participants correctly avoiding negative prospects, but often incorrectly avoiding positive prospects as well. When participants were given full (foregone) information, and told the value of prospects they avoided, the asymmetry disappeared, supporting the idea that they had entered a learning trap in which their behavior prevented further experience and learning. A related simulation study found that the same effect emerged in a connectionist network that generalized prospect values based on similarity (Eiser, Fazio, Stafford, & Prescott, 2003).

While this learning trap was first analyzed as a way of understanding the development of social attitudes, it is likely to emerge in many of the situations in which the hot stove effect has been observed. To take one example, Niv, Joel, et al. (2002) found that simulated bees came to avoid flowers that produced a variable amount of nectar, even if they produced more nectar on average than stable flowers. One would predict that bees would also avoid flowers of a species that looks similar to a low-nectar species, even if the species itself produces a large, stable amount of nectar.

3.3 Trapped by Selective Attention

For the bulk of this paper we will focus on a relatively unexplored learning trap that develops due to selective attention, one of the most basic forms of rapid generalization that extend beyond similarity. A classic finding in category learning is that the ease of learning a category structure depends not only on the similarity of the exemplars within and between categories, but on the number of dimensions required to distinguish categories (Shepard, Hovland, & Jenkins, 1961; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). Subsequent theories of categorization have posited that people adapt their allocation of attention to optimize performance, selectively attending those dimension that discriminate categories while ignoring those that are irrelevant (Nosofsky, 1986; Kruschke, 1992). Thus, upon encountering an item, an agent with selective attention will weigh past experiences with items that are similar on attended dimensions strongly, while weighing experiences with items that are similar on non-attended dimensions weakly or not at all. Recent work in reinforcement learning has shown that selective attention plays a role in value based decision making as well (Niv, Daniel, et al., 2015).

While selective attention often aids learning, it creates a bias toward attention to fewer dimensions. That is, a learner with selective attention will initially expect there to be few dimensions of a prospect relevant to its outcome, and will only be induced to believe there are additional relevant dimension with more data. This bias means

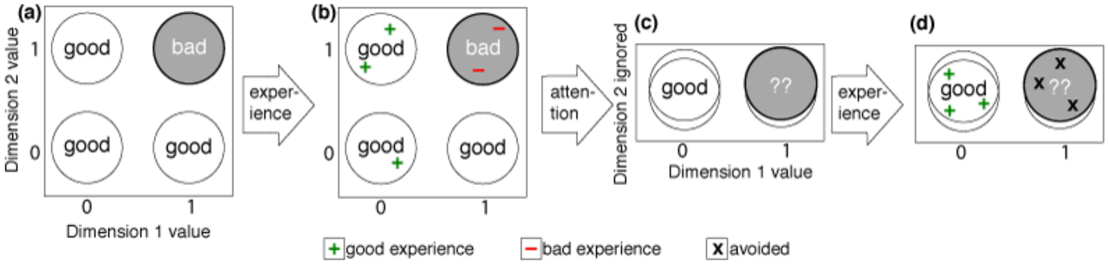


Figure 3.1: *a:* A deterministic environment containing prospects with two binary features, where only prospects possessing both features are negative. By attending to both features, a decision maker can avoid all negative prospects while exploiting all positive prospects. *b:* Early experience happens to highlight the relevance of Feature 1. *c:* The agent begins to attend only Feature 1, and to ignore Feature 2, making items with and without Feature 2 appear equivalent. Items without Feature 1 are positive, while the value of items with Feature 1 appear stochastic. *d:* The agent now avoids items with Feature 1, since some are negative. This prevents the agent from gaining feedback that would cause it to change its behavior.

there are cases where selective attention can inhibit later learning, particularly when a person comes to ignore a dimension that is later useful. This can occur in blocking and backwards blocking, phenomena in which associating a single cue with an outcome prevents learning about other, concurrently presented cues (Mackintosh, 1975; Kruschke & Blair, 2000). It also happens in cases where a person learns first one category structure and then a second in which a previously-irrelevant dimension must be attended (Kruschke, 1996; Hoffman & Rehder, 2010).

The bias towards categories with few relevant dimensions can also lead to a distinct learning trap. For example, consider the schematic diagram in Figure 3.1. Here, a hypothetical agent repeatedly encounters prospects that vary on two binary dimensions. These dimensions jointly determine the prospect quality in a somewhat complex manner, with prospects that have a value of 1 on both dimensions yielding negative outcomes and all other prospects yielding positive outcomes [2]. Prospects that are approached yield a deterministic positive or negative outcome, while those that are avoided provide no information. The agent must explore the environment by approaching uncertain prospects to learn about them, but also must eventually exploit what it has learned and avoid prospects it expects to be negative if it wishes maximize positive outcomes.

In this setting, an agent with selective attention will attempt to learn not just which encountered prospects are positive and negative, but also which dimensions are useful and worth attending when deciding what to approach and what to avoid. As the agent gains experience, it will adjust its allocation of attention to optimize its performance based on its observations.

The true structure of the environment incentivizes attention to both dimensions. Despite this, variation in the prospects the agent happens to explore early on may cause one dimension to be attended more than the other and believed to have a stronger effect on prospect outcome. If this tendency is strong enough, the agent may begin to exploit its perceived knowledge and act based on that dimension alone, approaching prospects that have a value of 0 on that dimension, and avoiding those with a value of 1.

Once learned attention influences behavior in this way, the agent has entered a learning trap. All prospects with a value of 1 on the attended dimension are avoided, including those that have a value of 0 on the unattended dimension, with which the agent has had no negative experiences. The bias away from these prospects is persistent, since the agent avoids all prospects that would provide evidence of the importance of the sec-

[2] While we assume a interactive, deterministic relationship between the dimensions and the binary outcome, a similar structure can occur if two dimensions have independent, logistic relationships to the outcome (e.g., $p(\text{negative}) = 1/(1 + \exp(-3(2d_1 + 2d_2 - 3)))$).

ond dimension. The agent may avoid a positive region of the environment indefinitely, and may also consequently hold false beliefs about how the environment is divided into meaningful categories.

3.4 Trapped by Simplicity: A Unifying View

In the following section, we will use a version of ALCOVE (Kruschke, 1992), a well-known model of human category learning with a selective attention mechanism, to explore the conditions under which an attentional learning trap can develop. However, we hypothesize that other models of categorization, including models based on very different principles than ALCOVE and lacking explicit selective attention mechanisms, would likely fall into a similar trap for slightly different computational reasons.

For example, we expect that rule-based models of categorization, such as the rule-plus-exception model (Nosofsky, Palmeri, & McKinley, 1994) and the rational rules model (Goodman, Tenenbaum, Feldman, & Griffiths, 2008), would fall into the "attentional" learning trap quickly and easily, because they prefer simple, one-dimensional rules. In the category structure described in Figure 3.1, either model would be likely to form a rule to respond based on dimension 1 or dimension 2 exclusively. While both models will adopt more complex rules if experience warrants, the use of a unidimensional rule would prevent the proper feedback from being obtained if feedback is choice-dependent. It has long been appreciated that selective attention and rule-based categorization are related phenomena (Kruschke, 1992).

Clustering models such as SUSTAIN (Love, Douglas L. Medin, & Gureckis, 2004) and the rational model (Anderson, 1991) could also fall into the trap. Both models tend towards simple category representations with few clusters, and could become trapped if they attempted to form a two-cluster representation of the environment in Figure 3.1 and assigned a region of positive items to the negative cluster. Again, this mistake would not be corrected because prospects in the negative cluster would be avoided, preventing the model from learning that its representation was incorrect and that another cluster should be recruited. (In the case of SUSTAIN, the trap would be exacerbated by an attentional tuning mechanism similar to ALCOVE's.)

The common thread joining these disparate models and making them susceptible to learning traps is their preference for simplicity. Thus, what we call an "attentional" learning trap could also be considered more generally as a "simplicity" learning trap; an agent comes to believe the environment has a simpler structure than it truly does, which prevents it from exploring further to come to a more complex, more correct belief. Other learning traps, including those based on stochasticity and similarity, could be derived from a simplicity-based bias as well. An environment that is stable is simpler than one that changes, and an environment where similar prospects produce similar outcomes is simpler than one where they do not. Indeed, many aspects of cognition seem to have an inductive bias towards simple inferences and generalizations (Chater & Vitányi, 2003; Feldman, 2003).

As researchers in machine learning have shown, some degree of inductive bias is necessary for learning to proceed—without it, a huge number of generalizations will be equally plausible from any finite set of experiences (Mitchell, 1980; Geman, Bienenstock, & Doursat, 1992). Thus the tendency to believe the world is simple may be a necessary component for effective learning, and one that leads to better performance in many environments even as it inevitably leads to worse performance in others (Wolpert, 1996). This differentiates learning traps from other phenomena such as confirmation bias that prevent belief revision (Nickerson, 1998). Confirmation bias is generally (though not always, see Navarro and Perfors (2011)) thought of as reflecting a flaw in inference and reasoning, but learning traps, in their most basic form, are not a true suboptimality. Rather, they are an unavoidable byproduct of learning—whether by people, machines, animals, or organizations—in situations where the need for additional information must be traded off with the need for reward maximization.

3.5 Computational Explorations of the Attentional Learning Trap

Our discussion so far has highlighted intuitive reasons why learning traps form as well as cited empirical evidence that humans are susceptible to some these traps (e.g., the hot stove effect). However, it is helpful to explore this issue in slightly more detail through computer simulation. Computer simulations allow us to conduct counterfactual analyses that help isolate necessary and sufficient aspects of the learner–environment interaction that provoke learning traps. They also help us to set the stage for our later empirical studies. We particularly focus our work on the novel attentional learning trap introduced above.

3.5.1 ALCOVE-RL: A Computational Model of Learning and Generalization from Experience

While the attentional learning trap could be fruitfully analyzed from multiple perspectives, we approach it from the angle of selective attention, using a modified version of the ALCOVE (Kruschke, 1992) model of categorization, for several reasons. ALCOVE is a successful and widely used model of human category learning (see e.g., Nosofsky, Gluck, et al., 1994), and can be conveniently modified for use in a reinforcement-learning context (see Jones & Cañas, 2010). Additionally, ALCOVE shares deep formal structure with the simple adaptive learning models used to study stochasticity-based learning traps (Denrell & March, 2001) and the connectionist models used to study similarity-based learning traps (Eiser et al., 2003). Like these models, ALCOVE learns in an incremental, mechanistic manner to reduce error (or increase reward). This makes ALCOVE an interesting model in which to demonstrate a learning trap, because unlike models that directly aim for simple representations, its biases are implicit. Just as the update rules of adaptive learning models imply a "built in" assumption that the environment is fairly stable, the attentional learning rules of ALCOVE may imply a built in assumption that there are few relevant dimensions, which will cause a learning trap when this assumption is incorrect and feedback is choice-dependent.

ALCOVE was originally designed for learning in supervised category-learning tasks with corrective feedback. To use it to study learning traps, we created a modified model, ALCOVE-RL, for use in a reinforcement-learning task where feedback is continuous and action-dependent, following the work of Jones and Cañas (2010).

As in the original ALCOVE model, when a new prospect is presented to our model it first activates the set of input nodes a^{in} based on its value on each dimension. This activation then spreads to a set of hidden nodes a^{hid} representing the space of exemplars (i.e., possible prospects). Each hidden node a_j^{hid} has a position h_j in the psychological space defined by the input dimensions, and is activated based on its similarity to the input node values.

$$a_j^{hid} = \exp[-c(\sum_i \alpha_i |h_{ji} - a_i^{in}|)]$$

The similarity function is parameterized by a specificity constant c, as well as learnable attention parameters α determining the breadth of generalization along each dimension. Essentially, internal representations of past prospects are activated based on their similarity to the current input, with similarity counting more on dimensions with high attention settings than on those with low ones.

The activation of the hidden nodes then spreads to output nodes a^{out}, which are each associated with an action. As in ALCOVE the output activation of a_k^{out} is determined by first taking the sum of the a^{hid} activations, weighted by the learnable parameters w_k. The w_k parameters represent the model's association of each hidden node with a reward value. The summed outputs are then divided by the sum of a^{hid} activations to produce a weighted average. This normalization is not necessary in ALCOVE for category learning, where the only goal is for the correct output node to have the highest activation. But in reinforcement learning, normalization ensures that output activation reflects predicted reward and does not over-shoot or under-shoot this value due to overall activation strength.

$$a_k^{out} = \sum_j w_{kj} a_j^{hid} / \sum_j a_j^{hid}$$

Following ALCOVE and classic models of decision making (Luce, 1959), the model's choice of action then follows a probabilistic choice rule, with its degree of determinism controlled by the parameter ϕ.

$$Pr(K) = \exp(\phi a_K^{out}) / \sum_k \exp(\phi a_k^{out})$$

The feedback mechanism in ALCOVE-RL again diverges from the original ALCOVE implementation. In ALCOVE, feedback is received for all outputs, and there is no penalty for outputs with greater magnitude than the correct values. In our model, upon receiving a reward r, the feedback t_k to the model is

$$t_k = \begin{cases} r & \text{if } k \text{ is the chosen action} \\ a_k^{out} & \text{otherwise} \end{cases}$$

In other words the model is told that the predicted reward for the chosen action should move towards r, and that no change is needed for the predictions on other actions since no information about their outcomes was gained.

Given this action-dependent feedback, the model then updates its exemplar-action weights and attention weights through the same mechanisms at ALCOVE in order to decrease future error.

$$\Delta w_{kj}^{out} = \lambda_w (t_k - a_k^{out}) a_j^{hid}$$

$$\Delta \alpha_i = -\lambda_\alpha \sum_j [\sum_k (t_k - a_k^{out}) w_{kj}] a_j^{hid} c |h_{ji} - a_i^{in}|$$

3.5.2 Exploring the Determinants of the Attentional Learning Trap through Simulation

The computational ALCOVE-RL model makes it possible to test through simulation whether the attentional learning trap described above can emerge through learning with choice-contingent feedback, and whether ALCOVE's selective attention mechanism is in fact crucial for the trap to develop. Successful production of a learning trap through simulation will set the stage for studies with human participants.

To explore the behavior of ALCOVE-RL across a range of conditions, we devised a task similar to Figure 3.1. In particular, we presented the model with a four-feature category learning problem, where approaching prospects (i.e., exemplars) with both Features 1 and 2 yielded a payoff of -4 and approaching any other prospect yielded a payoff of 1. This environment matches exactly the structure depicted in Figure 3.1, but with two added irrelevant features. Each run of the model lasted for a 20 block learning phase of 16 trials each. Within each block all prospects were observed in a pseudo-random order, with the condition that each sub-block of eight trials included two negative prospects and six positive prospects. The learning phase was followed by a final test block in which ALCOVE-RL made choices but underwent no learning updates.

We simulated five distinct conditions in order to understand what properties of model and of the environment were sufficient to produce an attentional learning trap. The <*contingent, att*> condition instantiated the kind of agent and environment described in our introduction to the learning trap. In this condition, the model's attentional learning capability was active, and the model only received feedback on the value of a prospect when it approached. In the <*full-info, att*> condition, we tested whether partial feedback was critical to the learning trap by modifying the environment so that the model received feedback on the value of all prospects regardless of its choice (i.e., foregone rewards, see Love & A. R. Otto, 2010). The <*contingent, no-att*> and <*full-info, no-att*> conditions tested the role of selective attention in the learning trap. These conditions mirrored the first two, but with the attention to all dimensions yoked to remain equal. Finally, in the <*random-info, att*> condition we tested whether a learning trap might be produced not because an agent receives *choice-contingent* feedback but simply because it receives *less* feedback. In this condition the model received feedback on 50% of trials, but the feedback trials were randomly selected and independent of the model's choices.

Each condition was simulated for 1000 model runs. All simulations were run with a specificity constant $c = 6$, a temperature parameter $\phi = 15$, an output-weight learning rate $\lambda_w = 0.1$, and an attention learning rate $\lambda_\alpha = 0.1$. These parameters were selected to create an agent with a high propensity to fall into learning traps. In particular, the high λ_w and λ_α cause the agent to form beliefs quickly, while the high ϕ causes the agent to act fairly deterministically based on those beliefs, reducing the amount of corrective feedback it can receive. Thus, the results of the simulation should not be interpreted as a

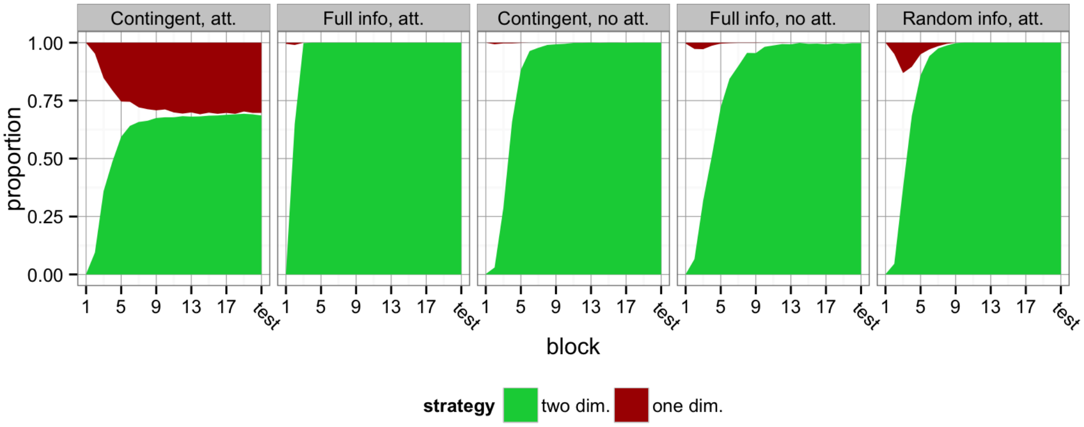


Figure 3.2: ALCOVE-RL model simulations of approach–avoid decision making in five attentional/informational conditions. Panels show the proportion of model runs adopting the correct two-dimensional strategy or one of the one-dimensional learning traps in each of the 20 learning blocks and the test block.

claim about the degree to which ALCOVE-RL will fall into an attentional learning trap under all possible parameter settings, but rather as an exploration of how one particular model with one parameter setting behaves as we alter the feedback from the environment and its ability to deploy selective attention.

3.5.3 Results

Figure 3.2 shows a breakdown of ALCOVE-RL's behavior over 20 blocks of experiential learning and the final test block. Models in each condition were classified based on whether or not they perfectly followed the correct two-dimensional rule or one of the two one-dimensional rules indicative of an attentional learning trap.

Over a quarter of agents in the <*contingent, att*> condition fell into the learning trap, with their behaviors influenced by only one of the relevant dimensions. In all other conditions, there was little tendency to learn an incorrect, unidimensional rule, and the correct rule tended to be learned quickly even in cases where an incorrect pattern of behavior was temporarily adopted. Providing full feedback in the <*full-info, att*> condition allowed for very swift learning, while maintaining contingent feedback but removing selective attention in the <*contingent, no-att*> condition lead to accurate learning as well. Interestingly, when given full feedback the model learned more quickly with selective attention than without it, showing that selective attention is only detrimental to learning this category structure when feedback is contingent.

It is also not the case that the bias of the <*contingent, att*> condition was caused by a general lack of information. In the <*random-info, att*> condition, where the model received feedback on half of the trials, randomly chosen, learning proceeded quickly and without bias. Thus, the attentional learning trap occurred due not to an overall poverty of information but to a specific pattern of behavior that prevented information from being gained about prospects that could correct the model's misallocated attention.

Finally, it is worth noting that in our simulations thus far we have assumed that nothing is encoded in the absence of feedback. As we have shown, learning traps can develop simply through the interplay of generalization and exploratory choice. However, data from recent experiments—among the few to study category learning with choice-contingent feedback—have suggested that when people do not experience feedback from a prospect they employ *constructivist* coding. In constructivist coding, rather than learning nothing upon avoiding an exemplar, the learning agent updates its beliefs as though it had approached the prospect and received the expected or predicted (likely negative) outcome, reinforcing existing beliefs with additional learning (Elwin, Juslin, Olsson, & Enkvist, 2007; Henriksson, Elwin, & Juslin, 2010). With this coding scheme, we would expect learning traps to become more intense, as the negative belief that lead a prospect to be avoided would be strengthened with each repeated avoidance. To test this we simulated a version of ALCOVE-RL in which upon avoiding a prospect the model updated its beliefs as though it had approached and received a small negative outcome of -0.5. The results of this <*constructivist, att*> condition, plotted against the <*contingent, att*> condition in Figure 3.3, confirm that even a small amount of constructivist coding leads to a more pronounced learning trap. (A stronger constructivist coding scheme in which avoided items are assumed to have produced the full -4 negative outcome causes the model to avoid all prospects entirely.) To preview our experimental results, the more severe learning trap caused by constructivist coding is in line with the behavior we observed among our human participants.

3.6 Experiment 1

Our simulations verify the pernicious nature of learning traps and the necessary aspects of the learner-environment interaction that encourage their development. To test the degree to which people are susceptible to the attentional learning trap, we performed a simple experiment similar to the category-learning task described above [3]. Given the

[3] Experiment and analysis code for all experiments is available at https://github.com/NYUCCL/LearningTrap. Data from all experiments is available at https://osf.io/hrb3u/?view_only=d374ae90615d4e5993ee1c502493673d. All experiments were approved by the NYU Institutional Review Board (IRB-FY2016-231 - Active Learning in Dynamic Task Environments).

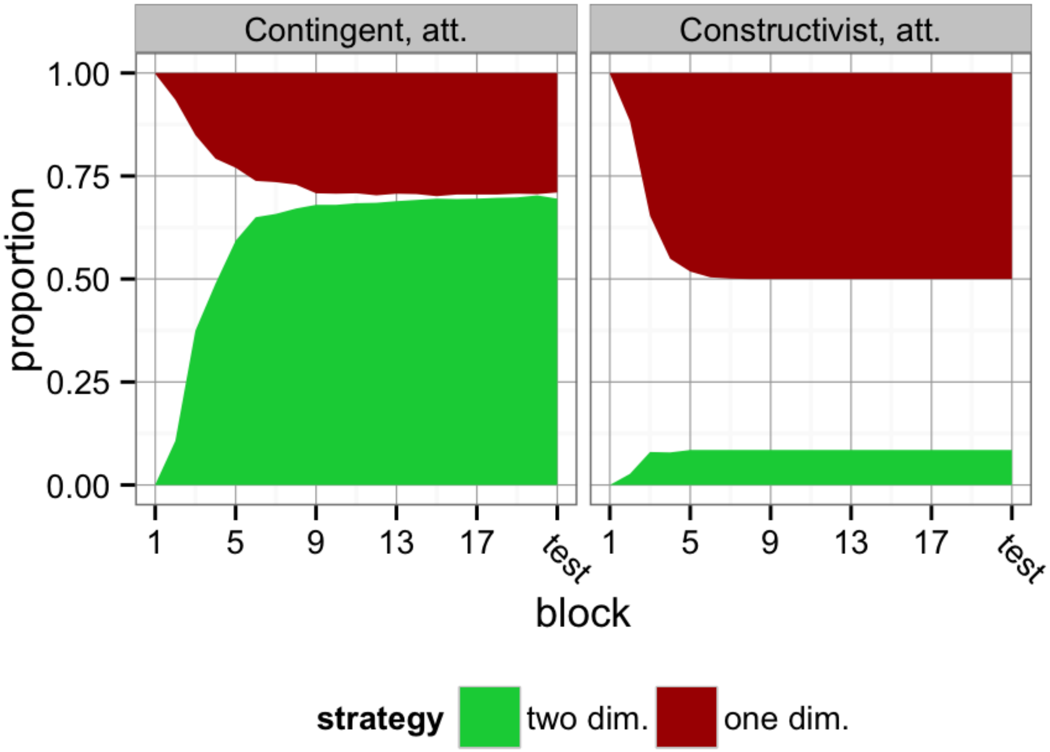


Figure 3.3: ALCOVE-RL model simulations with constructivist coding in the absence of feedback, compared with the standard contingent-information condition. Panels show the proportion of model runs adopting the correct two-dimensional strategy or one of the one-dimensional learning traps in each of the 20 learning blocks and the test block. Compared with the standard model, constructivist coding causes the model to fall into the attentional learning trap far more often.

broad support for ALCOVE in the category learning literature we expected that a similar learning trap would affect human learners. However, it is possible that human learners have more sophisticated learning strategies that help them avoid these types of costly errors when adapting their behavior to a task environment.

3.6.1 Method

Participants

One hundred one participants (48 female; 53 male) were recruited via Amazon Mechanical Turk. Preliminary power calculations indicated that 50 participants per condition would allow us to consistently detect a difference of 30% or greater of participants falling into the learning trap. Participants received $1.25 for participation and received a performance-based bonus that ranged up to $1.80. Two participants indicated that they used an external memory device (e.g., pen and paper) during the task, and four required more than two attempts to pass a post-instructions quiz. These participants were excluded from further analyses.

Stimuli

Stimuli were computer-generated cartoon bees that varied on four binary dimensions; they had two or six legs, a striped or spotted body, single or double wings, and antennae or no antennae, for a total of 16 unique stimuli. Example stimuli are shown in Figure 3.4. Two of the four dimensions were chosen as relevant, counterbalanced across participants. Of the four possible combinations of values on these two dimensions, one was chosen at random; stimuli with this combination of values were “dangerous”, and the

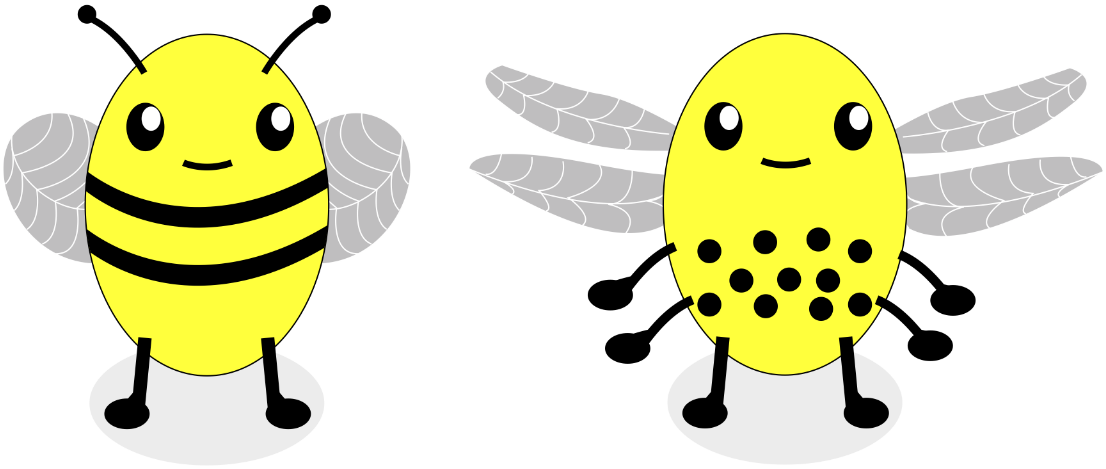


Figure 3.4: Examples of the stimuli used in Experiment 1, with opposite values on all four binary dimensions.

remaining stimuli were “friendly.”

Procedure

The experiment resembled a standard category-learning paradigm (e.g., Nosofsky, Gluck, et al., 1994), but with an added component of approach–avoid decision making. Participants played the role of a beekeeper collecting honey from several beehives. They were told that each hive contained a single variety of bees, and that while most hives contained friendly bees that would give them honey, some hives had been invaded by dangerous bees that would sting them if they tried to harvest.

On each trial, participants visited a new beehive, and were shown one of the bees in the hive. Based on the bee’s appearance, they then had to choose either to attempt to harvest honey from the bee variety in that hive or to avoid the hive. When participants chose to harvest, they received honey and added $0.02 to their bonus if the bee variety was friendly, but were stung and lost \$0.10 from their bonus if it was dangerous. When participants chose to avoid a hive, they gained \$0.00. Participants started the game with a bonus of \$0.40.

In the learning phase, participants encountered each of the 16 bee varieties 4 times, for a total of 64 trials. They were informed of the number of trials, and a the number of remaining trials was displayed throughout learning. While trials were not overtly separated into blocks, each of the 16 bee varieties was encountered in each block of 16 trials. Within a block, stimuli were randomized with the condition that every eight stimuli contained two dangerous and six friendly bee varieties.

Participants were split into two conditions, which differed in the feedback received upon avoiding a beehive in the learning phase. In the contingent condition, no feedback was provided when a participant avoided a hive. In the full-information condition, participants were informed of whether the bee variety was friendly or dangerous and of what their payoff *would have been* had they harvested the hive.

The learning phase was followed by a 32-trial surprise test phase. During the test phase, participants encountered each variety twice and chose to harvest or avoid hives as before, but received no feedback about the outcomes of their actions and were not able to see changes to their bonus. Stimuli were ordered using the same randomization procedure as the learning phase. This phase provided a comparison of learning under equivalent conditions.

After the test phase, participants were informed of their total bonus, and were asked two final questions: “About what percentage of beehives do you think contained dangerous bees?” and “Which features do you think were useful in deciding whether a bee variety was friendly or dangerous?". For the first question, participants entered a percentage between 0 and 100, and for the second question participants could choose any combination of the four features using check boxes.

3.6.2 Results

After exclusions, there were 48 participants in the continent-information and 47 participants in the full-information condition. Figure 3.5 shows participants' behavior over the four blocks of learning and during the test phase. While all statistical analyses use continuous measures of performance, for the purpose of visualization we categorized participants in each block based on whether they followed the correct two-dimensional rule or a one-dimensional rule indicative of a learning trap. Unlike in our model simulations, where we required perfect adherence to a rule for a model to be given a classification, we classified participants as following a given rule if 15 of 16 choices for a block adhered to the rule (30 of 32 choices for the test phase). Alternative cutoff thresholds yield qualitatively similar results.

We statistically compared the conditions using a Bayesian parameter estimation approach (Andrew Gelman, Carlin, et al., 2013). For Experiment 1, we used Bayesian equivalents of standard two-sample t and z tests. For continuous measures we modeled the data from the two conditions as being drawn from independent normal distributions with unknown mean and standard deviation. We gave the condition means weakly-informative priors of $Normal(.5, 1)$, and the condition standard deviations priors of $Normal(0, 1)$ truncated at zero. For binary measures we adopted a beta-binomial

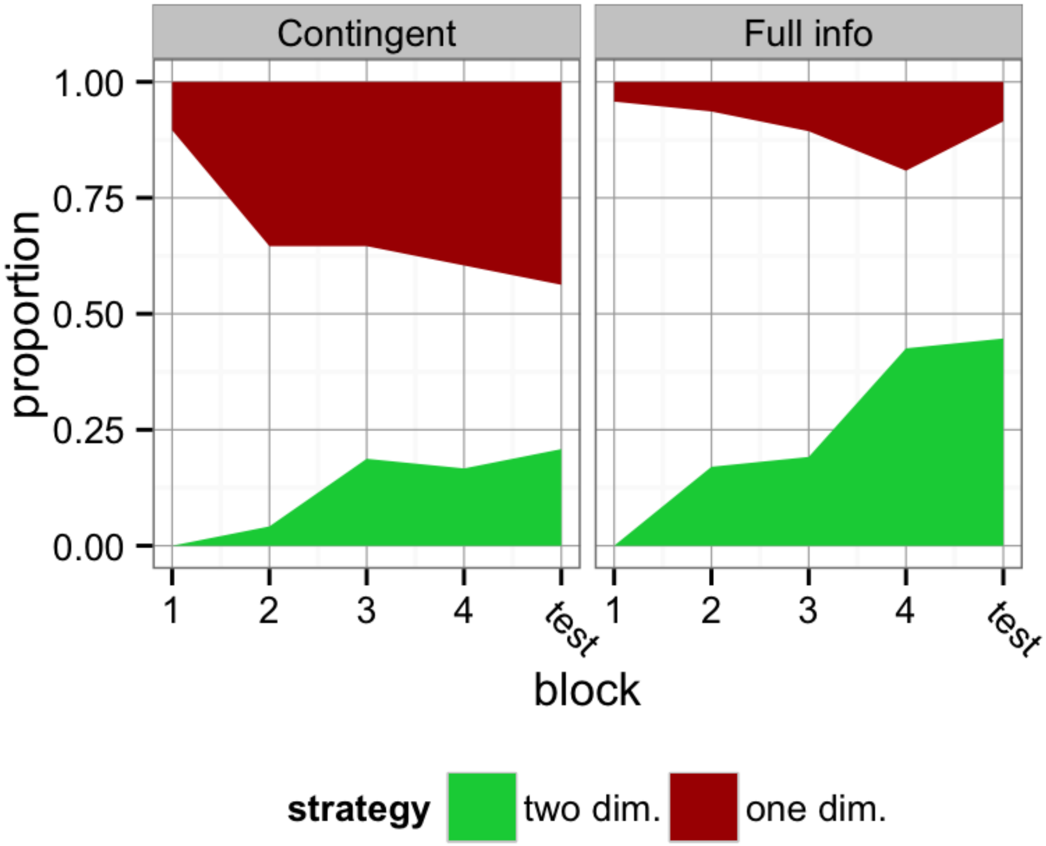


Figure 3.5: Participant behavior in the Experiment 1 beehive decision-making game. Panels show proportion of participants adopting the correct two-dimensional strategy or one of the one-dimensional learning traps in each of the 4 learning blocks and in the test phase. Participants were coded as using a two-dimensional or one-dimensional strategy if at least 15/16 of their choices in a block were consistent with that strategy.

model in which each condition's proportion was drawn independently from a $Beta(2, 2)$ prior, and the data was modeled as Bernoulli trials given the condition's proportion.

All models were implemented and fit using the Stan modeling language (Carpenter et al., 2017), which performs Bayesian inference using Hamiltonian Monte Carlo sampling. For each model we ran 4 independent chains of Monte Carlo sampling for 10,000 samples, the first 5,000 samples of which we discarded as "burn-in". Model convergence was confirmed using Stan's built-in $\hat{R}$ statistic.

For comparisons between conditions we report 95% posterior credible intervals (CI) for the difference between the conditions. Credible intervals that exclude zero can be interpreted as indicating high confidence that two conditions differ.

Early in the task, participants had little information and had to accrue experience to behave effectively later. In the first block of learning, participants in the contingent-information condition approached prospects on 75% of trials, while those the full-information condition approached only 58% of the time, $CI = [.07, .22]$. This suggests participants valued the information that was gained by approaching, in line with other recent findings that people are information-seeking in simple decision-making tasks (Speekenbrink & Konstantinidis, 2015; Rich & Todd M. Gureckis, 2017; Wilson et al., 2014).

As Figure 3.5 shows, participants increasingly adopted either the correct strategy or an inferior one-dimensional strategy as learning progressed. In the contingent-information condition, where no feedback was received about avoided prospects, participants tended towards one-dimensional strategies. In the full-information condition, they tended towards the correct two-dimensional strategy. Interestingly, the degree of dissimilarity between full-information and contingent-information behavior, and degree to which people fall into the learning trap, are both greater than in our simulations of the ALCOVE-RL model. While our explorations of ALCOVE-RL were intended to refine our understanding of the learning trap, and not to quantitatively fit human behavior, this hints that some aspects of people's behavior in the task, are not captured by the standard ALCOVE-RL model. Intriguingly, the simulations of ALCOVE-RL with constructivist coding appear qualitatively closer to people's behavior in both the current experiment and Experiment 2.

To create a numerical measure of this divergence of behavior, we calculated two behavioral "scores". The 2D score denoted the proportion of a participant's choices that were consistent with the true, two-dimensional task structure, and was equivalent to proportion correct choices. The 1D score denoted the proportion of participant's choices that were consistent with an attentional learning trap. This score was calculated by finding the proportion of responses that were consistent with each of the two one-dimensional rules that participants might form on the relevant task dimensions, and then taking the maximum over these two proportions. A value of 1 on the 2D or 1D score indicates that participants followed the respective rule perfectly, and a value of $.5$ is expected if participants behaved randomly. If participants followed a 2D rule perfectly they would receive a 1D score of $.75$, and vice versa.

The Bayesian model posteriors plotted in the upper panels of Figure 3.6 confirm the qualitative results in Figure 3.5. In the test phase, participants in the contingent-information condition had an average 1D score of .83, while those in the full-information condition had an average score of .75, $CI = [.03, .14]$. In contrast, participants in the contingent-information condition had an average 2D score of .72, while those in the

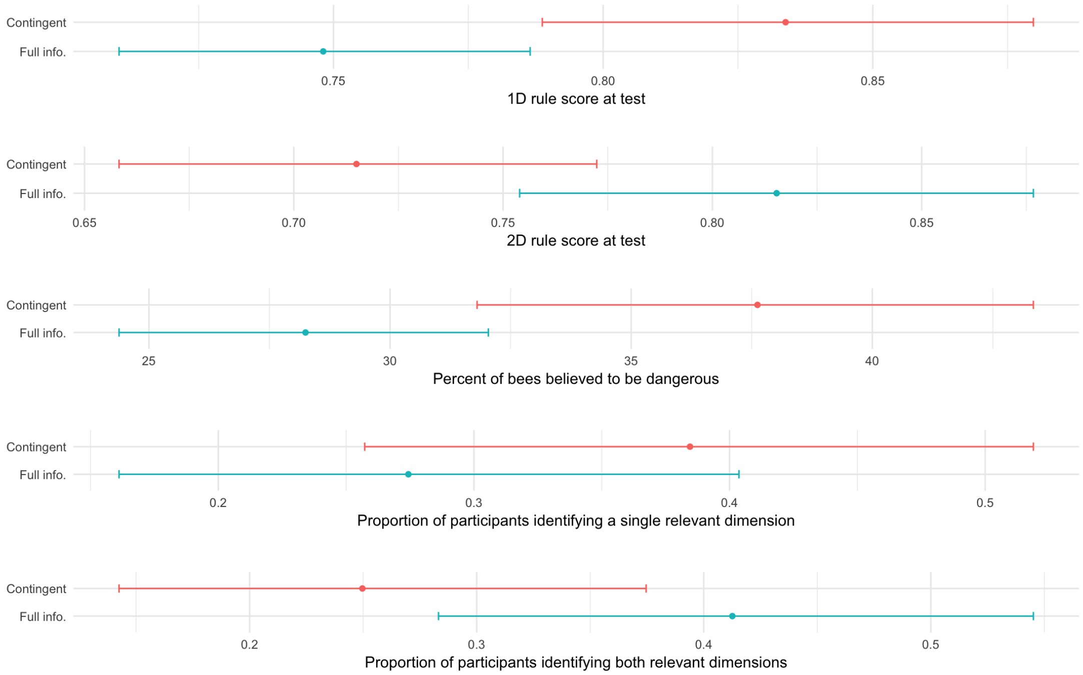


Figure 3.6: Comparisons of several measures of behavior between the contingent-information and full-information conditions of Experiment 1. Points indicate posterior population mean from Bayesian inference, and error bars indicate 95% credible intervals. All measures support the conclusion that participants with contingent information fell into the attentional learning trap more readily than those with full information.

full-information condition had a higher score of .82, $CI = [.02, .18]$.

This tendency to fall into a learning trap in the contingent-information but not full-information condition extended, albeit less clearly, to the explicit post-task questions, as seen in the lower panels of Figure 3.6. Participants in the contingent condition responded on average that 37.6% of prospects were bad (one participant was excluded for providing a negative response), while participants in the full-info condition responded that only 28.2% were bad, $CI = [.02, .16]$. The true proportion was 25%. This supports the conjecture that action-dependent feedback can affect a person's beliefs about the environment, and is consistent with the findings of Russell H Fazio et al. (2004) that approach–avoid learning leads to the belief that the environment is more negative than reality. In addition, only 22.9% of participants in the contingent-information condition identified the right combination of relevant features, while 40.4% of participants in the full-info condition did so, though the true difference is plausibly zero, $CI = [-.02, .33]$. Contingent-information participants identified only one of the relevant dimensions (and no irrelevant ones) in 37.5% of cases, while full-information participants did so only 25.5% of the time, although this difference also did not lie outside the 95% credible interval, $CI = [-.07, .29]$.

In summary, Experiment 1 provided preliminary evidence that people are susceptible to an attentional learning trap that emerges in cases of choice-contingent feedback. In the following sections we test the robustness and generality of the learning trap, while also exploring ways to prevent people from entering the trap.

3.7 Testing the Robustness of the Attentional Learning Trap

In Experiment 2, we introduced three interventions that we hypothesized might affect the severity of the attentional learning trap. Our goals in testing these interventions were twofold: first, to test whether the attentional learning trap is robust over a variety of stimulus conditions, and second, to investigate whether any of these interventions may prove effective at preventing the learning trap, and perhaps serve as a prototype for interventions in applied settings.

As described in the introduction, learning traps can have harmful consequences in a remarkably wide range of domains, making it important to study methods of diminishing them. To provide one example where the attentional learning trap could have harmful effects, and to preview the cover story for Experiment 2, consider a company hiring new employees. If there are multiple attributes, each of which is sufficient to make an applicant suitable for the job, the fact that learning an applicant's true value is contingent on their being hired may mean that only one or a subset of these factors is ever attended to and learned. This can obviously have a negative effect on the company's success. Just as importantly, if the company learns to attend excessively to a feature like college attendance that covaries with socioeconomic variables, it can have negative long-run societal effects as well.

Because the role of exploration in the development of learning traps is relatively well established, while the role of generalization has rarely been studied, we focused our three interventions on slowing generalization from past experience to novel prospects, rather than on directly increasing exploration. While generalization is a vital aspect of intelligent behavior, and reducing generalization is certainly not always beneficial, we hypothesized that decreasing the speed of generalization would cause people to sample prospects more exhaustively before beginning to avoid those prospects they believed to be negative, thus reducing the likelihood of falling into the attentional learning trap.

In the following sections, we first describe the intuitions behind each intervention, explore them computationally using ALCOVE-RL, and then test them experimentally with people.

3.7.1 Individuating Prospects

One clear way to decrease generalization and potentially limit the attentional learning trap is to make stimuli increasingly distinct and idiosyncratic. When stimuli are more distinctive, people tend to treat them more as individuals and show increased ability to learn identification compared to categorization. While identification learning is more difficult than categorization with generic artificial stimuli (Shepard et al., 1961; Love, Douglas L. Medin, & Gureckis, 2004), Douglas L. Medin, Dewey, and Murphy (1983) found that people were more easily able to pair unique first names than categorical last names with photographs of faces. Love, Douglas L. Medin, and Gureckis (2004) argued that this phenomenon could be accounted for with the SUSTAIN model of categorization by assuming that the faces had many distinctive features beyond those manipulated by the experimenters, which decreased the similarity among stimuli and thus increased the odds of representing each stimulus individually.

In an approach–avoid decision-making task, increased individuation of stimuli should make a person less likely to generalize information gained from experience with one prospect to decisions about another. Attention paid to idiosyncratic features will slow the biasing of attention towards a single dimension, giving the person more opportunity to explore a variety of stimuli and learn the true structure of the environment. Essentially, increased individuation of prospects shifts the task away from category-learning, and towards learning about whether to approach individual prospects.

3.7.2 Occluding Feature Information

A second approach to decreasing the attentional learning trap may be to restrict information by randomly occluding some features of a prospect such that the decision maker can't observe their values. While this intervention could of course impair a person's decision-making ability, it could actually improve performance in the long run by causing a greater spread of attention. E. G. Taylor and Ross (2009) found that participants learned more about non-diagnostic features in a category-learning task when features were randomly occluded, and hypothesized that feature occlusion discourages rapid narrowing of selective attention and promotes a broader attentional strategy. In the context of approach decisions, if a person is attending strongly to a dimension that is occluded, he or she may be forced to use other features, which may lead to the discovery that they are relevant. Even when the favored features is not occluded, the possibility of their future absence may cause people to be less quick to rely solely on one feature.

3.7.3 Increasing Noise

Stochasticity in prospect outcomes is the driving force behind the hot stove effect and underweighting of rare events; without it, experience is never misleading, and there is no possibility for an incorrect belief about a prospect to develop. When negative beliefs formed about one prospect can generalize to another prospect, rather than simply to the same prospect at a later time, stochasticity is no longer required for biased behavior to result. Instead, it is plausible that a small degree of noise might aid the learning process in the long term. Noise is used in optimization algorithms such as simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) to overcome learning-trap-like local minima. Todd M. Gureckis and Love (2009) found that noise improved human performance in a dynamic decision-making task that required people to discover a non-obvious solution to a problem. A noisy outcome, while potentially triggering false belief in the region the experienced prospect, might also cause a reallocation of attention that could have globally beneficial consequences. Such an experience might cause a non-attended but useful dimension to attract attention, setting the agent on a new trajectory of behavior and learning and pulling it out of a learning trap.

3.7.4 Modeling Debiasing Interventions

To test the potential effect of these interventions on generalization and behavior, we performed model simulations comparing them to the basic contingent-information condition. To modify the model for the individuation condition, we added an extra dimension with 16 nominal values representing idiosyncratic features of each stimulus (following

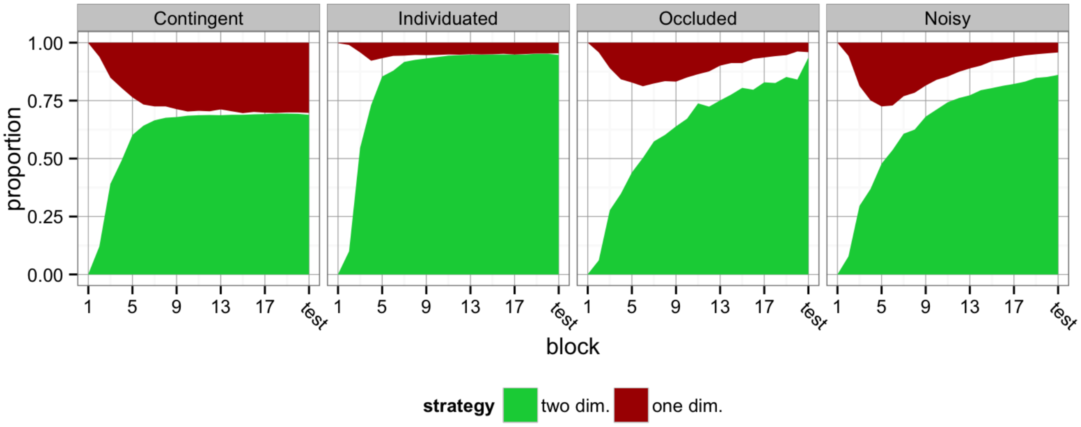


Figure 3.7: ALCOVE-RL model simulations of approach–avoid decision making in a basic contingent-information setting, and with three interventions. Panels show the proportion of model runs adopting the correct two-dimensional strategy or one of the one-dimensional learning traps in each of the 20 learning blocks and the test block.

Love, Douglas L. Medin, & Gureckis, 2004). For the occluded-dimension condition, on 25% of trials a randomly chosen dimension was masked such that it did not contribute to the model's network activation and its attention weight was not updated. For the noisy condition, on 5% of trials the potential outcome was flipped from the positive outcome to the negative outcome or vice versa.

The models were simulated for a learning phase of 20 blocks, followed by a one block test phase with no learning updates, where no dimensions were masked in the occluded-dimension condition and the individuating dimension was masked in the individuation condition. Results are plotted in Figure 3.7.

In the test phase, all three interventions lead to a greater proportion of the models adopting the correct two-dimensional rule. Interestingly, in early blocks the occlusion and increased noise interventions inhibit learning; while the model is less likely to enter the learning trap, it is also slower to learn the correct strategy. This indicates that these interventions represent a trade-off, allowing (potentially) superior long-term performance but at the cost of worse short-term performance.

3.8 Experiment 2

In Experiment 2, we tested whether individuating prospects, occluding feature information, or adding noisy outcomes would affect the degree to which participants fell into the attentional learning trap. In addition, to further test the generality of the learning trap we introduced a more life-like "job application" cover story, lengthened the training phase, and reduced the relative penalty for approaching negative prospects.

Prior to conducting Experiment 2, we conducted two large pilot experiments investigating different ways of implementing the interventions (see Appendix B). These pilot experiments showed the learning trap to be robust and did not show the interventions to be effective, foreshadowing the results of Experiment 2. Experiment 2 thus represents a final, large-n effort to replicate the learning trap and document the degree of effectiveness of the interventions.

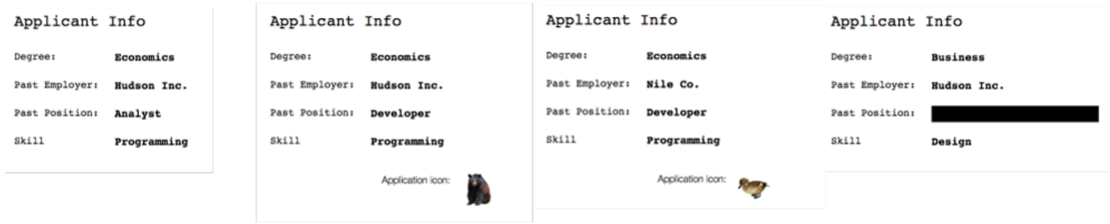
3.8.1 Method

Participants

Four hundred participants (176 female; 220 male) were recruited via Amazon Mechanical Turk. A power analysis showed that a sample size of about 80 participants per condition would allow us to reliably detect a 25% difference in percentage of people falling into the learning trap, or on continuous measures an effect size of .45. Participants received $2.00 for participation and received a performance-based bonus that ranged up to $1.68. Forty participants were excluded for requiring more than two attempts to pass a post-instructions quiz.

Stimuli

Stimuli were fake job applications that varied on four binary dimensions. Applicants had a “Degree” in “Business” or “Economics”, a “Past Employer” of either “Hudson Inc.” or “Nile Co.”, a “Skill” in either “Computer programming” or “Graphics editing”, and a “Past Position” of either “Product development” or “Market research”, for a total of 16 unique stimuli. Example stimuli are shown in Figure 3.8. Two of the four dimensions were chosen as relevant, counterbalanced across participants. Of the four possible combinations of values on these two dimensions, one was chosen at random; stimuli with this combination of values were “Unsuitable” applicants, while the remaining stimuli were “Suitable”.



Applicant Info

Degree: Economics
Past Employer: Hudson Inc.
Past Position: Analyst
Skill Programming

Applicant Info

Degree: Economics
Past Employer: Hudson Inc.
Past Position: Developer
Skill Programming

Application icon:

Applicant Info

Degree: Economics
Past Employer: Nile Co.
Past Position: Developer
Skill Programming

Application icon:

Applicant Info

Degree: Business
Past Employer: Hudson Inc.
Past Position: [redacted]
Skill Design

Figure 3.8: Examples of the stimuli used in Experiment 2. From left: an example stimulus from the contingent-information or full-information conditions, two example stimuli from the individuated condition, and an example stimulus from the occluded condition.

Procedure

Experiment 2's procedure is similar to that of Experiment 1. In this experiment, participants played the role of a recruiter considering a series of job applications. They were instructed that their goal was to generate revenue for their company.

In the learning phase, participants encountered each of the 16 unique stimuli eight times, for a total of 128 trials. The number of applications (i.e., trials) remaining was displayed throughout the learning phase. As in Experiment 1, stimuli were ordered such that each block of 16 contained all 16 stimuli, and each sub-block of 8 contained two negative and six positive stimuli, with stimulus order otherwise randomized.

On each trial, participants were presented with an application. The application started out blank, and participants had to press the space bar four times to reveal each of the four dimensions in a random order. This was done to reduce any bias towards attending to dimensions near the top of the application. The participant then had to choose whether to accept or reject the application. Accepting a suitable applicant generated revenue of $1 thousand for the company, while accepting an unsuitable applicant caused a loss of

$3 thousand. Rejecting an applicant caused no change in revenue. Revenue began at $50 thousand, and was converted to a cash bonus at the rate of $0.01 per $1 thousand.

Participants were split into five conditions. In the full-information condition, participants who rejected an applicant were informed of whether the applicant *would have been* suitable or unsuitable, and how the company's revenue would have changed. In the contingent condition, participants were given no feedback upon rejecting an applicant.

In the three intervention conditions, feedback during the learning phase was contingent as in the contingent condition. However, as shown in Figure 3.8, the stimuli were modified in ways hypothesized to reduce the learning trap. In the individuation condition, participants were instructed that the application system had a feature that assigned each of the 16 unique dimension value combinations a random unique icon to help them keep track of what they had observed. These icons were small pictures of animals that on some trials were displayed below the four dimensions, as shown in Figure 3.8. The icons were shown on 90% of trials in the first block of learning, which gradually decreased to 0% by the last block, so that participants could not rely solely on the icons.

In the occluded condition, on some trials one of the four dimensions was chosen at random and covered with a black bar (see Figure 3.8). If the participant hired the applicant, the hidden dimension was then revealed. This intervention was applied to 50% of trials in the first five blocks of training, and then was removed for the last three blocks.

In the noisy condition, applicant outcomes were changed from suitable to unsuitable or unsuitable to suitable on some randomly selected trials. Participants were informed that participants who appear suitable might occasionally be unsuitable, and vice versa.

This intervention was applied to 10% of trials in the first five blocks of training, and then was removed for the last three blocks.

The exact design of these interventions was informed by the data collected in our pilot experiments (described in the supplement). However, determining the exact difference in effect between subtly different interventions would require prohibitively large sample sizes, and the interventions were thus also guided by our own intuitions.

In all conditions, the learning phase was followed by a surprise 32-trial test phase, using the same randomization procedure as the learning phase. Participants chose to accept or reject as before, but received no feedback about the outcomes of their actions and were not able to see changes to revenue. Both interventions were also removed during the test phase so that it was equivalent across all five conditions.

After the test phase, participants were informed of their total bonus. As in Experiment 1, they were asked “About what percentage of applicants do you think were unsuitable?” and “Which fields do you think were useful in deciding whether an applicant was suitable or unsuitable?” We also added a third post-task question, which asked participants whether they believed they had learned completely how to use the applicant features to determine which applicants were suitable. Participants chose from a drop-down list either “I think I learned completely how the features determined suitability”, “I think there may have been aspects of applicant suitability that I did not learn”, or “I think there were definitely aspects of applicant suitability that I did not learn.”

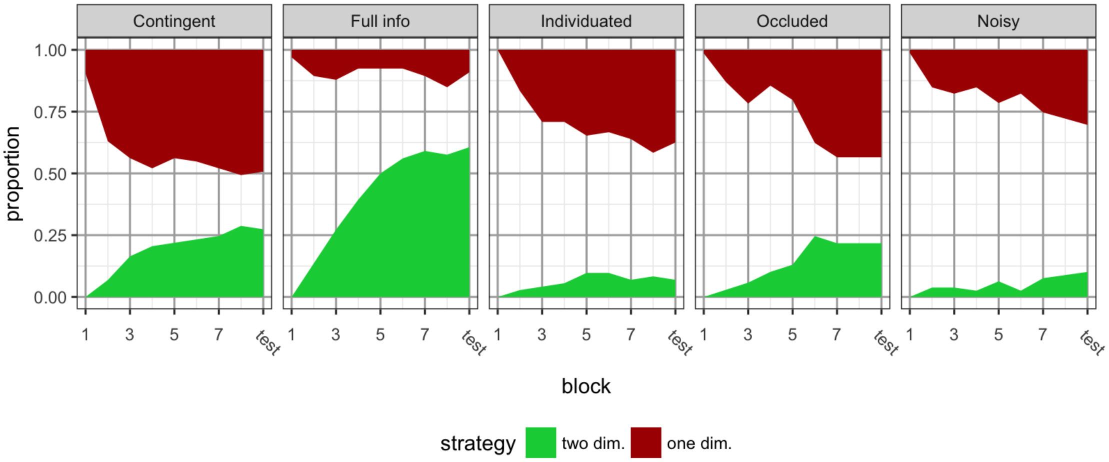


Figure 3.9: Participant behavior in the Experiment 2 job application decision-making game. Panels show proportion of participants adopting the correct two-dimensional strategy or one of the one-dimensional learning traps in each of the 8 learning blocks and in the test phase. Participants were coded as using a two-dimensional or one-dimensional strategy if at least 15/16 of their choices in a block were consistent with that strategy.

3.8.2 Results

After exclusions, there were 73 participants in the contingent-information condition, 66 in the full-information condition, 72 in the individuation condition, 70 in the occluded condition and 79 in the noisy condition. None of the experiment's results are qualitatively affected when excluded participants are included. Figure 3.9 shows participants' behavior over the eight blocks of learning and the test phase, using the same threshold for classifying participants as following a two-dimensional or one-dimensional strategy described in Experiment 1. We report first on the replication of the contingent-information and full-information conditions, followed by the results of the interventions.

3.8 Experiment 2

Experiment 2 included five conditions, increasing the risk of what from a frequentist perspective would be considered Type I errors due to multiple comparisons. To reduce this risk and improve the quality of our estimates, we adopted a Bayesian multilevel modeling approach to our analyses of Experiment 2, which assumes that the group means for each condition are drawn from the same overarching population distribution. This causes the group mean posteriors to be drawn towards each other during inference to a degree determined by the variability of the data, resulting in better estimates and fewer "false positives" (Andrew Gelman, Hill, & Yajima, 2012).

For continuous measures, rather than modeling the data as being drawn from *independent* normal distributions, we assumed that the condition means were themselves drawn from a population-level normal distribution with unknown mean and variance. We gave the population distribution mean a $Normal(0.5, 1)$ prior, and the population distribution standard deviation a $Normal(0, 0.5)$ distribution truncated at zero.

Similarly, for binary measures we posited that rather than the condition proportions being drawn independently, they were drawn from the same population-level beta distribution with unknown parameters. We reparameterized the standard $Beta(\alpha, \beta)$ distribution in terms of mean μ and precision κ, where $\alpha = \mu\kappa$ and $\beta = (1 - \mu)\kappa$, and specified a $Beta(2, 2)$ prior on μ and a $Gamma(1, .1)$ prior on κ.

We again estimated all models using Stan, with 4 independent chains of 10,000 samples.

As in Experiment 1, participants with contingent information were highly likely to accept candidates in the first block of learning, accepting 71% of applicants. Participants with full information were again less likely to do so, accepting 64%, $CI = [.02, .12]$, and were less likely to accept candidates than participants in the intervention conditions as well, with all CI's excluding zero.

Measures of the degree to which participants fell into the learning trap across conditions are plotted in Figure 3.10. Comparing the contingent and full-information conditions, the results fully replicate Experiment 1. In the test phase, participants in the two conditions diverged in their tendencies to adopt an incorrect one-dimensional strategy or the correct two-dimensional strategy . Full-information participants had an average 1D score of .75, while contingent-information participants averaged $.86$, $CI = [.04, .15]$. The pattern for the 2D score score was reversed; Full-information participants reached an average 2D score of $.87$, while those in the contingent-information condition reached a lower score of $.78$, $CI = [.03, .14]$.

While 25$\%$ of applicants were in fact unsuitable, contingent-information participants on average estimated this percentage to be 36.5$\%$. Full-information participants estimated the percentage to be 24.4$\%$, much closer to the true value and lower than the contingent-information condition, $CI = [.07, .17]$. Full-information participants were also more likely to correctly identify the two relevant dimensions, $CI = [.16, .46]$, while contingent-information were more likely to identify a single relevant dimension, $CI = [.27, .55]$.

Only around 10$\%$ of participants in the contingent and full-information conditions reported that there were "definitely" aspects of applicant suitability they did not learn, so we pooled these participants with those who reported that there "may" have been aspects they did not learn. In the contingent-information condition 45.2$\%$ of participants believed they "learned completely" how features determined suitability, while in the full-information condition 60.6% of participants believed they had learned completely. Full information may cause slightly more confident learning, but the credible interval for the difference contingent information does not exclude zero, $CI = [-.02, .29]$. Of the contingent-information participants who thought they had fully learned the task, 57.6% (19 out of 33) had in fact fallen into the learning trap. This suggests that in many cases participants entered a learning trap not just because it yielded "good-enough" performance, even while suspecting it to be incomplete, but because they believed it to represent the true structure of the environment.

Over all, as seen in Figure 3.10, the three interventions did not prevent the learning trap, or "prevented" the learning trap only insofar as they prevented effective learning altogether.

Specifically, the occlusion intervention appeared to have little effect, creating no difference outside the 95% credible interval from the contingent condition on any measure. The individuation and noise interventions did make people less likely to adopt a one-dimensional rule, $CI = [.01, .13]$ and $CI = [.03, .16]$ respectively. Compared to contingent-information condition participant, people in these conditions had lower 1D scores during the test phase. However, this was a hollow victory in that participants in these conditions also both had lower 2D scores at test, $CI = [.05, .17]$ and $CI = [.06, .17]$. In other words, participants who experienced these interventions had difficulty learning any stable structure of the task, correct or incorrect. This impression was reinforced when we asked participants whether they had learned the task completely. Participants in both the individuated and noisy conditions were less likely than those in the contingent-information condition to report that they had, $CI = [.07, .36]$

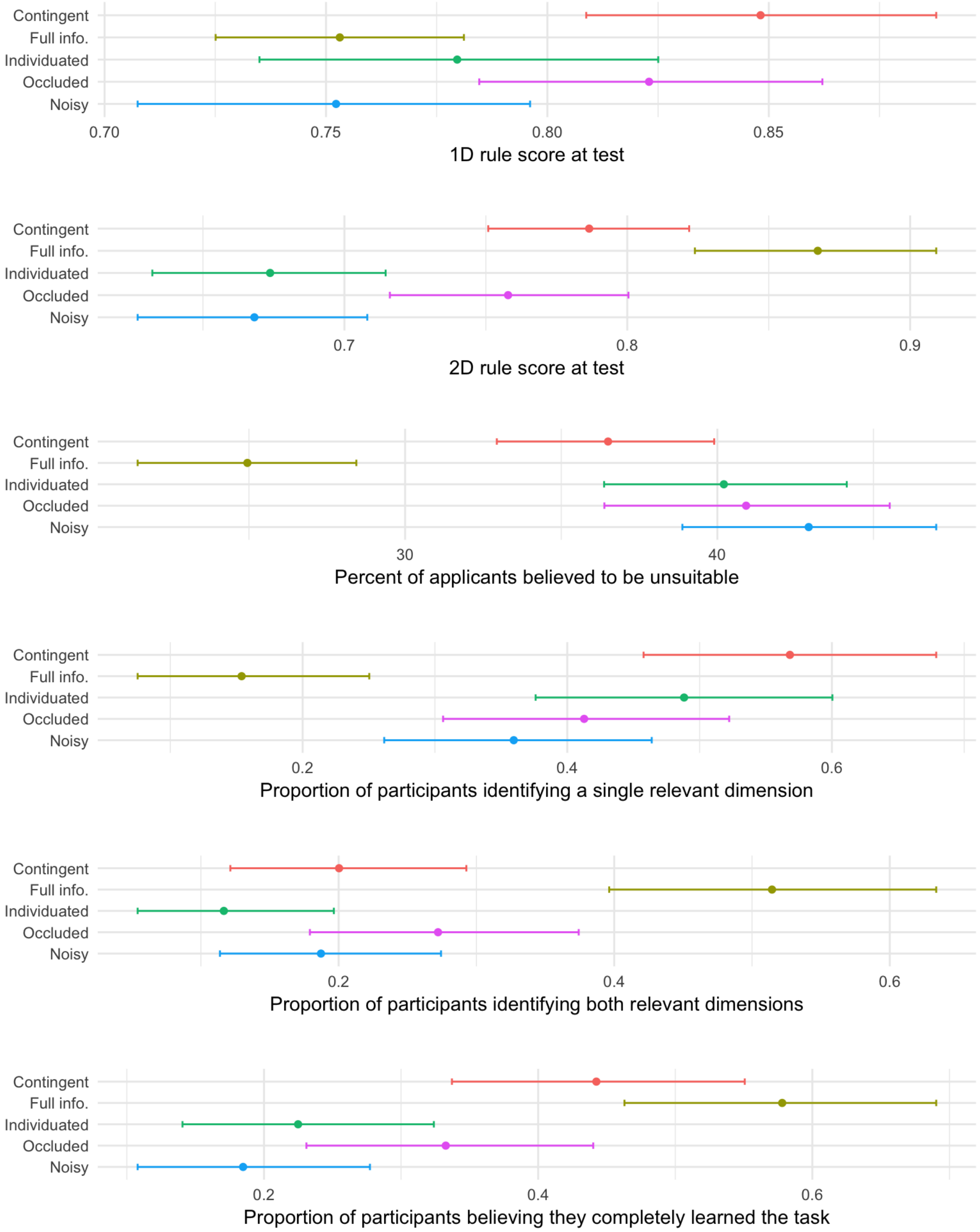


Figure 3.10: Comparisons of several measures of behavior across the five conditions of Experiment 2. Points indicate posterior population mean from Bayesian inference, and error bars indicate 95% credible intervals. All measures are support the conclusion that participants with contingent information fell into the attentional learning trap more readily than those with full information, but that the interventions did not lead to more robust learning compared to the contingent-information condition.

and $CI = [.12, .39]$.

Finally, compared to participants in the full-information condition, participants in all three intervention conditions had lower 2D scores during the test phase, were less likely to accurately identify the two relevant dimensions following the test, believed a greater proportion of applicants were unsuitable, and were less likely to report believing that they had learned completely how to determine applicant suitability, with the 95% credible intervals for all differences excluding zero. In other words, none of the interventions allowed people to obtain the accurate understanding of the environment possible through receiving full feedback following each choice.

3.9 General Discussion

Across two (plus Supplementary pilot) experiments and hundreds of participants we found evidence for a robust and novel learning trap. The attentional learning trap we describe is the joint product of the limits on exploration presented by choice-contingent feedback tasks and the inductive biases inherent in selective attention. Because the learning trap is related to the information restrictions of limited exploration, we found that it rarely occurred when people received full feedback regardless of their choices. Interestingly, the learning trap persisted when prospects were individuated, when prospect features were occasionally occluded, and when noise was added to prospect outcomes. Below, we address why these interventions did not prevent the learning trap and how learning traps might be diminished, as well as additional situations where we might expect learning traps to form.

3.9.1 The Potential for Reducing Learning Traps

The three interventions implemented in Experiments 2 did not, in any significant way, prevent participants from falling into the attentional learning trap. To the extent that they prevented rapid generalization across a single dimension, they also seemed to stop effective generalization entirely. That is, they prevented people from learning both the true, two-dimensional structure and the incomplete one-dimensional structure. This was evidenced both by participants' poor performance and their self-reports. While we believe future attempts at these sorts of interventions may prove more fruitful, our results do offer one possible lesson: decreasing generalization may be the wrong, or at least a risky, approach to reducing learning traps. As mentioned in the introduction, learning cannot occur without generalization (Mitchell, 1980). Thus, the challenge is to balance reduced generalization such that incorrect beliefs aren't quickly formed, but not so much that the correct beliefs are never formed.

An alternative approach is to directly increase exploration. Rather than slowing the process of learning, this approach seeks to increase people's willingness to continue enduring costs to gain information. The full-information conditions of both experiments experiments showed that accurate learning is possible in our tasks given enough information, suggesting this approach could be effective.

Exploration might be increased in multiple ways. First, recent studies show that human exploration is sensitive to the future usefulness of gaining information (Wilson et al., 2014; Rich & Todd M. Gureckis, 2017). Thus, increasing the salience or perceived number of future choices in an environment might increase exploration and decrease false beliefs. Second, exploration can be prompted by curiosity, which acts as an innate drive towards information-seeking (Berlyne, 1966; Loewenstein, 1994; Kidd & Hayden, 2015). This means that causing a greater intrinsic interest in a domain may be another tool for decreasing learning traps. Third, one might increase exploration by decreasing, at least temporarily, the drive for reward. Experiments with both animals (Tolman, 1948) and humans (Schwartz, 1982) have shown that excessive drive for external reward can prevent decision makers from learning the true structure of the world. If people can be given the opportunity to explore an environment in a disinterested manner, they will have a greater opportunity to build an accurate world model and will be less susceptible to learning traps.

Finally, keep in mind that the complete prevention of learning traps is often implausible and undesirable, since it requires excessive exploration at the cost of exploitation (Levinthal & March, 1993). If the primary goal is to maximize positive outcomes, rather than simply possess the most accurate beliefs about the world, it may make sense to allow the possibility of false beliefs (Le Mens & Denrell, 2011). Further research is needed to reduce biases in organizational and social behavior as much as possible, but some degree of bias may be an unavoidable byproduct of learning from experience.

3.9.2 The Possibilities of New Learning Traps

While the review in our paper attempted to integrate many past phenomena that are interpretable as a "learning trap," one interesting question is if there are other situations where such phenomena may occur. A number of possibilities come to mind, all having to do with more complex forms of generalization than we considered so far. For instance, people represent some real-world categories in terms of taxonomic hierarchies. Within these hierarchies there is often a psychologically prioritized level of abstraction, the *basic* level, that tends to be most naturally used when naming an object (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Waxman, 1990). *Dog*, for example, is a basic level category for most people, while *animal* is a superordinate category and *golden retriever* is a subordinate category. Because people tend to generalize at the basic level, there is a risk that negative attributes from single experience will be learned for an entire basic-level category, even if it truly applied only to a subordinate category.

Categories are also often connected to naive theories and causal beliefs, which can shape a person's understanding of the category and its members (Murphy & Douglas L Medin, 1985; Rehder & Hastie, 2001). One common kind of causal theory is that members of a category share a deep underlying cause, or *essence*, that gives rise to their other properties (Douglas L. Medin & Ortony, 1989; Susan A. Gelman, 2004). Essentialist beliefs are observed for natural kinds and for social categories such as race and gender (Gordon W. Allport, 1954; Hirschfeld, 1995; M. G. Taylor, 1996; M. G. Taylor, Rhodes, & Gelman, 2009). Since the attributes of instances of essentialized categories are believed to be caused by a shared essence, people tend to generalize a property of one instance to other instances more for essentialized categories than for non-essentialized ones (Susan A. Gelman, 1988; Susan A Gelman & Coley, 1990). This may make essentialized categories particularly susceptible to learning traps. In the domain of social categories, Denrell (2005) has described how the hot stove effect can lead to the development of negative perceptions of outgroups. Future work could examine whether essentialism plays a role in this type of social learning trap, and whether reductions in essentialist beliefs might help to prevent it.

3.9.3 Conclusion

Learning allows people to behave adaptively in a world that cannot be completely known *a priori*. But as we show in this report, learning processes cannot be relied on to converge steadily to true belief as a learner gains experience. When learning from experience influences reward-seeking choices, and reward-seeking choices produce the experience for further learning, the entire learning–choosing system can become stuck in patterns of poor decisions and false beliefs. In the words of Levinthal and March (1993), "learning has its own traps."

In this paper, we have tried to clarify the link between choice processes and learning processes in the development of learning traps. We have shown how, in environments with explore–exploit trade-offs created by choice-contingent feedback, many apparent suboptimalities are natural consequences of people's inductive biases. These inductive biases, rather than being flaws in the learning system, are prerequisites for effective generalization (Mitchell, 1980).

3.10 Context of research

This project emerged out of a broader family of projects examining how humans balance exploration and exploitation in complex environments. While some of our other studies have focused on how people use environmental cues to determine when to explore more or less (e.g., Rich & Todd M. Gureckis, 2017), this project aimed to understand how patterns of exploration affect beliefs when the environment is more complex than a simple repeated choice task. In doing so, it also addressed a perceived gap in the categorization literature, which has rarely looked closely at the challenges of learning via choice-contingent feedback. We believe a more complete understanding of exploratory choice can be gained by unpacking the two-way interaction between exploration and beliefs about the environment, and hope to continue to investigate this relationship as outlined in the Discussion.

This project also intersects with active and self-directed learning, another research focus of our lab (T. M. Gureckis & Markant, 2012). While active control over the contents of learning often produces gains over passive learning, learning traps provide a counterpoint: if the learner can choose to avoid stimuli that are expected to be unpleasant, active learning may lead to worse learning outcomes than passive exposure.

Appendices

A

APPENDIX A

A.1 Proof of the non-decreasing relative value of approaching

Following preliminary definitions, we show, first for the finite-horizon case and then for the infinite-horizon case, that the value of approaching relative to the value of avoiding increases or remains constant as the horizon becomes longer.

A.1.1 Preliminary definitions and assumptions

Let (F, A) denote a one-armed bandit problem with belief distribution F over the distribution of rewards on the uncertain arm, and discount sequence $A = (\alpha_1, \alpha_2, \dots)$. In the following proofs we assume the mean reward on the known arm is 0, but all results can easily be generalized to a bandit with arbitrary mean reward λ on the known arm by subtracting λ from all rewards on both arms to yield an equivalent bandit that fits our assumption. For a given bandit problem (F, A), let $W(F, A, \tau)$ be the summed expected reward obtained by following strategy τ. Let $V(F, A)$ be the expected value obtained by following an optimal strategy, such that $V(F, A) = \max_\tau W(F, A, \tau)$. Let $V_{ap}(F, A)$ be the expected value obtained by choosing the uncertain arm first and subsequently following an optimal strategy, and let $V_{av}(F, A)$ be the expected value obtained by choosing the known arm first and subsequently following an optimal strategy. For a given discount sequence $A = (\alpha_1, \alpha_2, \dots)$, let $A_{(1)} = (\alpha_2, \alpha_3, \dots)$.

Following Berry and Fristedt (1979), we define a *regular* discount sequence as one where for each m, $\gamma_m \gamma_{m+2} \leq \gamma_{m+1}^2$, where $\gamma_p = \sum_{i=p}^{\infty} \alpha_i$. We note that both finite-horizon and infinite-horizon discount sequences, as described below, are regular. For any bandit with a regular discount sequence, there exists an optimal strategy under which every choice of the known arm is followed by another choice of the known arm (see Berry & Fristedt, 1979).

A.1.2 Finite Horizon

Let (F, A) and (F, A^+) be two one-armed bandit problems that share a belief distribution F over the distribution of rewards on the uncertain arms, and a mean reward 0 on the known arms, and have discount sequences $A = (\alpha_1, \alpha_2, \dots)$ and $A^+ = (\alpha_1^+, \alpha_2^+, \dots)$. Let α_i equal 1 for $i \leq n$ and 0 for $i > n$, and let α_i^+ equal 1 for $i \leq n^+$, and 0 for $i > n^+$, where $n^+ > n$. Then

$$V_{ap}(F, A^+) - V_{av}(F, A^+) \geq V_{ap}(F, A) - V_{av}(F, A) \tag{A.1}$$

Proof

We complete the proof for the situation where $n^+ = n + 1$. If Equation A.1 holds in this situation, then by induction it must hold for $n^+ > n + 1$.

We first demonstrate that for all F, A and A^+ matching the definitions above, $V(F, A^+) \geq V(F, A)$. Let τ be an optimal strategy for (F, A) that always chooses the known arm following trial n, final consequential decision. (The elements of the discount sequence are 0 beginning at trial $n+1$, so choices beginning at this trial do not affect the strategy's worth.) Then the worth of following τ for (F, A^+) is $W(F, A^+, \tau) = V(F, A)$. Thus, $V(F, A^+) \geq V(F, A)$.

Proceeding to the main proof, there are four cases to consider. It may be optimal to choose the uncertain arm for (F, A) but not (F, A^+), it may be optimal to choose the uncertain arm for (F, A^+) but not (F, A), it may be optimal to choose the uncertain arm for both problems, or it may not be optimal to choose the uncertain arm for either problem. We address each of these cases in turn.

Uncertain arm is optimal for A but not A^+. If the uncertain arm is optimal for A but not A^+, then Equation A.1 fails because $V_{ap}(F, A) \geq V_{av}(F, A)$ but $V_{ap}(F, A^+) < V_{av}(F, A^+)$. We thus prove that this case cannot occur.

If the uncertain arm is not optimal for A^+, then the known arm is the sole optimal first choice. Because A^+ is regular, this implies that the strategy that always chooses the known arm is optimal. (If this were not the case, then there would be no optimal strategy for which every choice of the known arm is followed by another choice of the known arm.) This in turn means that $V(F, A^+) = 0$, and that $V_{ap}(F, A^+) < 0$.

Thus,

$$\begin{aligned} V_{ap}(F, A) &= V(F, A) = \mathbb{E}[X_1 + V(F', A_{(1)})] \\ V_{ap}(F, A^+) &= \mathbb{E}[X_1 + V(F', A^+_{(1)})] < 0 \\ V_{av}(F, A^+) &= V(F, A^+) = 0 \end{aligned}$$

Where F' is the belief distribution following observing the outcome X_1. We note that $V(F, A) \geq 0$, because it is always possible to always choose the known arm for a guaranteed payoff of zero. However, since F', $A_{(1)}$, and $A^+_{(1)}$ match our problem definition above, we know that $V(F', A^+_{(1)}) \geq V(F', A_{(1)})$. This means it cannot be the case that $\mathbb{E}[X_1 + V(F', A^+_{(1)})] < 0$ and $\mathbb{E}[X_1 + V(F', A_{(1)})] \geq 0$. Thus, it cannot be the case that the uncertain arm is optimal for A but not A^+.

Uncertain arm is optimal for A^+ but not A. If the uncertain arm is optimal for A^+ but not A, then $V_{ap}(F, A^+) \geq V_{av}(F, A^+)$ and $V_{ap}(F, A) < V_{av}(F, A)$, so Equation A.1 clearly holds.

Uncertain arm is optimal for neither problem. If the uncertain arm is not optimal for either problem, then

$$\begin{aligned} V_{ap}(F, A) &= \mathbb{E}[X_1 + V(F', A_{(1)})] \\ V_{ap}(F, A^+) &= \mathbb{E}[X_1 + V(F', A^+_{(1)})] \\ V_{av}(F, A) &= V(F, A) = 0 \\ V_{av}(F, A^+) &= V(F, A^+) = 0 \end{aligned}$$

Where F' is the belief distribution following observing the outcome X_1.

Since $V(F', A^+_{(1)}) \geq V(F', A_{(1)})$, it must be the case that

$$\mathbb{E}[X_1 + V(F', A^+_{(1)})] \geq \mathbb{E}[X_1 + V(F', A_{(1)})]$$

Then because $V_{av}(F, A^+) = V_{av}(F, A) = 0$, Equation A.1 holds.

A APPENDIX A

Uncertain arm is optimal for both problems. If the uncertain arm is optimal for both problems, then

$$V_{ap}(F, A) = V(F, A)$$

$$V_{ap}(F, A^{+}) = V(F, A^{+})$$

$$V_{av}(F, A) = V(F, A_{(1)})$$

$$V_{av}(F, A^{+}) = V(F, A^{+}_{(1)})$$

Observe that $A^{+}_{(1)} = A$. Thus, we want to show that

$$V(F, A^{+}) - V(F, A) \geq V(F, A) - V(F, A_{(1)})$$

Let τ be the optimal strategy for (F, A) that yields the greatest expected reward at the final choice, choice n, and let Z_n be that expected reward. Then $V(F, A) \leq V(F, A_{(1)}) + Z_n$, otherwise $W(F, A_{(1)}, \tau) > V(F, A_{(1)})$. In other words optimizing expected reward on all trials up to but not including the final trial n cannot yield expected reward less than Z_n smaller than $V(F, A)$.

For (F, A^{+}), by following τ and then making the same selection on choice $n + 1$ as on choice n, we show that $V(F, A^{+}) \geq V(F, A) + Z_n$. In other words, optimizing expected rewards through trial n plus one additional trial yields expected reward at least Z_n greater than $V(F, A)$. Thus

$$V(F, A^{+}) - V(F, A) \geq V(F, A) - V(F, A_{(1)})$$

and Equation A.1 holds.

A.1.3 Infinite Horizon

Let (F, A) and (F, A^+) be two one-armed bandit problems that share a belief distribution F over the distribution of rewards on the uncertain arms, and a mean reward 0 on the known arms, and have discount sequences $A = (\alpha_1, \alpha_2, \dots)$ and $A^+ = (\alpha_1^+, \alpha_2^+, \dots)$. Let $\alpha_i = d^{i-1}$ and $\alpha_i^+ = e^{i-1}$, where d and e are between 0 and 1 and $e > d$. Then

$$V_{ap}(F, A^+) - V_{av}(F, A^+) \geq V_{ap}(F, A) - V_{av}(F, A)$$

Proof

We again begin by demonstrating that for all F, A and A^+ matching the definitions above, $V(F, A^+) \geq V(F, A)$.

First, suppose $V(F, A) = 0$. Then clearly $V(F, A^+) \geq V(F, A)$ because always choosing the known arm guarantees a payoff of 0.

Now suppose $V(F, A) > 0$. Following Berry and Fristedt (1979), we observe that because A is regular there must be an optimal strategy τ that chooses the uncertain arm at stages 1 through N and chooses the known arm subsequently, where N is a random variable with $p(N \geq 1) = 1$. N may depend on the history of observations and may be infinite with positive probability.

For (F, A), the expected payoff of τ is

A Appendix A

$$\mathbb{E}[\sum_{m=1}^{N} X_m \alpha_m]$$

Where X_m is the payoff on the uncertain arm realized at stage m.

We proceed by showing that the expected payoff of following τ is greater or equal for (F, A^+), and thus that $V(F, A^+) \geq V(F, A)$.

We seek to show that

$$\sum_{m=1}^{\infty} b_m \alpha_m^+ \geq \sum_{m=1}^{\infty} b_m \alpha_m$$

where $b_m = \mathbb{E}[X_m \mathbb{1}_{\{N \geq m\}}]$

Using the notation $\gamma_m = \sum_m^\infty \alpha_m$ and $\gamma_m^+ = \sum_m^\infty \alpha_m^+$, we rewrite the sums as

$$\sum_{m=1}^{\infty} b_m \alpha_m = \sum_{m=1}^{\infty} b_m (\gamma_m - \gamma_{m+1}) = b_1 \gamma_1 + \sum_{m=1}^{\infty} (b_{m+1} - b_m) \gamma_{m+1}$$

$$\sum_{m=1}^{\infty} b_m \alpha_m^+ = \sum_{m=1}^{\infty} b_m (\gamma_m^+ - \gamma_{m+1}^+) = b_1 \gamma_1^+ + \sum_{m=1}^{\infty} (b_{m+1} - b_m) \gamma_{m+1}^+$$

so we now seek to show that

$$\gamma_1^+ [b_1 + \sum_{m=1}^{\infty} (b_{m+1} - b_m) \frac{\gamma_{m+1}^+}{\gamma_1^+}] \geq \gamma_1 [b_1 + \sum_{m=1}^{\infty} (b_{m+1} - b_m) \frac{\gamma_{m+1}}{\gamma_1}]$$

We begin by showing that

$$b_1 + \sum_{m=1}^{\infty} (b_{m+1} - b_m) \frac{\gamma_{m+1}^+}{\gamma_1^+} \geq b_1 + \sum_{m=1}^{\infty} (b_{m+1} - b_m) \frac{\gamma_{m+1}}{\gamma_1}$$

A.1 Proof of the non-decreasing relative value of approaching

This follows from two facts. First, as shown by Berry and Fristedt (1979),

$$\begin{aligned} b_{m+1} - b_m &= \mathbb{E}[(-X_m)\mathbb{1}_{\{N=m\}}] + \mathbb{E}[(X_m - X_{m-1})(1 - \mathbb{1}_{\{N<m\}})] \\ &= \mathbb{E}[(-X_m)\mathbb{1}_{\{N=m\}}] \geq 0 \end{aligned}$$

so $b_{m+1} - b_m \geq 0$. Second, $\frac{\gamma_m^+}{\gamma_1^+} \geq \frac{\gamma_m}{\gamma_1}$, because $\frac{\gamma_m^+}{\gamma_1^+} = e^{m-1}$, $\frac{\gamma_m}{\gamma_1} = d^{m-1}$, and $e^{m-1} > d^{m-1}$ because $e > d$.

$$b_1 + \sum_{m=1}^{\infty}(b_{m+1} - b_m)\frac{\gamma_{m+1}^+}{\gamma_1^+} \geq b_1 + \sum_{m=1}^{\infty}(b_{m+1} - b_m)\frac{\gamma_{m+1}}{\gamma_1}$$

Finally, $\gamma_1^+ > \gamma_1$, and in fact $\gamma_1^+ = \frac{1}{1-e}$ and $\gamma_1 = \frac{1}{1-d}$. Thus not only do we find that

$$\gamma_1^+[b_1 + \sum_{m=1}^{\infty}(b_{m+1} - b_m)\frac{\gamma_{m+1}^+}{\gamma_1^+}] \geq \gamma_1[b_1 + \sum_{m=1}^{\infty}(b_{m+1} - b_m)\frac{\gamma_{m+1}}{\gamma_1}]$$

which means

$$\sum_{m=1}^{\infty} b_m \alpha_m^+ \geq \sum_{m=1}^{\infty} b_m \alpha_m$$

we in fact find that

$$\sum_{m=1}^{\infty} b_m \alpha_m^+ \geq \frac{1-d}{1-e}\sum_{m=1}^{\infty} b_m \alpha_m$$

Thus $V(F, A^+) \geq V(F, A)$, and in fact in all cases $V(F, A^+) \geq \frac{1-d}{1-e}V(F, A)$.

Proceeding to the main proof, we consider the same four cases. It may be optimal to choose the uncertain arm for (F, A) but not (F, A^+), it may be optimal to choose the uncertain arm for (F, A^+) but not (F, A), it may be optimal to choose the uncertain arm for both problems, or it may not be optimal to choose the uncertain arm for either problem.

Having shown that $V(F, A^+) \geq V(F, A)$, the first three cases are identical to the finite horizon proof above, so we focus on the final case. If it is optimal to choose the uncertain arm for A and A^+, then due to the exponential nature of the discount sequence

$$\begin{aligned}
V_{ap}(F, A) &= V(F, A) \\
V_{ap}(F, A^+) &= V(F, A^+) \\
V_{av}(F, A) &= dV(F, A) \\
V_{av}(F, A^+) &= eV(F, A^+)
\end{aligned}$$

We thus seek to show that

$$V(F, A^+) - eV(F, A^+) \geq V(F, A) - dV(F, A)$$

which means we must show that

$$V(F, A^+) \geq \frac{1-d}{1-e} V(F, A)$$

This is precisely what we proved above, so Equation A.1 holds.

A.2 Description of data analysis

A.2.1 Main analysis

Participants' trial-by-trial data was modeled using a hierarchical Bayesian approach. In a hierarchical model, each individual is modeled using a unique set of parameters, but those parameters are assumed to be drawn from a common population distribution. This approach allows for individual differences while yielding robust estimates of population-level parameters of interest (A Gelman et al., 2014).

For each of the five coefficients in our hierarchical logistic regression (bias, expectation, log horizon, log trial, and log horizon × log trial), participants' individual coefficients were modeled as being drawn from a normal population distribution with unknown mean and standard deviation. We assumed vague priors for population-level parameters.

For each data set (each experiment or condition within an experiment) we sampled from the model posterior using the Stan modeling language (Stan Development Team, 2015) to run Hamiltonian Monte Carlo simulation. Each simulation was run in four chains of 1000 samples, with the first half discarded as burn-in, and we confirmed convergence using the $\hat{R}$ convergence diagnostic (A Gelman et al., 2014). Posterior means and 95% posterior intervals for the population distribution means of each coefficient are plotted in Figure 1.4.

For each sample of parameters from the posterior for each participant, we also simulated the model's behavior on the stimuli presented to that participant. This gives an estimate of how one would expect participants to behave if our model was a true description of their behavioral processes. The mean behavior of the model simulations is plotted in Figure **??** for Experiment Sequence 1 and in Figure **??** for Experiment Sequence 2. These graphs show that the model is able to capture the main patterns in our results.

A.2.2 Expanded model for Experiment Sequence 1

To test our assumption in Experiment Sequence 1 that the effect of horizon on behavior is logarithmic, we performed a supplemental analysis using a modified version of the hierarchical Bayesian model. In this expanded model, we added a vector of 32 freely-varying population-level parameters that measured the effect of horizons of each length, from 1 to 32. On trials with horizon n, rather than using $\log(n)$ as the predictor for horizon, the model used the nth element of the population-level horizon-length vector.

Due to computational and data constraints, this vector of horizon effects was estimated at the population, but not individual, level. Each individual was still assigned a single individual-level horizon parameter that determined the over-all sensitivity to horizon of the individual and scaled the vector of horizon effects up or down. The population mean of these individual parameters was fixed at one to prevent redundancy with the horizon-length parameters.

Figure **??** shows the means and posterior intervals of the 32 horizon parameters for Experiment 1a and the contingent-information condition of Experiment 1b. While allowed to freely vary, the parameters adopt a roughly logarithmic form, supporting our

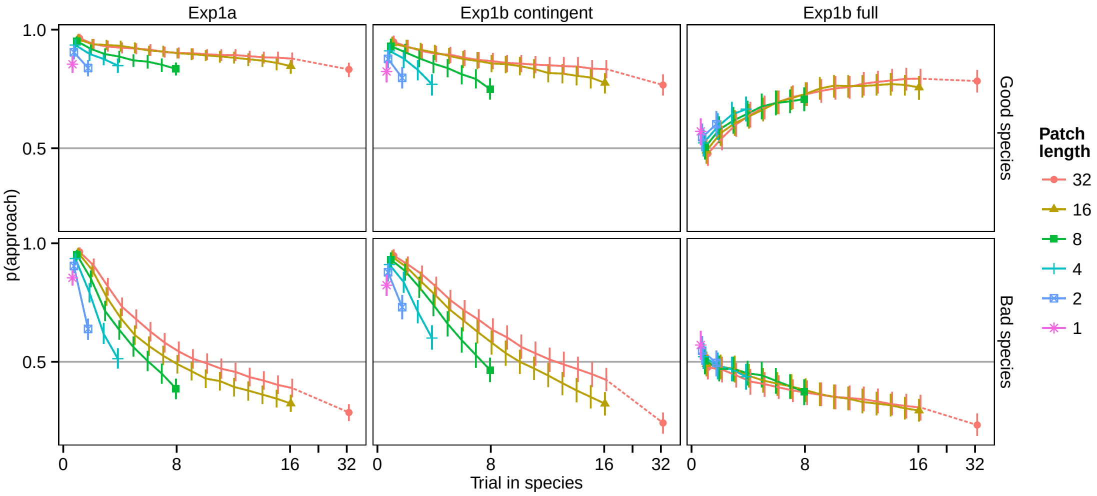


Figure A.1: Simulated probability of approaching a mushroom on each encounter within a patch for each finite-horizon experiment and condition. Data was generated using sampled parameters from the posteriors of all individuals from the Bayesian logistic regression model. Lines show mean behavior over the simulations and error bars show 95% posterior intervals.

supposition of a logarithmic effect of horizon. The model was not fit to the data from the full-information condition of Experiment 1b, as there was no effect of horizon in this condition.

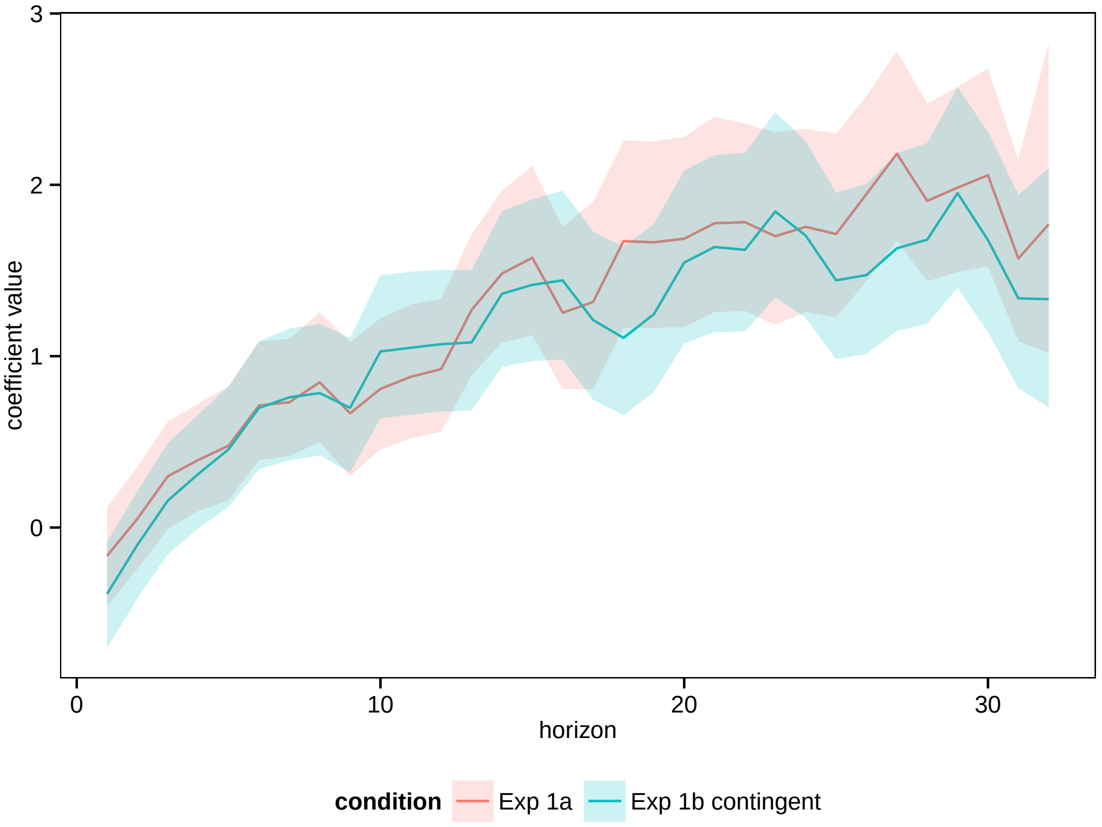


Figure A.2: Population-level effect of horizons from 1 to 32 trials in Experiment 1a and the contingent-information condition of Experiment 1b. Lines show posterior means and shaded regions show 95% posterior intervals. In both scenarios, a roughly logarithmically increasing effect of horizon is observed.

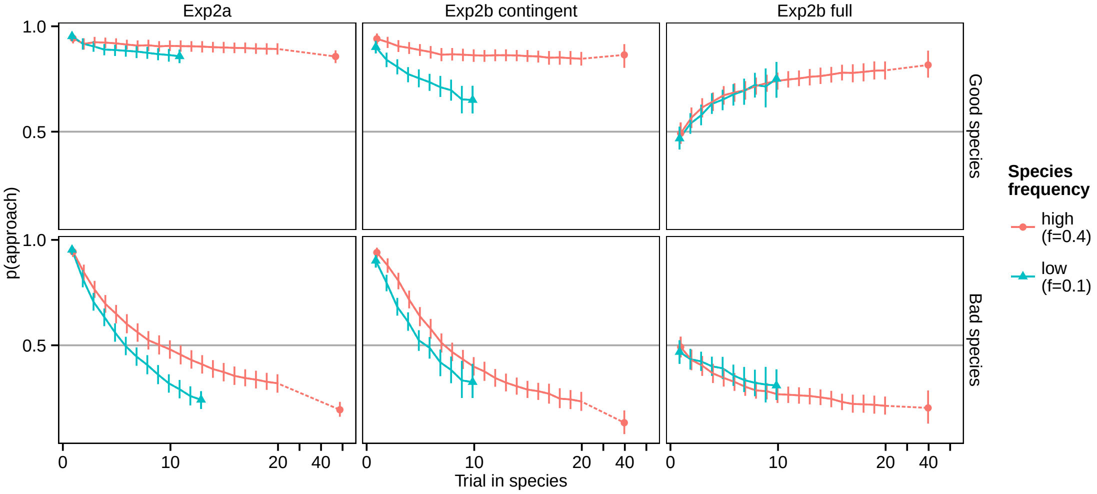


Figure A.3: Population-level effect of horizons from 1 to 32 trials in Experiment 1a and the contingent-information condition of Experiment 1b. Lines show posterior means and shaded regions show 95% posterior intervals. In both scenarios, a roughly logarithmically increasing effect of horizon is observed.

B

Appendix B

To test how effective different interventions (and different modes of implementation) were in preventing the attentional learning trap, we conducted two fairly large pilot experiments. In both of these experiments we observed the data during collection, and stopped data collection at about 30 participants per group when it became clear that the interventions did not lead to noticeably better performance than the contingent information condition. In the main text of the paper, we report on a final experiment we ran, with a fixed sample size of 80 participants per condition, to confirm the effectiveness or ineffectiveness of the interventions. In this section, we describe the two pilot experiments and their results.

B.1 Experiment S1

In Experiment S1, we tested whether individuating prospects or occluding feature information would affect the degree to which participants fell into the attentional learning trap (increasing noise is introduced in Experiment S2). In addition, to further test the generality of the learning trap we introduced a more life-like "job application" cover story, lengthened the training phase, and reduced the relative penalty for approaching negative prospects.

B.1.1 Method

Participants

One hundred twenty two participants (63 female; 58 male) were recruited via Amazon Mechanical Turk. We performed a preliminary analysis of our results after collecting 30 participants per group and halted data collection upon observing that the learning trap was robust and that the interventions did not improve learning. Participants received $2.00 for participation and received a performance-based bonus that ranged up to $1.64. Seven participants were excluded for requiring more than two attempts to pass a post-instructions quiz.

Stimuli

Stimuli were fake job applications which varied on four binary dimensions. Applicants had a "Degree" in "Business" or "Economics", a "Past Employer" of either "Hudson Inc." or "Nile Co.", a "Skill" in either "Computer programming" or "Graphics editing", and a "Past Position" of either "Product development" or "Market research", for a total of 16 unique stimuli. Example stimuli are shown in Figure B.1. Two of the four dimensions were chosen as relevant, counterbalanced across participants. Of the four possible combinations of values on these two dimensions, one was chosen at random; stimuli with this combination of values were "Unsuitable" applicants, while the remaining stimuli were "Suitable".

B Appendix B

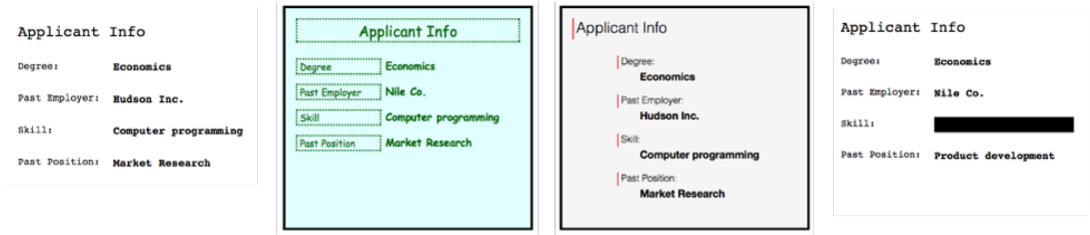


Figure B.1: Examples of the stimuli used in Experiment S1. From left: an example stimulus from the contingent-information or full-information conditions, two example stimuli from the individuated condition, and an example stimulus from the occluded condition.

Procedure

Experiment S1's procedure is similar to that of Experiment 1. In this experiment, participants played the role of a recruiter considering a series of job applications. They were instructed that their goal was to generate revenue for their company.

In the learning phase, participants encountered each of the 16 unique stimuli eight times, for a total of 128 trials. The number of applications (i.e., trials) remaining was displayed throughout the learning phase. As in Experiment 1, stimuli were ordered such that each block of 16 contained all 16 stimuli, and each sub-block of 8 contained two negative and six positive stimuli, with stimulus order otherwise randomized.

On each trial, participants were presented with an application. The application started out blank, and participants had to press the space bar four times to reveal each of the four dimensions in a random order. This was done to reduce any bias towards attending dimensions near the top of the application. The participant then had to choose whether to accept or reject the application. Accepting a suitable applicant generated revenue of $1 thousand for the company, while accepting an unsuitable applicant caused a loss of

$4 thousand. Rejecting an applicant caused no change in revenue. Revenue began at $50 thousand, and was converted to a cash bonus at the rate of $0.01 per $1 thousand.

Participants were split into four conditions. In the full-information condition, participants who rejected an applicant were informed of whether the applicant *would have been* suitable or unsuitable, and how the company's revenue would have changed. In the contingent condition, participants were given no feedback upon rejecting an applicant.

In the two intervention conditions, feedback during the learning phase was contingent as in the contingent condition. However, as shown in Figure B.1, the appearance of the stimuli was modified in ways hypothesized to reduce the learning trap. In the individuated condition, each of the 16 unique stimuli was given a distinct visual "style" by changing the font, colors, and layout. In the occluded condition, on half of the trials one of the four dimensions was chosen at random and covered with a black bar. This means that when the feature was revealed by pressing the space bar, the black bar became visible instead of the feature text. Two examples of individuated stimuli and an example of a stimulus with an occluded dimension are shown in Figure B.1.

In all conditions, the learning phase was followed by a surprise 32-trial test phase, using the same randomization procedure as the learning phase. Participants chose to accept or reject as before, but received no feedback about the outcomes of their actions and were not able to see changes to revenue. Both interventions were also removed during the test phase so that it was equivalent for all four conditions.

After the test phase, participants were informed of their total bonus. As in Experiment 1, they were asked two final questions: "About what percentage of applicants do you think were unsuitable?" and "Which fields do you think were useful in deciding

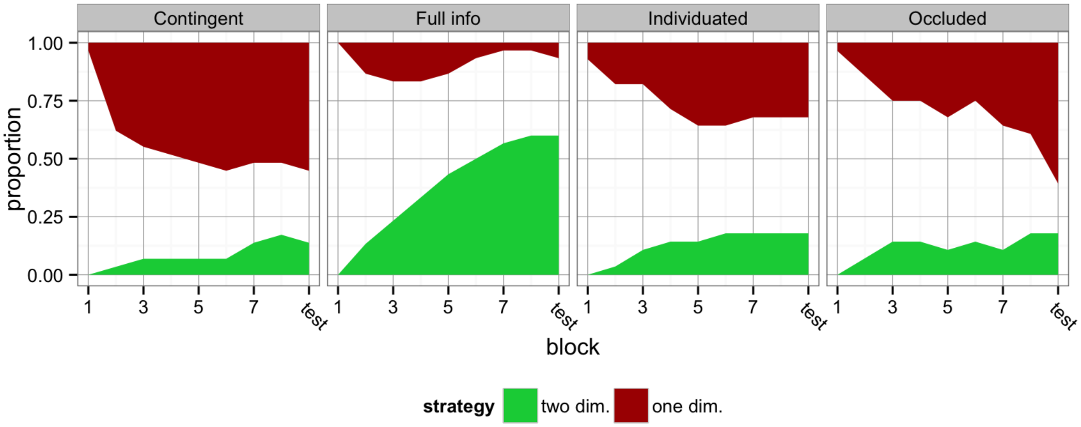


Figure B.2: Participant behavior in the Experiment S1 job application decision-making game. Panels show proportion of participants adopting the optimal two-dimensional strategy or one of the suboptimal one-dimensional learning traps in each of the 8 learning blocks and in the test phase. Participants were coded as using a two-dimensional or one-dimensional strategy if at least 15/16 of their choices in a block were consistent with that strategy.

whether an applicant was suitable or unsuitable?"

B.1.2 Results

Figure B.2 shows participants' behavior over the eight blocks of learning and the test phase, using the same threshold for classifying participants as following a two-dimensional or one-dimensional strategy described in Experiment 1. We report first on the replication of the contingent-information and full-information conditions, followed by the results of the interventions.

As described in the results section of Experiment 2, we analyzed the results of the experiment using a Bayesian multi-level modeling approach. Measures pertaining to whether participants entered a learning trap or learned the true structure of the environment are plotted in Figure B.6. As can be seen, the contingent and full-information conditions replicate Experiment 1, while the interventions did not appear to improve performance.

By the test phase, participants in the contingent and full-information conditions diverged in their tendencies to adopt the optimal two-dimensional strategy or a suboptimal one-dimensional strategy. Full-information participants reached an average 2D score of .87, while those in the contingent-information condition reached a lower score of .73, $CI = [.00, .20]$. The pattern for the 1D score was reversed; full-information participants averaged .73, while contingent-information participants averaged .85, $CI = [.02, .19]$.

While 25% of applicants were truly unsuitable, contingent-information participants on average estimated this percentage to be 43.7% (one participant was excluded for producing a negative response). Full-information participants estimated the percentage to be 25.7%, much closer to the true value, $CI = [.07, .25]$. Full-information participants were also more likely to correctly identify the two relevant dimensions, $CI = [.17, .61]$, while contingent-information participants were more likely to identify a single relevant dimension, $CI = [.09, .53]$.

Overall, the computationally principled interventions did not prevent the attentional learning trap (See Figures B.3–B.2). Examination of Figure B.2 shows that the behavior of participants in the individuated and occluded conditions mostly resembles behavior in the contingent-information condition. No differences in responses between participants in the contingent-information condition and those in the intervention conditions, either during the task or in the post-task questionnaire, had 95% credible intervals ex-

B Appendix B

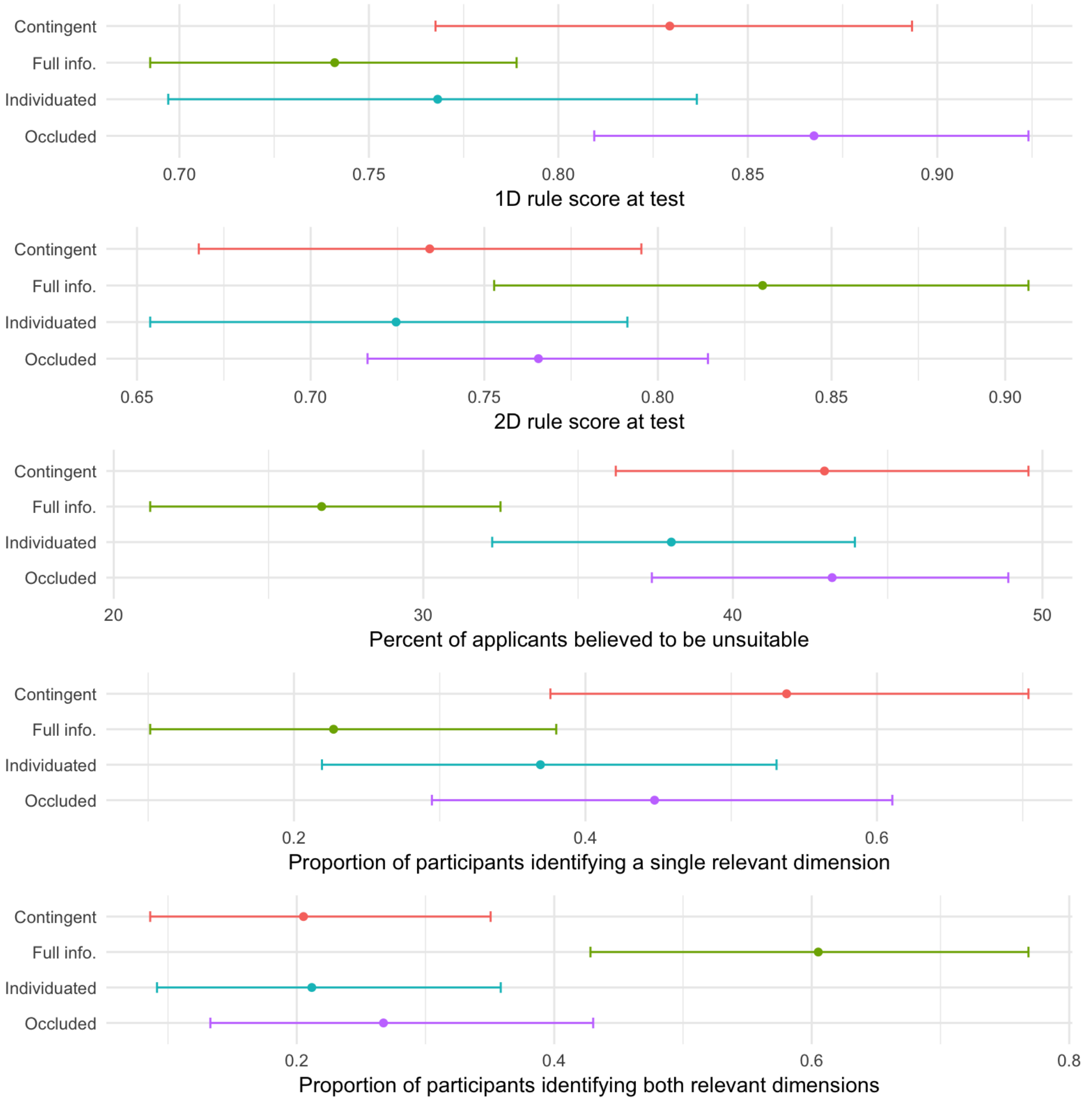


Figure B.3: Comparisons of several measures of behavior across the four conditions of Experiment S1. Points indicate posterior population mean from Bayesian inference, and error bars indicate 95% credible intervals. While the strength of these inferences are weakened by the experiment's small sample size and non-predetermined stopping rule, they support the conclusion that participants with contingent information fell into the attentional learning trap more readily than those with full information, but that the interventions did not lead to more robust learning compared to the contingent-information condition.

cluding zero. Interestingly, there was a trend towards participants in the individuated condition being less likely to report only a single relevant dimension in the questionnaire $CI = [-.05, .40]$, and these participants' behavior was marginally less consistent with a one-dimensional strategy during the test phase, $CI = [-.02, .18]$. Thus, the individuation intervention may have partially succeeded in reducing the rapid generalization that underlies the attentional learning trap, but without helping participants to eventually converge to the correct rule.

B.2 Experiment S2

In Experiment S1, we replicated the attentional learning trap in a new, more real-world domain, but found that the trap was resistant to both the individuation and occlusion interventions. The inability of the manipulations to decrease the prevalence of the learning trap may be because they were not strong enough. Thus, in Experiment S2, we modified and retested these two interventions with the goal of increasing their effectiveness. We additionally tested the effect of outcome noise which was not considered as a condition in Experiment S1.

We made two changes to the basic task structure. The first change was to decrease the penalty for hiring an unsuitable applicant from $4 thousand (4 cents) to $3 thousand (3 cents). This was predicted to encourage slightly more exploration by decreasing the potential cost of hiring an uncertain applicant. The second change was to give each dimension a "right" and "left" value (see Stimuli section and Figure B.4). This made the differences between stimuli more salient, somewhat analogous to the individuation condition, and may help people remember which stimuli they have had positive or negative experiences with.

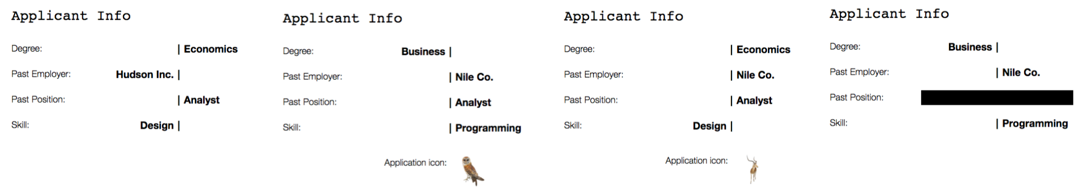
We also made several changes to the implementation of the interventions. In the individuated condition, rather than changing incidental stylistic aspects of the stimuli, we paired each of the sixteen stimuli with a unique animal icon, and explicitly told participants the icons were meant to serve as a memory aid. In the occluded condition, we revealed the value of the occluded dimension after a choice had been made so that the occlusion created an incentive to learn about multiple dimensions but did not impede learning after a choice. We also added a new, "noisy" condition to test whether the addition of outcome noise would improve performance, as suggested by our simulations.

Finally, for each interventions we phased the intervention in or out in a manner we expected to increase its effectiveness. In the individuated condition, we gradually removed the icons on later trials so participants could not learn solely by memorizing the icons. In the occluded and noisy conditions, we increased the number of stimuli with occluded dimensions or atypical outcomes over time so that the experiment was not too difficult early in the learning phase.

B.2.1 Method

Participants

One hundred fifty four participants (67 female; 87 male) were recruited via Amazon Mechanical Turk. As in Experiment S1, we performed a preliminary analysis of our results after collecting 30 participants per group and halted data collection upon observ-



Applicant Info

Degree: | Economics
Past Employer: Hudson Inc. |
Past Position: | Analyst
Skill: Design |

Applicant Info

Degree: Business |
Past Employer: | Nile Co.
Past Position: | Analyst
Skill: | Programming
Application icon:

Applicant Info

Degree: | Economics
Past Employer: | Nile Co.
Past Position: | Analyst
Skill: Design |
Application icon:

Applicant Info

Degree: Business |
Past Employer: | Nile Co.
Past Position: [redacted]
Skill: | Programming

Figure B.4: Examples of the stimuli used in Experiment S2. From left: an example stimulus from the contingent-information, full-information, or noisy conditions, two example stimuli from the individuated condition, and an example stimulus from the occluded condition.

ing that the learning trap was robust and that the interventions did not improve learning. Participants received $2.00 for participation and received a performance-based bonus that ranged up to $1.64. Fourteen participants were excluded for requiring more than two attempts to pass a post-instructions quiz.

Stimuli

In Experiment S1, the two values for a given dimension replaced each other in the same physical location. We modified the stimuli by assigning each dimension a "left" and a "right" value that were always printed in their respective locations, as shown in Figure B.4.

Procedure

The procedure was similar to that of Experiment S1. The penalty for accepting unsuitable candidates was reduced from $4 thousand (4 cents) to $3 thousand (3 cents).

In the individuation condition, stimuli were no longer differentiated by style. Instead, participants were instructed that the application system had a feature that assigned each of the 16 unique dimension value combinations a random unique icon to help them keep track of what they had observed. This icon was a small picture of an animal displayed below the four dimensions, as shown in Figure B.4. Participants were instructed that the icon would not always be available, and that they should still focus on learning how the dimensions predicted applicant suitability.

In the occlusion condition, one dimension was sometimes occluded as in Experiment S1 (see Figure B.4). However, if the participant hired the applicant, the hidden dimension was then revealed.

In the noisy condition, applicant outcomes were changed from suitable to unsuitable or unsuitable to suitable on some randomly selected trials. Participants were informed that participants who appear suitable might occasionally be unsuitable, and vice versa.

The strengths of the three interventions changed over the course of the training phase in an attempt to maximize their efficacy. In the individuation condition, the application icons were available on 90% of trials in the first block, which fell linearly to 0% of trials in the final block of training. In the occluded condition, 20% of trials had an occluded dimension in the first block, which increased to 50% by the final block. In the noisy condition, 10% of trials had a flipped outcome in the first through fourth blocks, which then fell linearly to 0% of trials in the final block. As in Experiment S1, all interventions were removed during the test phase.

As in Experiment S1, participants completed 128 learning trials, followed by 32 test trials with no feedback and in which the interventions were removed.

After the test phase, we asked “About what percentage of applicants do you think were unsuitable?” and “Which fields do you think were useful in deciding whether an applicant was suitable or unsuitable?" as in Experiment S1. We added a third post-task question asking participants whether they believed they had learned completely how to use the applicant features to determine which applicants were suitable. Participants chose from a drop-down list either "I think I learned completely how the features determined suitability", "I think there may have been aspects of applicant suitability that I did not learn", or "I think there were definitely aspects of applicant suitability that I did not learn."

B.2.2 Results

Figure B.5 shows participants' behavior over the eight blocks of learning and the test phase. As in the Results section of Experiment S1, we first report on the replication of the contingent-information and full-information conditions, including choice behavior and post-task questions, and then proceed to examine the effectiveness of the interventions.

Posterior estimates of quantities of interest are plotted in Figure B.6. The results of the contingent-information and full-information conditions partially replicate the previous experiments. In the test phase, the directional pattern of strategy use matched previous experiments. The average 2D score was .82 in the full-information condition, but only .76 in the contingent-information condition. The average 1D score was .73 in the full-information condition, but .84 in the contingent-information condition. However, while the credible interval for the difference in 1D scores excluded zero, $CI = [.01, .15]$, the credible interval for 2D scored did not, $CI = [-.05, .11]$. Comparing these results

B Appendix B

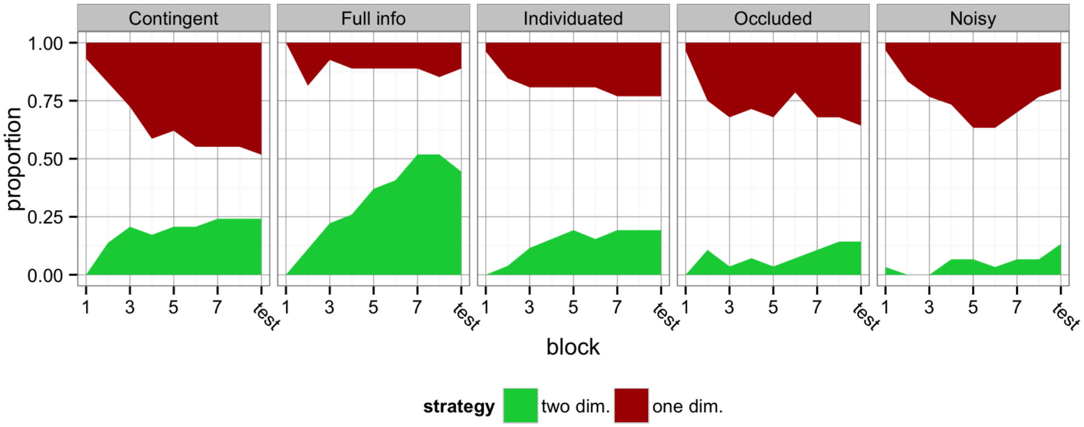


Figure B.5: Participant behavior in the Experiment S2 job application decision-making game. Panels show proportion of participants adopting the optimal two-dimensional strategy or one of the suboptimal one-dimensional learning traps in each of the 8 learning blocks and in the test phase. Participants were coded as using a two-dimensional or one-dimensional strategy if at least 15/16 of their choices in a block were consistent with that strategy.

to those of Experiment S1, it appears that in Experiment S2 there was an increase in the proportion of optimal choices in the contingent-information condition and a decrease in the full-information condition. While possibly due to noise, it may be that this discrepancy is related to our change to the use of left/right stimulus dimensions, which may have enhanced memory of individual exemplars (similar to our individuation intervention). This could have reduced the use of rigid rules, whether they are correct (in the two-dimension case) or incorrect (in the one-dimension case). Regardless, our basic finding that participants with contingent information are more susceptible to following an incomplete, one-dimensional rule is upheld in Experiment S2.

In the post-task questions, participants with contingent information estimated the proportion of unsuitable applicants to be 39.0%, while those with full information provided an average estimate of 28.2%, though the credible interval of the difference did not quite exclude zero $CI = [-.01, .17]$. Reflecting the results of the test phase, full-information participants were more likely to provide the correct dimensions but not definitively so, $CI = [-.13, .29]$, while contingent-information participants were more clearly more likely to report only a single relevant dimension, $CI = [.00, .41]$.

Few participants in the contingent and full information conditions reported that there were "definitely" aspects of applicant suitability they did not learn (4 and 3, respectively), so we pooled these participants with those who reported that there "may" have been aspects they did not learn. In the contingent-information condition, 62.0% of participants believed they "learned completely" how features determined suitability, while in the full-information condition 48.1% believed they had learned completely $CI = [-.12, .34]$. There was no large difference between the groups, suggesting that contingent-information participants did not enter the learning trap solely because it yielded “good enough” performance even while knowing that there was more to learn about the environment. Rather, it appears that participants adopted an incorrect rule even while believing they had learned successfully.

In Experiment S1, we found some evidence that the individuation intervention prevented the learning trap, but without helping participants to find the optimal two-dimensional rule. In Experiment S2, a similar pattern emerged (see Figures B.6–B.5). In the test phase, none of the interventions led to a greater proportion of responses consistent with the true two-dimensional rule. All three interventions appeared to cause reductions in responses consistent with a one-dimensional rule, though none of the credible intervals excluded zero.

In the post-task questions, participants in the intervention conditions did not strongly differ from those in the contingent-information condition in the proportion of applicants they believed were unsuitable, or in their probability of reporting both relevant dimensions or only a single dimension. But participants were less likely to report that they completely learned the determinants of suitability in the three intervention conditions, $CI = [.01, .46]$ for the individuation condition, $CI = [-.02, .43]$ for the occlusion condition (not quite excluding zero), and $CI = [.18, .62]$ for the noisy condition. Thus, it seems that the interventions succeeded in slowing the rapid learning that leads to the attentional learning trap, but this inhibited learning persisted rather than leading to an eventual convergence to the optimal strategy.

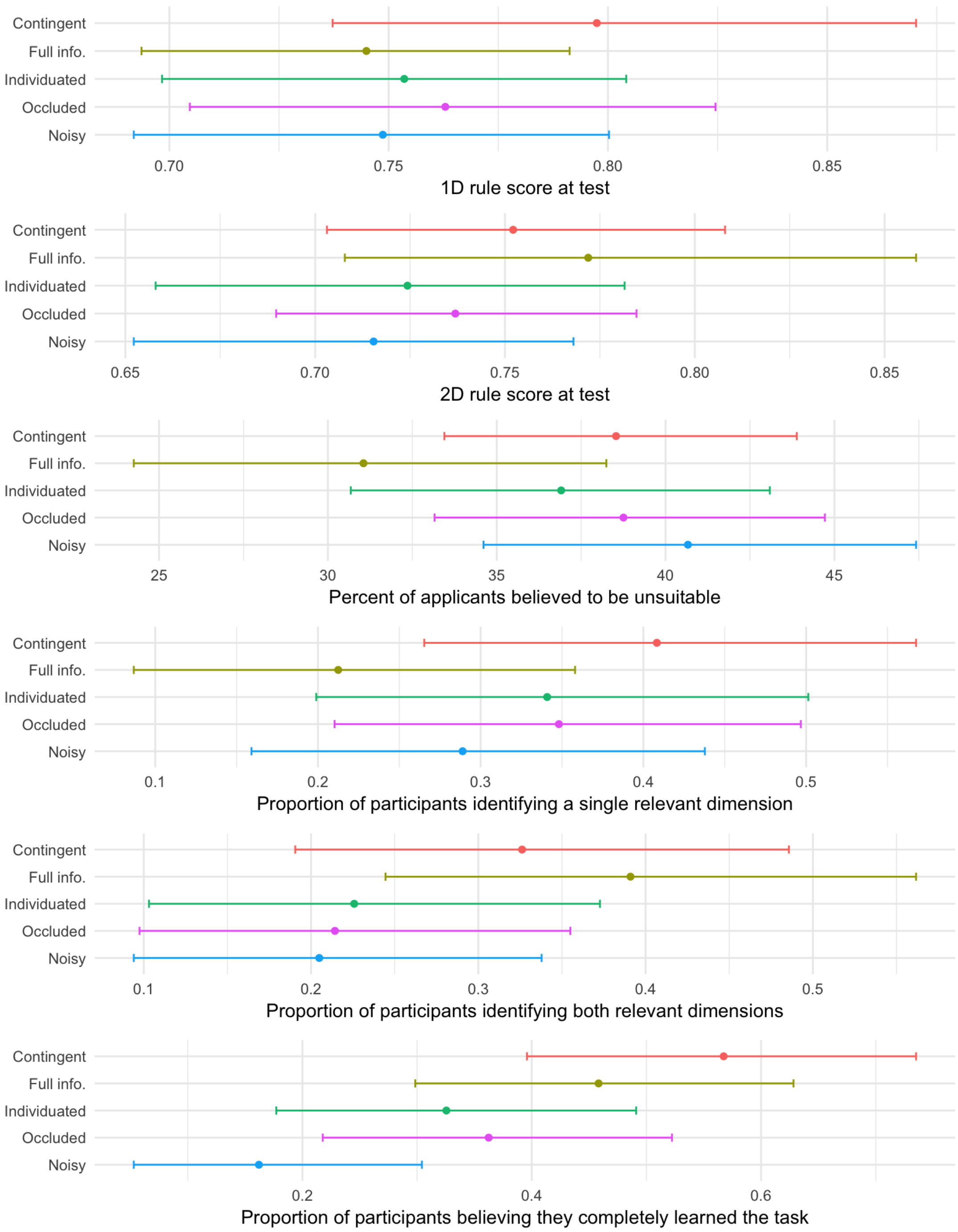


Figure B.6: Comparisons of several measures of behavior across the five conditions of Experiment S2. Points indicate posterior population mean from Bayesian inference, and error bars indicate 95% credible intervals. As with Experiment S1, the results support the conclusion that participants with contingent information fell into the attentional learning trap more readily than those with full information, but that the interventions did not lead to more robust learning compared to the contingent-information condition.

BIBLIOGRAPHY

Abramson, L. Y. [L Y], Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: critique and reformulation. *Journal of abnormal psychology*, *87*(1), 49–74.

Abramson, L. Y. [Lyn Y], Metalsky, G. I., & Alloy, L. B. (1989). Hopelessness Depression : A Theory-Based Subtype of Depression. *Psychological Review*, *96*(2), 358–372.

Aguirregabiria, V. & Mira, P. (2010). Dynamic Discrete Choice Structural Models: A Survey. *Journal of Econometrics*, *156*(1), 38–67

Allport, G. W. [Gordon W.]. (1954). *The Nature of Prejudice*. Garden City, NY: Doubleday.

Allport, G. W. [Gordon W]. (1979). *The Nature of Prejudice*. Basic Books.

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

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory*, *10*(1), 55–77.

Barocas, S. & Selbst, A. (2016). Big Data's Disparate Impact. *California law review*, *104*(1), 671–729

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*(9), 1214–1221.

Bellman, R. (1957). *Dynamic Programming* (1st ed.). Princeton, NJ, USA: Princeton University Press.

Berlyne, D. E. (1966). Curiosity and Exploration. *Science*.

Berry, D. A. & Fristedt, B. (1979). Bernoulli One-Armed Bandits–Arbitrary Discount Sequences. *The Annals of Statistics*, *7*(5), 1086–1105.

Berry, D. A. & Fristedt, B. (1985). *Bandit problems*. London: Chapman & Hall/CRC.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads , Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–215

Brown, A. L., Chua, Z. E., & Camerer, C. F. (2009). Learning and visceral temptations in dynamic saving experiments. *The Quarterly Journal of Economics*, (February), 197–231.

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. *Proceedings of Machine Learning Research*, *81*, 1–15. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Camerer, C. F. & Weigelt, K. (1996). An Asset Market Test of a Mechanism for Inducing Stochastic Horizons in Experiments. *Research in Experimental Economics*, *6*.

BIBLIOGRAPHY

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI Now 2017 Report. *AI Now Institute at New York University*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1)

Chater, N. & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Ching, A. T., Erdem, T., & Keane, M. P. (2013). Invited Paper —Learning Models: An Assessment of Progress, Challenges, and New Developments. *Marketing Science*, *32*(6), 913–938

Chintagunta, P. K., Goettler, R. L., & Kim, M. (2012). New Drug Diffusion when Forward-Looking Physicians Learn from Patient Feedback and Detailing. *Journal of Marketing Research*, *49*(December), 1–43.

Coenen, A. & Gureckis, T. M. [Todd M]. (2016). The distorting effect of deciding to stop sampling. In *Proceedings of the 38th annual conference of the cognitive science society*.

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

Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review*.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dawes, R. M. [R M], Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science (New York, N.Y.) 243*(4899), 1668–1674.

Dawes, R. M. [Robyn M]. (1993). Prediction of the Future versus an Understanding of the Past : A Basic Asymmetry. *The American Journal of Psychology*, *106*(1), 1–24.

Denrell, J. (2003). Vicarious Learning, Undersampling of Failure, and the Myths of Management. *Organization Science*, *14*(3), 227–243

Denrell, J. (2005). Why most people disapprove of me: experience sampling in impression formation. *Psychological review*, *112*(4), 951–978.

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

Denrell, J. & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological review*, *114*(2), 398.

Denrell, J. & Le Mens, G. (2011). Seeking positive experiences can produce illusory correlations. *Cognition*, *119*(3), 313–324

Denrell, J. & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538.

BIBLIOGRAPHY

Diener, C. I. & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. *Journal of Personality and Social Psychology*, *36*(5), 451–462.

Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2015). *OpenIntro Statistics* (3rd ed.).

Dougherty, C. (2015). Google Photos Mistakenly Labels Black People 'Gorillas'. Retrieved April 12, 2018, from https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/

Ebert, J. E. J. & Prelec, D. (2007). The Fragility of Time: Time-Insensitivity and Valuation of the Near and Far Future. *Management Science*, *53*(9), 1423–1438.

Einhorn, H. J. & Hogarth, R. M. [Robin M.]. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*(5), 395–416.

Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin*, *29*(10), 1221–1235

Elton, E. J., Gruber, M. J., & Blake, C. R. (1996). Survivorship Bias and Mutual Fund Performance. *The Review of Financial Studies*, *9*(4), 1097–1120.

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological science*, *18*(2), 105–110.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway Feedback Loops in Predictive Policing, 1–12. arXiv: 1706.09847. Retrieved from http://arxiv.org/abs/1706.09847

Erdem, T. & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing science*, *15*(1), 1–20.

Erev, I. (2014). Recommender Systems and Learning Traps. *Proceedings of the First International Workshop on Decision Making and Recommender Systems*, 5–8.

Evans, G. W., Gonnella, C., Marcynyszyn, L. a., Gentile, L., & Salpekar, N. (2005). The Role of Chaos in Poverty and Children's Socioemotional Adjustment. *Psychological Science*, *16*(7), 560–565.

Fang, C., Lee, J., & Schilling, M. a. (2010). Balancing Exploration and Exploitation Through Structural Design: The Isolation of Subgroups and Organizational Learning. *Organization Science*, *21*(3), 625–642.

Fazio, R. H. [Russell H], Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: valence asymmetries. *Journal of personality and social psychology*, *87*(3), 293–311.

Feiler, D. C., Tong, J. D., & Larrick, R. P. (2012). Biased Judgment in Censored Environments. *Management Science*.

Feldman, J. (2003). The Simplicity Principle in Human Concept Learning. *Current Directions in Psychological Science*, *12*(6), 227–232

Fiedler, K. [K]. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological review*, *107*(4), 659–676.

Fiedler, K. [Klaus]. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of experimental psychology. Learning, memory, and cognition*, *34*(1), 186–203.

BIBLIOGRAPHY

Fiedler, K. [Klaus], Brinkmann, B., Betsch, T., & Wild, B. (2000). A Sampling Approach to Biases in Conditional Probability Judgments : Beyond Base Rate Neglect and Statistical Format. *Journal of Experimental Psychology: General*, *129*(3).

Fiedler, K. [Klaus] & Unkelbach, C. (2014). Regressive Judgment : Implications of a Universal Property of the Empirical World. *Current Directions in Psychological Science*, *23*(5).

Fiedler, K. [Klaus], Walther, E., Freytag, P., & Plessner, H. (2002). Judgment Biases in a Simulated Classroom — A Cognitive – Environmental Approach. *Organizational Behavior and Human Decision Processes*, *88*(1), 527–561.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, *40*(2), 351–401.

Fuller, R. (1991). Behavior analysis and unsafe driving: Warning—learning trap ahead! *Journal of Applied Behavior Analysis*, *24*, 73–75.

Gelman, A. [A], Carlin, J., Stern, H., & Rubin, D. (2014). *Bayesian data analysis*. Retrieved from http://www.tandfonline.com/doi/full/10.1080/01621459.2014.963405

Gelman, A. [Andrew], Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Taylor & Francis.

Gelman, A. [Andrew], Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. arXiv: 0907.2478

Gelman, S. A. [Susan A.]. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*(1), 65–95.

Gelman, S. A. [Susan A.]. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–409.

Gelman, S. A. [Susan A] & Coley, J. D. (1990). The Importance of Knowing a Dodo Is a Bird: Categories and Inferences in 2-Year-Old Children. *Developmental Psychology*, *26*(5), 796–804.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural computation*, *58*.

Gill, D. & Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *The American Economic Review*, *102*(1), 469–503.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science: A Multidisciplinary Journal*, *32*(1), 108–154

Gopher, D., Weil, M., & Siegel, D. (1989). Practice under changing priorities: An approach to the training of complex skills. *Acta Psychologica*, *71*(1-3), 147–177.

Green, L. & Myerson, J. (2004). A Discounting Framework for Choice With Delayed and Probabilistic Rewards. *Psychological Bulletin*, *130*(5), 769–792

Grosskopf, B., Erev, I., & Yechiam, E. (2006). Foregone with the wind: Indirect payoff information and its implications for choice. *International Journal of Game Theory*, *34*, 285–302.

Guez, A., Silver, D., & Dayan, P. (2013). Scalable and efficient bayes-adaptive reinforcement learning based on Monte-Carlo tree search. *Journal of Artificial Intelligence Research*, *48*, 841–883.

Gureckis, T. M. [T. M.] & Markant, D. B. (2012, September). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science*, *7*(5), 464–481

Gureckis, T. M. [Todd M.] & Love, B. C. (2009). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, *53*(3), 180–193. arXiv: NIHMS150003

Gureckis, T. M. [Todd M], Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*.

Henriksson, M. P., Elwin, E., & Juslin, P. (2010). What is coded into memory in the absence of outcome feedback? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 1.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

Hirschfeld, L. A. (1995). Do children have a theory of race? *Cognition*, *54*, 209–252.

Hoffman, A. B. & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340

Hogarth, R. M. [R. M.], Lejarraga, T., & Soyer, E. (2015). The Two Settings of Kind and Wicked Learning Environments. *Current Directions in Psychological Science*, *24*(5), 379–385

Howard, R. A. (1966). Information Value Theory. *Systems Science and Cybernetics, IEEE Transactions on*, *2*(1), 22–26.

Huys, Q. J. M. & Dayan, P. (2009). A Bayesian formulation of behavioral control. *Cognition*, *113*(3), 314–328.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2016). Fairness in Reinforcement Learning, 1–23. arXiv: 1611.03071. Retrieved from http://arxiv.org/abs/1611.03071

Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., . . . Prince, S. E. (1996). A Component Analysis of Cognitive-Behavioral Treatment for Depression. *Journal of Consulting and Clinical Psychology*, *64*(2), 295–304.

Jones, M. & Cañas, F. (2010). Integrating Reinforcement Learning with Models of Representation Learning. (4).

Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2016). Fairness in Learning: Classic and Contextual Bandits. *Advances in Neural Information Processing Systems*. arXiv: 1605.07139

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). *Simple Rules for Complex Decisions*.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2016). Information Sampling Behavior With Explicit Sampling Costs. *Decision*.

Kidd, C. & Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. *Neuron*, *88*(3), 449–460

Kirby, K. N. & Herrnstein, R. (1995). Preference Reversals Due To Myopic Discounting of Delayed Reward. *Psychological Science*, *6*(2), 83–89.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680.

Kleinberg, J., Ludwig, J., Cohen, M., Crohn, A., Cusick, G. R., Dierks, T., … John, J. (2017). Human Decisions and Machine Predictions. *NBER Working Paper*.

Knox, W. B., Otto, a. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *3*(JAN), 1–12.

Koehler, J. J. & Mercer, M. (2009). Selection Neglect in Mutual Fund Advertisements. *Management Science*, *55*(7), 1107–1121.

Kreps, D. M. & Porteus, E. L. (1978). Temporal Resolution of Uncertainty and Dynamic Choice Theory. *Econometrica*, *46*(1), 185–200.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44

Kruschke, J. K. (1996, June). Dimensional Relevance Shifts in Category Learning. *Connection Science*, *8*(2), 225–248

Kruschke, J. K. & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*, *112*(2), 443–447.

Le Mens, G. & Denrell, J. (2011). Rational learning and information sampling: on the "naivety" assumption in sampling explanations of judgment biases. *Psychological review*, *118*(2), 379–392.

Le Mens, G., Kareev, Y., & Avrahami, J. (2016). The Evaluative Advantage of Novel Alternatives: An Information-Sampling Account. *Psychological Science*, *27*(2), 161–168

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174.

Lempert, K. M. & Phelps, E. A. (2015). The Malleability of Intertemporal Choice. *Trends in Cognitive Sciences*, *20*(1), 64–74

Lerman, J. (2013). Big Data and its Exclusions. *Stanford Law Review*, *66*, 55–63.

Levinthal, D. A. & March, J. G. (1993). The myopia of learning. *Strategic Management Journal*, *14*, 95–112.

Liu, C., Eubanks, D. L., & Chater, N. (2015). The weakness of strong ties: Sampling bias, social ties, and nepotism in family business succession. *The Leadership Quarterly*

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandel, D., & Parkes, D. C. (2017). Calibrated Fairness in Bandits. In *Fairness, accountability, and transparency in machine learning*.

Loewenstein, G. (1987). Anticipation and the Valuation of Delayed Consumption*. *The Economic Journal*, *97*(387), 666–684.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98

Love, B. C., Medin, D. L. [Douglas L.], & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological review*, *111*(2), 309–332.

Love, B. C. & Otto, A. R. (2010). You Don't Want To Know What You're Missing: When Information about Forgone Rewards Impedes Dynamic Decision Making. *Judgment and Decision Making*, *5*(1), 1–10.

Luce, R. D. (1959). *Individual Choice Behavior: a Theoretical Analysis*. John Wiley and sons.

Lum, K. & Isaac, W. (2016). To predict and serve? *Significance*

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298

Maier, S. F. & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General*, *105*(1), 3–46.

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, *2*(1), 71–87.

Mcclure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time Discounting for Primary Rewards. *27*(21), 5796–5804.

Medin, D. L. [Douglas L.], Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 607–625.

Medin, D. L. [Douglas L.] & Ortony, A. (1989). Psychological Essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge, MA: Cambridge University Press.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., … Fiedler, K. (2015). Unpacking the Exploration – Exploitation Tradeoff : A Synthesis of Human and Animal Literatures. *Decision*, *2*(3), 191–215.

Meyer, R. J. & Shi, Y. (1995). Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed-Bandit Problem. *Management Science*, *41*(5), 817–834.

Millar, A. & Navarick, D. J. (1984). Self-control and choice in humans: Effects of video game playing as a positive reinforcer. *Learning and Motivation*, *15*(2), 203–218.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, *110*(512), 1399–1411. arXiv: arXiv:1011.1669v3

Murphy, G. L. & Medin, D. L. [Douglas L]. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3). Retrieved from http://psycnet.apa.org/journals/rev/92/3/289/

Myerson, J. & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the experimental analysis of behavior*, *64*(3), 263–276.

Navarick, D. (1998). Impulsive choice in adults: How consistent are individual differences? *The Psychological Record*, *48*, 665–674.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world : An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77.

Navarro, D. J. & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological review*, *118*(1), 120–134.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

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.

Nosofsky, R. M. (1986). Attention , Similarity , and the Identification-Categorization Relationship. *Journal of experimental psychology. General*, *115*(1), 39–57.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & cognition*, *22*(3), 352–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Otto, a. R., Markman, a. B., & Love, B. C. (2012). Taking More, Now: The Optimality of Impulsive Choice Hinges on Environment Structure. *Social Psychological and Personality Science*, *3*(2), 131–138.

Oudeyer, P. Y. & Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *3*(NOV), 1–14.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774.

Payzan-LeNestour, E. & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology*, *7*(1), e1001048.

Rehder, B. & Hastie, R. (2001). Causal Knowledge and Categories: The Effects of Causal Beliefs on Categorization, Induction, and Similarity.

Rich, A. S. & Gureckis, T. M. [Todd M.]. (2017). Exploratory choice reflects the future value of information. *Decision*, *(in press)*.

Riefer, P. S., Prior, R., Blair, N., Pavey, G., & Love, B. C. (2017). Coherency Maximizing Exploration in the Supermarket. *Nature Human Behaviour*, *1*, 1–12.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies*, *4*(2), 155–161.

Sang, K., Todd, P. M., & Goldstone, R. L. (2011). Learning near-optimal search in a minimal explore/exploit task. *Proceedings of the Thirty-third Annual Conference of the Cognitive Science Society*, 2800–2805.

Schwartz, B. (1982). Reinforcement-induced behavioral stereotypy: How not to teach people to discover rules. *Journal of Experimental Psychology: General*, *111*(1), 23–59.

Seale, D. A. & Rapoport, A. (2000). Optimal stopping behavior with relative ranks: the secretary problem with unknown population size. *Journal of Behavioral Decision Making*, *13*(4), 391–411.

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

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(9), 1689–1699. arXiv: arXiv:1011.1669v3

Shook, N. J. & Fazio, R. H. [Russell H.]. (2008). Interracial roommate relationships: An experimental field test of the contact hypothesis: Research article. *Psychological Science*, *19*(7), 717–723.

Simon, H. A. (1990). Invariants of Human Behavior. *Annual review of psychology*, *41*, 1–19. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.ms.10.080180.000245

Singh, S., Barto, A., & Chentanez, N. (2004). Intrinsically motivated reinforcement learning. *18th Annual Conference on Neural Information Processing Systems (NIPS)*, *17*(2), 1281–1288.

Solnick, J. V., Kannenberg, C. H., Eckerman, D. A., & Waller, M. B. (1980). An experimental analysis of impulsivity and impulse control in humans. *Learning and Motivation*, *11*(1), 61–77.

Speekenbrink, M. & Konstantinidis, E. (2014). Uncertainty and exploration in a restless bandit task. *Proceedings of the 36rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society*.

Speekenbrink, M. & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit task. *Topics in cognitive science*, *7*.

Stan Development Team. (2015). Stan: A C++ Library for Probability and Sampling, Version 2.7. Retrieved from http://mc-stan.org/

Stange, K. M. (2012). An empirical investigation of the option value of college enrollment. *American Economic Journal: Applied Economics*, *4*(1), 49–84.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge Univ Press.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest*, *1*(1).

Taylor, E. G. & Ross, B. H. (2009). Classifying partial exemplars: seeing less and learning more. *Journal of experimental psychology. Learning, memory, and cognition*, *35*(5), 1374–1380.

Taylor, M. G. (1996). The Development of Children's Beliefs about Social and Biological Aspects of Gender Differences. *Child Development*, *67*.

Taylor, M. G., Rhodes, M., & Gelman, S. A. (2009). Boys will be boys; cows will be cows: Children's essentialist reasoning about gender categories and animal species. *Child Development*, *80*(2), 461–481.

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *The Behavioral and brain sciences*, *24*(4), 629–640, 629–640.

Teodorescu, K. & Erev, I. (2014a). Learned Helplessness and Learned Prevalence: Exploring the Causal Relations Among Perceived Controllability, Reward Prevalence, and Exploration. *Psychological Science*, *25*(10), 1861–1869.

Teodorescu, K. & Erev, I. (2014b). On the Decision to Explore New Alternatives: The Coexistence of Under- and Over-exploration. *Journal of Behavioral Decision Making*, *27*, ‘

Tolman, E. C. (1948). Cognitive Maps in Rats and Man. *Psychological Review*, *55*(4), 189–208.

Tversky, A. (1977). Features of Similarity. *Psychological Review*, *84*(4).

Tversky, A. & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680.

Tversky, A. & Kahneman, D. (1971). Belief in the Law of Small Numbers. *Psychological Bulletin*, *76*(2), 105–110.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, *38*(4), 599–637

Wainer, H. (2005). *Graphic discover: A trout in the milk and other visual adventures*. Princeton University Press.

Waxman, S. R. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development*, *5*(2), 123–150.

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*.

Wolpert, D. H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, *8*(7), 1341–1390

Wulff, D. U., Hills, T. T., & Hertwig, R. (2015). How short- and long-run aspirations impact search and choice in decisions from experience. *Cognition*, *144*, 29–37

Yechiam, E. [E], Erev, I., & Gopher, D. (2001). On the potential value and limitations of emphasis change and other exploration-enhancing training methods. *Journal of experimental psychology. Applied*, *7*(4), 277–285.

Yechiam, E. [Eldad] & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, *19*, 1–16.

Zhang, S. & Yu, A. J. (2013). Forgetful Bayes and myopic planning : Human learning and decision-making in a bandit setting. *Advances in neural information processing systems*, *26*, 2607–2615.

Zwick, R., Rapoport, A., Lo, A. K. C., & Muthukrishnan, a. V. (2003). Consumer Sequential Search: Not Enough or Too Much? *Marketing Science*, *22*(4), 503–519.