UNIVERSITY OF CALIFORNIA,
IRVINE

Rational structure inference in learning and across development

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Nora Harhen

Dissertation Committee:
Professor Aaron Bornstein, Chair
Professor Catherine Hartley
Professor Nadia Chernyak

2024

DEDICATION

To my family and friends

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

Completing a PhD is equally an emotional journey as it is an intellectual one. I am immensely grateful for the friends, family, and mentors who helped me navigate both journeys.

I thank Aaron Bornstein for believing in my ideas (even when I did not) and fostering their development from inchoate thoughts into full-fledged projects through many multi-hour long meetings.

I thank Cate Hartley for warmly welcoming me into her lab as a visiting grad student. I aspire to one day approach your level of intellectual adventurousness and kindness. I guess that's what the post doc is for.

I thank Anne Collins, my undergraduate thesis advisor, for introducing me to the joys of computational modeling despite my initial trepidation. I will always treasure the clarity of thought that formalization brings.

I thank all my brilliant and hilarious labmates from both the Bornstein and Hartley labs (and in no particular order): Nidhi Banavar, Jungsun Yoo, Ari Khoudary, Sharon Noh, Mohit Nadkarni, Dale Zhou, Alexa Booras, Bianca Leonard, Yifei Chen, Kate Nussenbaum, Ali Cohen, Noam Goldway, Susan Benear, Hanxiao Lu, Rheza Budiono, Alice Zhang, Alejandra Martínez, Julie Lee, Naiti Bhatt, and Greer Bizzell-Hatcher. Thank you for making science fun.

I thank my grad school friends for keeping me cackling for all five years and offering support the many times I needed it. Priyam Das, for the many baking triumphs and failures and for the long conversations on the car ride between Orange County and the Bay. Aakriti Kumar, for being the best roommate anyone could ask for. Lauren Montgomery, for your infinite kindness, calmness, and stash of tea. Adriana Felisa Chávez De la Peña, for your infectious laugh and spot-on insights. Jaime Islas Farias, for your amazing facial expressions and cooking. And most of all, Nidhi Banavar, for making me laugh so hard that I cried and for sharing in the many difficulties and joys that come with grad school.

I thank my friends Olivia, Jess, and Paola for constantly reminding me that academia is a weird bubble, and it's good to experience the real world sometimes.

I thank my partner Fred Callaway for his love, patience, goofiness, and optimism. I think this is possibly the most romantic thing one scientist can say to another: without a doubt, you are the person I most enjoy discussing science with.

I thank Karen and Dan for taking me in as a family member and housing and feeding me for much of the pandemic.

I thank my sister Brenda for always being outraged at the same things as me.

And, I thank my parents for supporting and loving me through this emotional rollercoaster.

VITA

Nora Harhen

EDUCATION

Doctor of Philosophy in Cognitive Neuroscience **2019–2024**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Cognitive Science **2014–2018**
University of California, Berkeley *Berkeley, CA*

RESEARCH EXPERIENCE

Visiting Graduate Student **Fall 2021, 2022, 2023**
New York University *New York, NY*

Neuroplasticity & Development Lab Manager **2018–2019**
Johns Hopkins University *Baltimore, MD*

TEACHING EXPERIENCE

Psych 111/112 Teaching Assistant **2019–2020**
University of California, Irvine *Irvine, CA*

Letters & Sciences 22 Teaching Assistant **2016**
University of California, Berkeley *Berkeley, CA*

REFEREED JOURNAL PUBLICATIONS

Harhen, N.C., Bornstein A.M. Interval timing as a computational pathway from early life adversity to affective disorders. *Topics in Cognitive Science* (2024).

Harhen, N.C., Bornstein A.M. Overharvesting in human patch foraging reflects rational structure learning and adaptive planning. *Proceedings of the National Academy of Sciences* (2023).

Arcos, K.*, **Harhen, N.***, Loiotile, R., Bedny, M. Superior verbal but not nonverbal memory in congenital blindness. *Exp Brain Res* (2022). https://doi.org/10.1007/s00221-021-06304-4

Loiotile, R., Kanjlia, S., **Harhen, N.**, Bedny, M. ”Visual” cortices of congenitally blind adults are sensitive to response selection demands in a go/no-go task. *Neuroimage* (2021). https://doi.org/10.1016/j.neuroimage.2021.118023.

REFEREED CONFERENCE PUBLICATIONS

Harhen, N.C., Bornstein A.M. Learning to expect change: Volatility during early experience alters reward expectations in a model of interval timing. *Proceedings of the 20th International Conference on Cognitive Modeling* (2022). Selected as one of the best papers of *ICCM.*

Harhen, N.C., Bornstein A.M. Humans adapt their foraging strategies and computations to environment complexity. *Proceedings of the 5th Multidisciplinary Conference on Reinforcement Learning and Decision Making* (2022).

Harhen, N.C., Bornstein A.M. Structure learning as a mechanism of overharvesting. *Proceedings of the 19th International Conference on Cognitive Modeling* (2021).

Harhen, N.C., Hartley, C.A, Bornstein, A.M. Model-based foraging using latent-cause inference. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (2021).

SERVICE

Application Statement Feedback Program **2021-present**
Editor

UCI Cognitive Sciences Colloquium Organizing Committee **2021-2022**
Student Organizer

Competitive Edge **2020**
Peer Mentor

AWARDS

F31 Ruth L. Kirschstein National Research Service Award, NIMH	2023-2024
Memory, Space, & Time Workshop Travel Award	2022
Sloan-Nomis Cognitive Foundations of Economic Behavior Summer School	2022
Reinforcement Learning & Decision-making Conference Travel Award	2022
National Defense Science & Engineering Graduate Fellowship	2020-2023
Robert J. Glushko Prize for Outstanding Undergraduate Research	2018
Summer Undergraduate Research Fellowship	2017

ABSTRACT OF THE DISSERTATION

Rational structure inference in learning and across development

By

Nora Harhen

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2024

Professor Aaron Bornstein, Chair

Humans often fail to accord with the predictions of optimal decision making models. How can we be such poor decision makers in simple task environments and yet also be deft navigators of the real world? This dissertation presents work suggesting that these two observations are likely related: Our decision making strategies have been shaped by the complexity and uncertainty of real-world decision problems, and we apply these strategies even in much simpler decision contexts. In each of the dissertation's chapters, we consider a presumed suboptimal behavior and demonstrate how it can emerge from rational responses to uncertainty. In chapters 2 and 3, we focus on the case of patch foraging. We show that both adults' over-exploitation and children's over-exploration can stem from rational inference of the environment's underlying structure. Our results reveal that these two opposing behaviors emerge from different structural priors. Finally, in chapter 4, we focus on the reward learning deficits that characterize anhedonia. Using a reinforcement learning model, we show that simulated agents who have rationally adapted to an unpredictable early life environment produce anhedonia-like behavior when later placed in a predictable environment. Collectively, this work demonstrates how multi-scale learning processes work to mitigate the many forms of uncertainty present in real-world decisions. By taking these learning processes into account, we are able to rationalize multiple "suboptimal" behaviors.

Chapter 1

Introduction

> "Evolution strikes me as infinitely more spiritually profound than Genesis."
>
> Maggie Nelson, *The Argonauts*

How does the mind acquire its form? Cognitive scientists have historically turned to the surrounding environment for answers. Illustrating why, Roger Shepard, famously likened the mind to a "mirror" reflecting the world's invariant features [174]. Underlying his metaphor is an insight that underpins rational approaches to modeling cognition: the environment poses problems to the mind, and to solve them, the mind internalizes the environment. Rational frameworks, exemplified by John Anderson's rational analysis [8] and David Marr's computational level of analysis [127], take the problem posed by the environment and derive the problem's optimal solution to obtain a model of a cognitive process. What is unique about a rational model is that it provides insights into the purpose of a particular cognitive process. Put differently, it tells us *why* the process occurs the way it does.

Most rational models start with the working assumption that the mind is *already* well-adapted to the environment. To ensure this assumption is met, many decision making tasks provide participants with full knowledge of the task environment [76, 132]. And yet, even with an abundance of information, participants continue to violate the predictions of optimal models of decision making [196]. How is it possible that we can fail at such simple choice tasks and at the same time, thrive in the real world?

While the decision making literature suggests we are poor decision makers, the learning literature paints a more optimistic picture, suggesting we are exceptional learners. Humans excel at uncovering the hidden structure of novel environments [71] and using this structure knowledge to learn from unobserved, counterfactual outcomes [52, 37]. Sometimes the desire to find structure is so great, that learners occasionally "see" structure where there is, in fact, only noise [208]. Further demonstrating humans' impressive learning abilities, they flexibly modulate their learning in response to environmental changes, in ways consistent with theoretical predictions of rational learning models [17, 136, 204, 109, 138]. Perhaps most strikingly and pertinently for our work, similar models have been used to demonstrate the normative advantages of our noisy and sometimes error prone learning processes [2, 141, 207].

How can we reconcile our poor decision making with our excellent learning? To gain traction on this question, we begin by reflecting on how the decision problems encountered in task environments differ from those found in real-world environments. Task environments tend to be simple and static, whereas real world environments are complex and dynamic. Thus, behavior that seems suboptimal in the context of a single task may actually reflect rational adaptation to the agent's previous environment, and perhaps even to the entire distribution of environments the agent has experienced across their lifetime.

The work in this dissertation uses rational learning models to re-frame seemingly "irrational" decisions. Central to our approach is the recognition that our experiences in past environments shapes learning in the current environment. Even when task environments are much simpler, we transfer over learning and decision making strategies that are adaptive in real-world environments (Chapters 2 and 3). Moreover, we demonstrate particular persistence in using strategies that were adaptive early in life. (Chapter 4). Unlike many optimal models, we do not start with the assumption of already being well-adapted to the environment. We instead model how it might rationally occur over multiple timescales – within-task and across development. This work suggests that the mind should be well-adapted not to a single environment, but many.

In Chapter 2, we use a Bayesian structure learning model to explain the widely observed suboptimality of "overharvesting" in patch foraging. Foragers, from rodents to humans, stay in patches of resources longer than the optimal decision rule prescribes. Importantly, this optimal model assumes the forager has perfect knowledge of their environment. By relaxing this assumption and replacing it with the assumption that foragers are Bayes-optimal structure learner, we show that even a Bayes-optimal learner can overharvest. They specifically do so when they expect the environment to be more complex than the simple environment they end up in. We test the critical predictions of the model against participants' behavior in a novel patch foraging task and find that our model provides a superior explanation of participants' adaptation to the richness and dynamics of the environment, relative to the optimal policy and other comparison models, .

In Chapter 3, we demonstrate that the model from Chapter 2 can also explain the overexploration of children and adolescents. In the patch foraging task presented in Chapter 2, we find that children and adolescents are more exploratory than adults, acting in greater alignment with the optimal decision policy. Our model proposes that this behavioral difference stems from a difference in structural priors – children and adolescents expect environments to be simpler.

In Chapter 4, we look over a longer timescale of adaptation – development. We examine how early life environments' persistent role in shaping how we learn and act can give rise to behaviors that are considered maladaptive. In this chapter, we focus on the deficits in reward learning which characterize anhedonia. We use a reinforcement learning model that models how an agent internalizes their environment. We demonstrate that a simulated agent exposed to an unpredictable environment, during a period of heightened plasticity, will develop internal representations that produce anhedonia-like behavior in predictable environments.

By introducing models that take into account learning processes as they unfold over multiple time scales, we not only provide better explanations of participants' behavior, but we are also able rationalize it. Through widening the scope of rational modeling to more expansive timescales, we extend its set of possible use cases to include development. We can begin to ask questions such as: why does the developmental trajectory take this form?

Chapter 2

Overharvesting in human patch foraging reflects rational structure learning and adaptive planning

The contents of this chapter were published in Harhen, N.C., Bornstein A.M. Overharvesting in human patch foraging reflects rational structure learning and adaptive planning. *Proceedings of the National Academy of Sciences* (2023).

2.1 Introduction

Many real world decisions are sequential in nature. Rather than selecting from a set of known options, a decision-maker must choose between accepting a current option or rejecting it for a potentially better future alternative. Such decisions arise in a variety of contexts including choosing an apartment to rent, a job to accept, or a website to browse. In ethology, these decisions are known as patch leaving problems. Optimal foraging theory suggests that the current option should be compared to the quality of the overall environment [132]. An agent using the optimal choice rule given by Marginal Value Theorem (MVT [35]) will leave once the local reward rate of the current patch, or concentration of resources, drops below the global reward rate of the environment.

Foragers largely abide by the qualitative predictions of MVT, but deviate quantitatively in systematic ways - staying longer in a patch relative to MVT's prescription. Known as overharvesting, this bias to overstay is widely observed across organisms [40, 91, 101, 140**?** , 173, 203, 34]. Despite this, how and why it occurs remains unclear. Proposed mechanisms include a sensitivity to sunk costs [203, 34], diminishing marginal utility [40], discounting of future rewards [25, 34, 40], and underestimation of post-reward delays [101]. Critically, these all share MVT's assumption that the forager has accurate and complete knowledge of their environment, implying that deviations from MVT optimality emerge in spite of this knowledge. However, an assumption of accurate and complete knowledge often fails to be met in dynamic real world environments [100]. Relaxing this assumption, how might foragers learn the quality of the local and global environment?

Previously proposed learning rules include recency-weighted averaging over all previous experiences [40, 67] and Bayesian updating [105]. In this prior work, learning of environment *quality* is foregrounded while knowledge of environment *structure* is assumed. In a homogeneous environment, as is nearly universally employed in these experiments, this is a reasonable assumption as a single experience in a patch can be broadly generalized from across other patches. However, it may be less reasonable in more naturalistic heterogeneous environments with regional variation in richness. To make accurate predictions within a local patch, the forager must learn the heterogeneous structure of the broader environment. How might they rationally do so?

Prior work has found that humans act in accordance with rational statistical inference of environment structure [2, 179, 39]. Here, we build on this work and extend it to a foraging context. We show that apparent overharvesting in these tasks can be explained by combining structure learning [165, 71] with adaptive planning, a combination of mechanisms with potentially broad applications to many complex behaviors performed by humans, animals, and artificial agents [107].

We tested the model's predictions with a novel variant of a serial stay-switch task (Fig. 3.1A; [40]). Participants visited different planets to mine for "space treasure" and were tasked to collect as much space treasure as possible over the course of a fixed length game. TOn each trial, they had to decide between staying on the current planet to dig from a depleting treasure mine or traveling to a new planet with a replenished mine at the cost of a time delay. To mimic naturalistic environments, we varied planet richness across the broader environment while locally correlating richness in time. More concretely, planet richness was drawn from a multimodal distribution (Fig. 3.1B) and transitions between planets of a similar richness were more likely (Fig. 3.1C). Our model predicted distinct behavioral patterns from structure learning individuals versus their non-structure learning counterparts in our task. Specifically, within the multimodal environment, non-structure learners are predicted to underharvest on average, while structure learners overharvest. Furthermore, structure learners' extent of overharvesting are predicted to vary across the task — decreasing with experience and increasing following rare transitions between planets. In contrast, non-structure learners should consistently underharvest.

We found that principled inference of environment structure and adaptation to this structure can 1) produce key deviations from MVT that have been widely observed in participant data across species and 2) capture patterns of behavior in a novel patch foraging task that cannot be explained by previously proposed models. Taken together, these results reinterpret overharvesting: Rather than reflecting irrational choice under a fixed representation of the environment, it can be seen as rational choice under a dynamic representation.

2.2 Methods

2.2.1 Participants

We recruited 176 participants from Amazon Mechanical Turk (111 male, ages 23-64, Mean=39.79, SD=10.56). Participation was restricted to workers who had completed at least 100 prior studies and had at least a 99% approval rate. This study was approved by the institutional review board of the University of California, Irvine, under Institutional Review Board (IRB) Protocol 2019-5110 ("Decision-making in time"). All participants gave informed consent in advance. Participants earned $6 as a base payment and could earn a bonus contingent on performance ($0-$4). We excluded 60 participants according to one or more of three criteria: 1. having average planet residence times 2 standard deviations above or below the group mean (36 participants) 2. failing a quiz on the task instructions more than 2 times (33 participants) or 3. failing to respond appropriately to one or more of the two catch trials (17 participants). On catch trials, participants were asked to press the letter "Z" on their keyboard. These questions were meant to "catch" any participants repeatedly choosing the same option (using key presses "A" or "L") independent of value.

2.2.2 Task

Participants completed a serial stay-switch task adapted from previous human foraging studies [40, 117]. With the goal of collecting as much space treasure as possible, participants traveled to different planets to mine for gems. Upon arrival to a new planet, they performed an initial dig and received an amount of gems sampled from a Gaussian distribution with a mean of 100 and standard deviation (SD) of 5. Following this initial dig, participants had to decide between staying on the current planet to dig again or leaving to travel to a new planet (Fig 3.1A). Staying would further deplete the gem mine while leaving yielded a replenished gem mine at the cost of a longer time delay. They made these decisions in a series of five blocks, each with a fixed length of 6 minutes. Blocks were separated by a break of participant-controlled length, up to a maximum of 1 minute.

On each trial, participants had 2 seconds to decide via key press whether to stay ("A") or leave ("L"). If they decided to stay, they experienced a short delay before the gem amount was displayed (1.5 s). The length of the delay was determined by the time the participant spent making their previous choice (2 - RT s). This ensured participants could not affect the environment reward rate via their response time. If they decided to leave, they encountered a longer time delay (10 s) after which they arrived on a new planet and were greeted by a new alien (5 s). On trials where a decision was not made within the allotted time (2 s), participants were shown a timeout message for two seconds.

Unlike previous variants of this task, planets varied in their richness within and across blocks, introducing greater structure to the task environment. Richness was determined by the rate at which the gem amount exponentially decayed with each successive dig (Fig. 3.1B). If a planet was "poor", there was steep depletion in the amount of gems received. Specifically, its decay rates were sampled from a beta distribution with a low mean (mean = 0.2; sd = 0.05; $\alpha = 13$ and $\beta = 51$). In contrast, rich planets depleted more slowly (mean = 0.8; sd = 0.05; $\alpha = 50$ and $\beta = 12$). Finally, the quality of the third planet type — neutral — fell in between rich and poor (mean = 0.5; sd = 0.05; $\alpha = 50$ and $\beta = 50$). The environment dynamics were designed such that planet richness was correlated in time. When traveling to a new planet, there was an 80% probability of it being the same type as the prior planet ("no switch"). If not of the same type, it was equally likely to be of one of the remaining two types ("switch", Fig. 3.1C). This information was not communicated to participants, requiring them to infer the environment's structure and dynamics from rewards received alone.

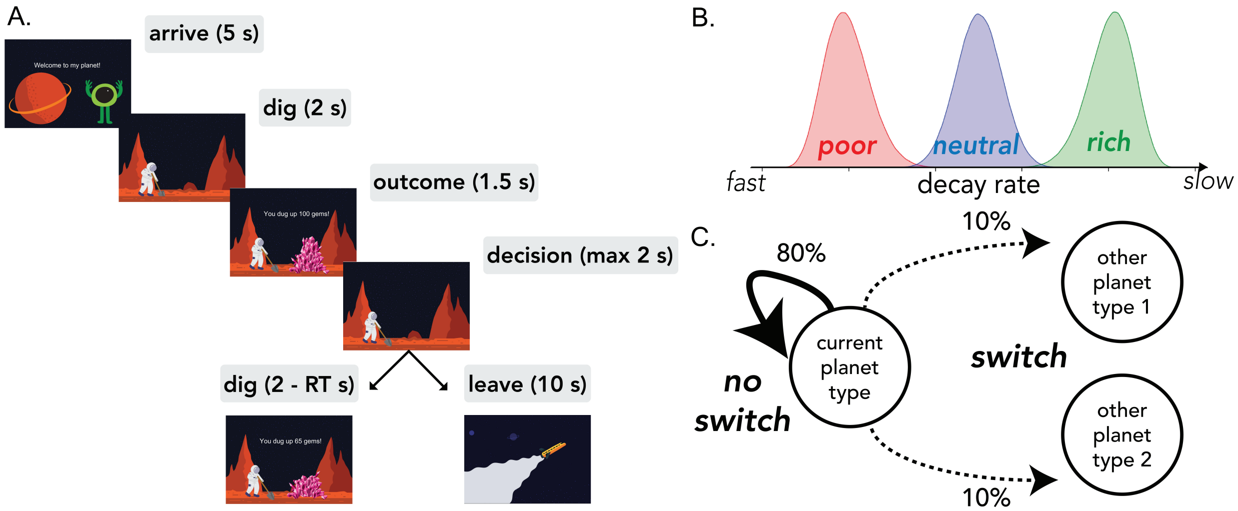

Figure 2.1: A. Serial stay-switch task. Participants traveled to different planets and mined for space gems across 5 6-minute blocks. On each trial, they had to decide between staying to dig from a depleting gem mine or incurring a time cost to travel to a new planet. **B. Environment structure.** Planets varied in their richness or, more specifically, the rate at which they exponentially decayed with each dig. There were three planet types — poor, neutral, and rich — each with their own characteristic distribution over decay rates. **C. Environment dynamics.** Planets of a similar type clustered together. A new planet had an 80% probability of being the same type as the prior planet ("no switch"). However, there was a 20% probability of transitioning or "switching" to a planet of a different type.

2.2.3 Marginal Value Theorem

Participants' planet residence times, or PRTs, were compared to those prescribed by MVT. Under MVT, agents are generally assumed to act as though they have accurate and complete knowledge of the environment. For this task, that would include knowing each planet type's unique decay rate distribution and the total reward received and time elapsed across the environment.

Knowledge of the decay rate distributions is critical for estimating V_{stay}, the anticipated reward if the agent were to stay and dig again.

$$V_{stay} = r_t * d \tag{2.1}$$

where r_t is the reward received on the last dig and d is the upcoming decay.

$$d = \begin{cases} 0.2 & \text{if planet is poor} \\ 0.5 & \text{if planet is neutral} \\ 0.8 & \text{if planet is rich} \end{cases}$$

V_{leave} is estimated using the total reward accumulated, r_{total}, total time passed in the environment, t_{total}, and the time delay to reward associated with staying and digging, t_{dig}.

$$V_{leave} = \frac{r_{total}}{t_{total}} * t_{dig} \tag{2.2}$$

$\frac{r_{total}}{t_{total}}$ estimates the average reward rate of the environment. Multiplying it by t_{dig} gives the opportunity cost of the time spent exploiting the current planet.

Finally, to make a decision, the MVT agent compares the two values and acts greedily, always taking the higher valued option.

$$\text{choice} = \text{argmax}(V_{stay}, V_{leave}) \tag{2.3}$$

2.2.4 Structure Learning & Uncertainty Adaptive Discounting Model

Making the stay-leave decisions

We assume that the forager compares the value for staying, V_{stay}, to the value of leaving V_{leave}, to make their decision. Similar to MVT, we assume foragers act greedily with respect to these values.

Learning the structure of the environment

Learning the structure of the environment affords more accurate and precise predictions which support better decision-making. Here, the forager predicts how many gems they'll receive if they stay and dig again and this determines the value of staying, V_{stay}. To generate this prediction, a forager could aggregate over all past experiences in the environment [40]. This may be reasonable in homogeneous environments but less so in heterogeneous ones where it could introduce substantial noise and uncertainty. Instead, in these varied environments, it may be more reasonable to cluster patches based on similarity and only generalize from patches belonging to the same cluster as the current one. This selectivity enables more precise predictions of future outcomes.

Clusters are latent constructs. Thus, it is not clear how many clusters a forager *should* divide past encounters into. Non-parametric Bayesian methods provide a potential solution to this problem. They allow for the complexity of the representation — as measured by the number of clusters — to grow freely as experience accumulates. These methods have been previously used to explain phenomena in category learning [83, 165], task set learning [39], fear conditioning [71], and event segmentation [176].

To initiate this clustering process, the forager must assume a model of how their observations, decay rates, are generated by the environment. The generative model we ascribe to the forager is as follows. Each planet belongs to some cluster, and each cluster is defined by a unique decay rate distribution:

$$d_k \sim Normal(\mu_k, \sigma_k) \tag{2.4}$$

where k denotes cluster number. The generative model takes the form of a *mixture model* in which normal distributions are mixed together according to some distribution $P(k)$ and observations are generated from sampling from the distribution $P(d|k)$.

Before experiencing any decay on a planet, the forager has prior expectations regarding the likelihood of a planet belonging to a certain cluster. We assume that the prior on clustering corresponds to a "Chinese restaurant process" [9]. If previous planets are clustered according to $p_{1:N}$, then for the current planet:

$$P(k) = \begin{cases} \frac{n_k}{N+\alpha} & \text{if k is old} \\ \\ \frac{\alpha}{N+\alpha} & \text{if k is new} \end{cases}$$

Where n_k is the number of planets assigned to cluster k, α is a clustering parameter, and N is the total number of planets encountered. The probability of a planet belonging to an old cluster is proportional to the number of planets already assigned to it. The probability of it belonging to a new cluster is proportional to α. Thus, α controls how dispersed the clusters are — the higher α is the more new cluster creation is encouraged. The ability to incrementally add clusters as experience warrants it makes the generative model an *infinite capacity mixture model.*

After observing successive depletions on a planet, the forager computes the posterior probability of a planet belonging to a cluster:

$$P(k|D) = \frac{P(D|k)P(k)}{\sum_{j=1}^{J} P(D|j)P(j)} \tag{2.5}$$

Where J is the number of clusters created up until the current planet, D is a vector of all the depletions observed on the current planet, and all probabilities are conditioned on prior cluster assignments of planets, $p_{1:N}$.

Exact computation of this posterior is computationally demanding as it requires tracking all possible clusterings of planets and the likelihood of the observations given those clusterings. Thus, we approximate the posterior distribution using a particle filter [60]. Each particle maintains a hypothetical clustering of planets which are weighted by the likelihood of the data under the particle's chosen clustering. All simulations and fitting were done with 1 particle which is equivalent to Anderson's local MAP algorithm [8].

With 1 particle, we assign a planet definitively to a cluster. This posterior then determines (a) which cluster's parameters are updated and (b) the inferred cluster on subsequent planet encounters.

If the planet is assigned to an old cluster, k, the existing μ_k and σ_k are updated analytically using the standard equations for computing the posterior for a normal distribution with unknown mean and variance:

$$\begin{aligned}
\bar{d} &= \frac{1}{n}\sum_{i=1}^{n} d_i \\
\mu_0' &= \frac{n_0\mu_0 + n\bar{d}}{n_0 + n} \\
n_0' &= n_0 + n \\
\nu_0' &= \nu_0 + n \\
\nu_0'\sigma_0^{2\prime} &= \nu_0\sigma_0^2 + \sum_{i=1}^{n}(d_i - \bar{d})^2 + \frac{n_0 n}{n_0 + n}(\mu_0 - \bar{d})^2
\end{aligned} \tag{2.6}$$

where d is a decay observed on the current planet, n is the total number of decays observed on the current planet, n_0 is the total number of decays observed across the environment before the current planet, μ_0 is the prior mean of the cluster-specific decay rate distribution and ν_0 is its precision. $\mu_0^{'}$ and $\nu_0^{'}$ are the posterior mean and variance respectively.

If the planet is a assigned to a new cluster, then a new cluster is initialized with the following distribution:

$$d_{new} \sim Normal(\mu = 0.5, \sigma = 0.5) \tag{2.7}$$

This initial distribution is updated with the depletions encountered on the current planet upon leaving.

The goal of this learning and inference process is to support accurate prediction. To generate a prediction of the next decay, the forager samples a cluster according to $P(k)$ or $P(k|D)$ depending on whether any depletions have been observed on the current planet. Then, a decay rate is sampled from the cluster specific distribution, d_k. The forager averages over these samples to produce the final prediction.

To demonstrate structure learning's utility for prediction, we show in simulation the predicted decay rates on each planet with structure learning (Fig. 2.2A) and without (Fig. 2.2B). With structure learning, the forager's predictions approach the mean decay rates of the true generative distributions. Without structure learning, however, the forager is persistently inaccurate, underestimating the decay rate on rich planets and overestimating it on poor planets.

Adapting the model of the environment

Because the inference process is an approximation and foragers' experience is limited, their inferred environment structure may be inaccurate. Theoretical work has suggested that a rational way to compensate for this inaccuracy is to discount future values in proportion to the agent's uncertainty over their representation of the environment[96]. We quantified an agent's uncertainty by taking the entropy of the approximated posterior distribution over clusters. We sample clusters 100 times proportional to the posterior. These samples are multinomially distributed. We represent them with the distribution, X:

$$X \sim Multinomial(100, K) \tag{2.8}$$

Where K is a vector containing the counts of clusters from sampling 100 times from the distribution, $P(k)$ or $P(k|d)$ depending on whether depletions on the planet have been observed. Uncertainty is quantified as the Shannon entropy of distribution X.

We implemented this proposal in our model by discounting the value of leaving as follows:

$$V_{leave} = \frac{r_{total}}{t_{total}} * t_{dig} * \gamma_{effective} \tag{2.9}$$

$$\gamma_{effective} = \frac{1}{1 + e^{(-\gamma_{base} + \gamma_{coef} * H(X))}} \tag{2.10}$$

where γ_{base} and γ_{coef} are free parameters and $H(X)$ is the entropy of the distribution X.

Model simulations in single patch type environments - parameter exploration

For each combination of α, γ_{coef}, and environment richness, we simulated the model 100 times, with γ_{base} held constant at 5. Decay rates in each patch in an environment were drawn from the same beta distribution. Critically, the parameters of the beta distribution varied between environments but not patches (poor - a = 13, b = 51; neutral - a = 50, b = 50; poor - a = 50, b = 12). This was done to create single patch type environments, similar to those commonly used in prior work on overharvesting [40, 91, 101, 41, 44, 45, 99]. Simulated agents' choices were compared to those that would be made if acting with an MVT policy (see *Comparison to Marginal Value Theorem*). The difference was taken between the agent's stay time in a patch and that prescribed by MVT, and these differences were averaged over to compute a a single average patch residence time (PRT) relative to MVT for each agent.

2.2.5 Model fitting

We compared participant PRTs on each planet to those predicted by the model. A model's best fitting parameters were those that minimized the difference between the true participant's and simulated agent's PRTs. We considered 1000 possible sets of parameters generated by quasi-random search using low-discrepancy Sobol sequences [184]. Prior work has demonstrated random and quasi-random search to be more efficient than grid search [21] for parameter optimization. Quasi-random search is particularly efficient with low-discrepancy sequence, more evenly covering the parameter space relative to true random search.

Because cluster assignment is a stochastic process, the predicted PRTs vary slightly with each simulation. Thus, for each candidate parameter setting, we simulated the model 50 times and averaged over the mean squared error (MSE) between participant PRTs and model-predicted PRTs for each planet. The parameter configuration that produced the lowest MSE on average was chosen as the best fitting for the individual.

2.2.6 Model Comparison

We compared three models: the structure learning and adaptive discounting model described above, a temporal difference model previously applied in a foraging context, and a MVT model that learns the mean decay rate and global reward rate of the environment.

MVT-Learning In this model, the agent learns a threshold for leaving which is determined by the global reward rate, ρ [40]. ρ is learned with a simple delta rule with α as a learning rate and taking into account the temporal delay accompanying an action τ. The value of staying is $d * r_t$ where d is the predicted decay and r_t is the reward received on the last time step. The value of leaving,V_{leave}, is the opportunity cost of the time spent digging, $\rho * t_{dig}$. The agent chooses an action using a softmax policy with temperature parameter, β which determines how precisely the agent represents the value difference between the two options.

$$
\begin{aligned}
P(a_t = dig) &= \frac{1}{(1 + e^{(-c-\beta(d*r_t-\rho*t_{dig}))})} \\
\delta_i &= \frac{r_i}{\tau_i} - \rho_t \\
\rho_{t+1} &= \rho_t + (1-(1-\alpha)^{\tau_t}) * \delta_t
\end{aligned}
\tag{2.11}
$$

TD-Learning The temporal difference (TD) agent learns a state-specific value of staying and digging, $Q(s, dig)$ and a non-state specific value of leaving, $Q(leave)$. The state, s is defined by the gem amounts offered on each dig. The state space is defined by binning the possible gems that could be earned from each dig. The bins are spaced are according to $log(b_{j+1})$ - $log(b_j) = log(\bar{k})$ where b_{j+1} and b_j are the upper and lower bounds of the bins and $\bar{d}$ is the mean decay rate. This state space specification is taken from [40]. We set b_{j+1} to 135 and b_j to 0 as these were the true bounds on gems received per dig. We set $\bar{k}$ to 0.5 because this would be the mean decay rate if one were to average the depletions experienced over all planets. The agent compares the two values and makes their choice using a softmax policy.

$$
\begin{aligned}
P(a_t = dig) &= \frac{1}{(1 + e^{(-c-\beta(Q_t(s_t,dig)-Q_t(leave)))})} \\
D_t &\sim Bernoulli(P(a_t)) \\
\delta_t &= r_t + \gamma^{\tau_t}(D_t * Q_t(s_t) + (1-D_t) * Q_t(leave)) - Q_t(s_{t-1}, a_{t-1}) \\
Q_{t+1}(s_{t-1}, a_{t-1}) &= Q_t(s_{t-1}, a_{t-1}) + \alpha * \delta_t
\end{aligned}
\tag{2.12}
$$

where c, α, β, γ are free parameters and t is the current time step. c is a perseveration term, α is the learning rate, β is the softmax temperature, and γ is the temporal discounting factor.

Cross Validation Each model's fit to the data was evaluated using a 10-fold cross validation procedure. For each participant, we shuffled their PRTs on all visited planets and split them into 10 separate training/test datasets. The best fitting parameters were those that minimized the sum of squared error (SSE) between the participant's PRT and the model's predicted PRT on each planet in the training set. Then, with the held out test dataset, the model was simulated with the best fitting parameters and the SSE was calculated between the participant's true PRT and the model's PRT. To compute the model's final cross validation score, we summed over the test SSE from each fold.

2.3 Results

2.3.1 Model simulations in single patch type environments

We examined the extent of over- and underharvesting as a function of the richness of the environment and the parameters governing structure learning (α) and uncertainty adaptive discounting (γ_{coef}). We simulated the model in single patch type environments to demonstrate that overharvesting could be produced through these two mechanisms in an environment commonly used in patch foraging tasks. It is important to note that, because of our definition of uncertainty, discounting adaptation is dependent on the structure learning parameter. We take uncertainty as the entropy of the posterior distribution over the current patch type. If a single patch type is assumed ($\alpha = 0$), then the entropy will always be zero and the discounting rate will be static. In our exploration of the parameter space, we find that as α increases over harvesting increases. Similarly, increasing γ_{coef} also increases overharvesting, however, only if $\alpha > 0$ (Fig 2.2E). Additionally, the overall richness of the environment interacts with the influence of these parameters on overharvesting — α and γ_{coef}'s influence is attenuated with increasing richness. The environment's richness also determines the baseline (when $\alpha = 0$ and $\gamma_{coef} \leq 0$) extent of over- and underharvesting. Because our model begins with a prior over the decay rate centered on 0.5, this produces overharvesting in the poor environment (mean decay rate = 0.2), optimal harvesting in the neutral (mean decay rate = 0.5), and underharvesting in the rich (mean decay rate = 0.8). In sum, we have shown, in multiple single patch type environments varying in richness, that overharvesting can be produced through a combination of mechanisms — structure learning and uncertainty adaptive discounting.

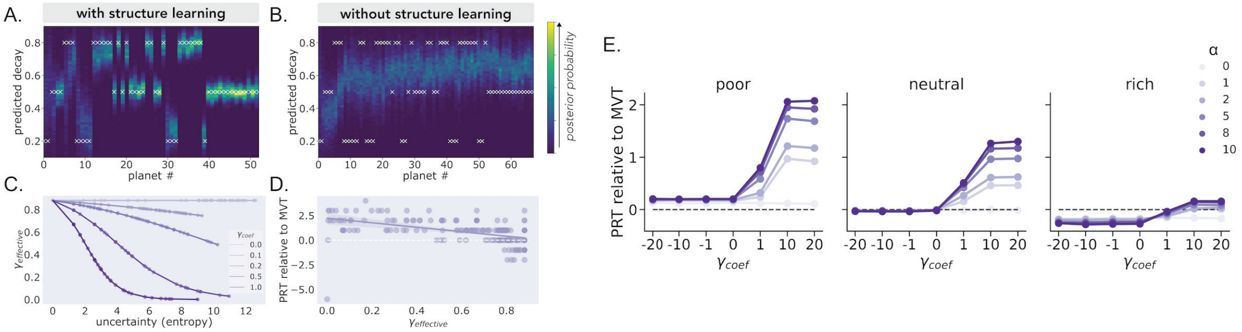

Figure 2.2: Structure learning improves prediction accuracy. A. With structure learning A simulated agent's posterior probability over the upcoming decay rate on each planet is plotted. If the forager's prior allows for the possibility of multiple clusters ($\alpha > 0$), they learn with experience the cluster-unique decay rates. Initially, the forager is highly uncertain of their predictions. However, with more visitations to different planets, the agent makes increasingly accurate and precise predictions. **B. Without structure learning** If the forager's prior assumes a single cluster ($\alpha = 0$), the forager makes inaccurate and imprecise predictions - either over or underestimating the upcoming decay, depending on the planet type. This inaccuracy persists even with experience because of the strong initial assumption. **Uncertainty adaptive discounting. C. The effect of γ_{coef}** The entropy of the posterior distribution over patch type assignment is taken as the forager's internal uncertainty and is used to adjust their discounting rate, $\gamma_{effective}$. The direction and magnitude of uncertainty's influence on the discounting rate is determined by the parameter, γ_{coef}. The more positive the parameter is, the more the discounting rate is reduced with increasing uncertainty, formalized as entropy. If negative, the discounting rate increases with greater uncertainty. **D. The effect of $\gamma_{effective}$ on overharvesting** Increasing γ_{base} increases the baseline discounting rate while increasing the slope term increases the extent the discounting rate adapts in response to uncertainty. **E. Overharvesting increases with α and γ_{coef} in single patch type environments** Simulating the model in multiple single patch type environments with varying richness, we find that increasing α and γ_{coef}, holding γ_{base} constant, increases the extent of overharvesting (PRT relative to MVT). The richness of the environment determines the extent of the parameters' influence, with it being greatest in the poor environment.

2.3.2 Model-free analyses

Participants adapt to local richness

We first examined a prediction of MVT — foragers should adjust their patch leaving to the richness of the local patch. In the task environment, planets varied in their richness or how quickly they depleted. Slower depletion causes the local reward rate to more slowly approach the global reward rate of the environment. Thus, MVT predicts that stay times should increase as depletion rates slow. As predicted, participants stayed longer on rich planets relative to neutral (t(115) = 19.77, $p < .0001$) and longer on neutral relative to poor (t(115) = 12.57, $p < .0001$).

Experience decreases overharvesting

Despite modulating stay times in the direction prescribed by MVT, participants stayed longer or overharvested relative to MVT when averaging across all planets (t(115) = 3.88, $p = .00018$). However, the degree of overharvesting diminished with experience. Participants overharvested more in the first two blocks relative to the final two (t(115) = 3.27, $p = .0014$). Our definition of MVT assumes perfect knowledge of the environment. Thus, participants approaching the MVT optimum with experience is consistent with learning the environment's structure and dynamics.

Local richness modulates overharvesting

We next considered how participants' overharvesting varied with planet type. As a group, participants overharvested only on poor and neutral planets while behaving MVT optimally on rich planets (Fig. 2.3A; poor - t(115) = 6.92, $p < .0001$; neutral - t(115) = 9.00, $p < .0001$;

rich - t(115) = 1.38, p = .17).

Environment dynamics modulate decision time and overharvesting

We also asked how participants adapted their foraging strategy to the environment's dynamics or transition structure. Upon leaving a planet, it was more common to transition to a planet of the same type (80%, "no switch") than transition to a planet of a different type ("switch"). Thus, we reasoned that switch transitions should be points of maximal surprise and uncertainty given their rareness. However, this would only be the case if the participant could discriminate between planet types and learned the transition structure between them.

If surprised, a participant should take longer to make a choice following a rare "switch" transition. So, we next examined participants' reaction times (z-scored and log-transformed) for the decision following the first depletion on a planet. We compared when there was a switch in planet type versus where there was none. As predicted, participants showed longer decision times following a "switch" transition suggesting they were sensitive to the environment's structure and dynamics (Fig. 2.3B; t(115) = 2.65, p = .0093).

If uncertain, our adaptive discounting model predicts that participants should discount remote rewards more heavily and, consequently, overharvest to a greater extent. To test this, we compared participants overharvesting following rare "switch" transitions to their overharvesting following the more common "no switch" transitions. Following the model's prediction, participants marginally overharvested more following a change in planet type (t(115) = 1.86, p = .065). When considering only planets that participants overharvested on on average (poor and neutral), overharvesting was significantly greater following a change (Fig. 2.3C; t(115) = 4.67, $p < .0001$).

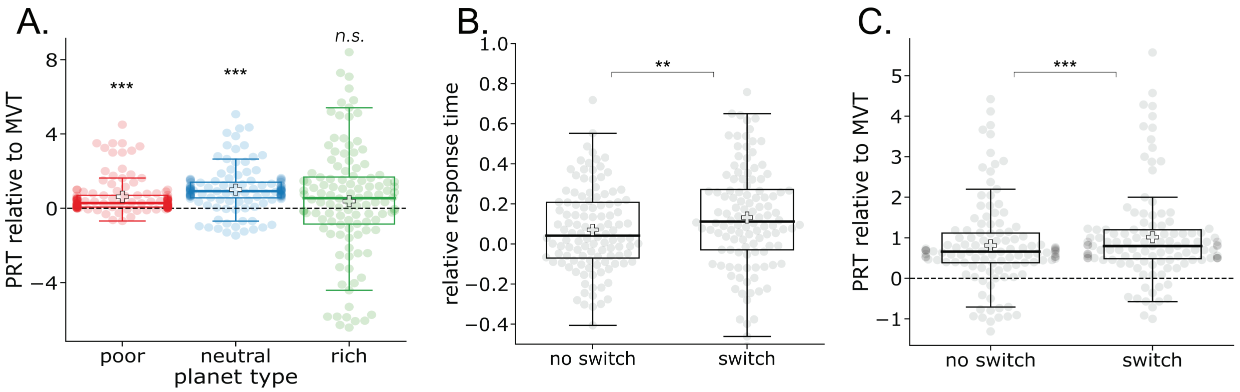

Figure 2.3: Model-free results A. Planet richness influences over and underharvesting behavior. Planet residence times (PRT) relative to Marginal Value Theorem's (MVT) prediction are plotted as the median (± one quartile) across participants. The grey line indicates the median while the white cross indicates the mean. Individuals' PRTs relative to MVT are plotted as shaded circles. In aggregate, participants overharvested on poor and neutral planets and acted MVT optimally on rich planets. **B. Decision times are longer following rare switch transitions.** If a participant has knowledge of the environment's planet types and the transition structure between them, then they should be surprised following a rare transition to a different type. Consequently, they should take longer to decide following these transitions. As predicted, participants spent longer making a decision following transitions to different types ("switch") relative to when there was transition to a planet of the same type ("no switch"). This is consistent with having knowledge of the environment's structure and dynamics. **C. Overharvesting increases following rare switch transitions.** On poor and neutral planets, participants overhavested to a greater extent following a rare "switch" transition relative to when there was a "no switch" transition. This is consistent with uncertainty adaptive discounting. Switches to different planet types should be points of greater uncertainty. This greater uncertainty produces heavier discounting and in turn staying longer with the current option.*p <0.05, **p <0.01, ***p <0.001

2.3.3 Model-based analyses

Structure learning with adaptive discounting provide the best account of participant choice

To check the models' goodness of fit, we asked whether the compared models could capture key behavioral results found in the participants' data. For each model and participant, we simulated an agent with the best fitting parameters estimated for them under the given model. Only the adaptive discounting model was able to account for overharvesting when averaging across all planets (Fig. 2.4A, t(115) = 8.87, $p < .0001$). The temporal-difference learning model predicted MVT optimal choices on average (t(115) = 1.30, p = .19) while the MVT learning model predicted underharvesting (t(115) = -7.26, p < .0001). These differences were primarily driven by predicted behavior on the rich planets (Fig. 2.4B).

Model fit was also assessed at a more granular level (stay times on individual planets) using 10-fold cross validation. Comparing cross validation scores as a group, participants' choices were best captured by the adaptive discounting model (Fig. 2.4C; mean cross validation scores — adaptive discounting: 16.55, TD: 22.47, MVT learn: 32.31). At the individual level, 64% of participants were best fit by the adaptive discounting model, 14% by TD, and 22% by MVT learn.

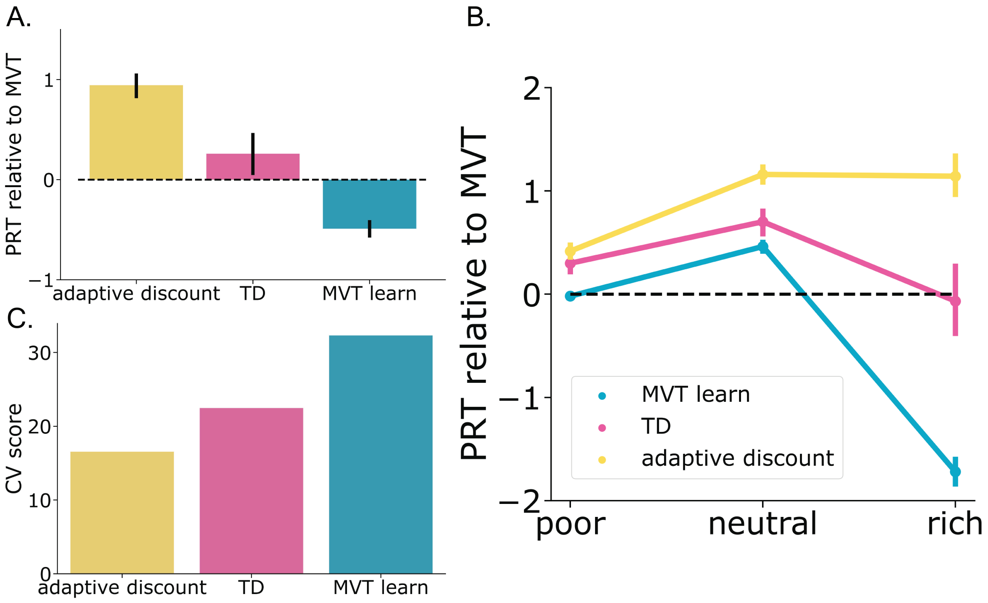

Figure 2.4: Modeling results A. The adaptive discounting model predicts overharvesting. Averaging across all planets, only the adaptive discounting model predicts overharvesting while the temporal-difference learning model predicts MVT optimal behavior and the MVT learning model predicts underharvesting. This demonstrates that overharvesting, a seemingly suboptimal behavior, can emerge from principled statistical inference and adaptation. **B. Model predictions diverge most on rich planets.** Similar to participants, the greatest differences in behavior between the models occurred on rich planets. **C. The adaptive discounting model provides the best account for participant choices.** The adaptive discounting model had the lowest mean cross validation score indicating it provided the best account of participant choice at the group level.

Adaptive discounting model parameter distribution

Because the adaptive discounting model provided the best account of choice for most participants, we examined the distribution of individuals' best fitting parameters for the model. Specifically, we compared participants' estimated parameters to two thresholds. These thresholds were used to identify whether a participant 1) inferred and assigned planets to multiple clusters and 2) adjusted their overharvesting in response to internal uncertainty.

The threshold for multi-cluster inference, 0.8, was computed by simulating the adaptive discounting model 100 times and finding the lowest value that produced multi-cluster inference in 90% of simulations. 76% of participants were above this threshold (Fig 2.5A). Thus, most participants were determined to be "structure learners" using our criteria.

The threshold for uncertainty-adaptive discounting was assumed to be 0. A majority of participants, 93%, were above this threshold (Fig 2.5C). These participants were determined to be "adaptive discounters", those who dynamically modulated their discounting factor in accordance with their internal uncertainty.

We next looked for relationships between parameters. Uncertainty should be greatest for individuals who have prior expectations that do not match the environment's true structure, whether too complex or too simple. Consistent with this, there was a non-monotonic relationship between the structure learning and discounting parameters. γ_{base} and γ_{coef} were greatest when α was near its lower bound, 0, and upper bound, 10 (γ_{base}: $\beta = 0.080$, $p < .0001$; γ_{coef}: $\beta = 0.021$, $p < .0001$). An individual's base level discounting constrains the range over which uncertainty can adapt the effective discounting. Reflecting this, the two discounting parameters were positively related to one another (τ = -0.33, $p < .0001$).

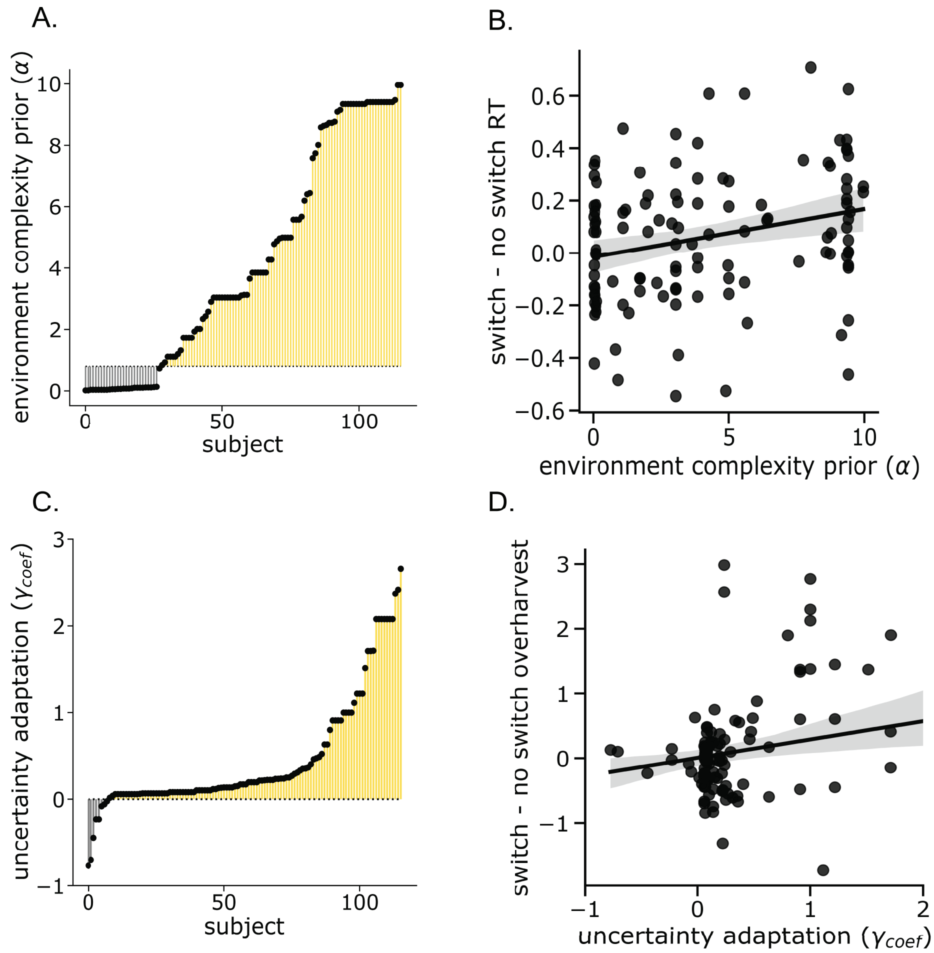

Figure 2.5: Parameter distributions A. Participants learned the structure of the environment. Distribution of participants' priors over environment complexity, α. Each individual's parameter is shown relative to a baseline threshold, 0.8. This threshold is the lowest value that produced multi-cluster inference in simulation. Most participants (76%) fall above this threshold indicating a majority learned the environment's multi-cluster structure. **B. Environment complexity parameters were positively related to reaction time sensitivity to transition frequency.** An individual must infer multiple planet types to be sensitive to the transition structure between them. In terms of the model, this would correspond to having a sufficiently high environment complexity parameter. Validating this parameter, it was positively correlated with individual's modulation of reaction time following a rare transition to a different planet type. **C. Participants adapted their discounting computations to their uncertainty over environment structure.** Distribution of participant's uncertainty adaptation parameter, γ_{coef}. Each individual's parameter is shown relative to a baseline of 0. A majority were above this threshold (93%) indicating most participants dynamically adjusted their discounting, increasing it when they experienced greater internal uncertainty. **D. Uncertainty adaptation parameters were positively related to overharvesting sensitivity to transition frequency.** If an individual increases their discounting to their internal uncertainty over environment structure, then they should discount more heavily following rare transitions and stay longer with the current option. Consistent with this, we found that the extent an individual increased their overharvesting following a rare transition was related to their uncertainty adaptation parameter.

Parameter validation

Correlations with model-free measures of task behavior confirmed the validity of the model's parameters. We interpret α as reflecting an individual's prior expectation of environment complexity. α must reach a certain threshold to produce inference of multiple clusters and consequently, sensitivity to the transitions between clusters. Validating this interpretation, participants with higher fit α demonstrated greater switch costs between planet types (Fig 2.5B, Kendall's $\tau = 0.17$, $p = .00076$). Moreover, this relationship was specific to α. γ_{base} and γ_{coef} were not significantly correlated with switch cost behavior (γ_{base}: τ = -0.036, p = .57; γ_{coef}: τ = -0.10, p = .11). This is a particularly strong validation as the model was not fit to reaction time data. Validating γ_{coef} as reflecting uncertainty-adaptive discounting, the parameter was correlated with the extent overharvesting increased following a rare transition or "switch" between different planet types (Fig 2.5D, τ = 0.15, p = .016). This was not correlated with α nor the baseline discounting factor γ_{base} (α: τ = -0.011, p = .86; γ_{base}: τ = 0.082, p = .20).

2.4 Discussion

While Marginal Value Theorem (MVT) provides an optimal solution to patch leaving problems, organisms systematically deviate from it, staying too long or overharvesting. A critical assumption of MVT is that the forager has accurate and complete knowledge of the environment. Yet, this is often not the case in real world contexts — the ones to which foraging behaviors are likely to have been adapted [90]. We propose a model of how foragers could rationally learn the structure of their environment and adapt their foraging decisions to it. In simulation, we demonstrate how seemingly irrational overharvesting can emerge as a byproduct of a rational dynamic learning process. In a heterogeneous, multimodal environment, we compared how well our structure learning model predicted participants' choices relative to two other models — one implementing a MVT choice rule with a fixed representation of the environment and the other a standard temporal-difference learning algorithm. Importantly, only our structure learning model predicted overharvesting in this environment. Participants' choices were most consistent with learning a representation of the environment's structure through individual patch experiences. They leveraged this structured representation to inform their strategy in multiple ways. One way determined the value of staying. The representation was used to predict future rewards from choosing to stay in a local patch. The other modulated the value of leaving. Uncertainty over the accuracy of the representation was used to set the discount factor over future value. These results suggest that in order to explain foraging as it occurs under naturalistic conditions optimal foraging may need to provide an account of how the forager learns to acquire accurate and complete knowledge of the environment, and how they adjust their strategy as their representation is refined with experience.

In standard economic choice tasks, humans have been shown to act in accordance with rational statistical inference of environment structure. Furthermore, by assuming humans must learn the structure of their environment from experience, seemingly suboptimal behaviors can be rationalized including prolonged exploration [2], melioration [179], social biases [176], and overgeneralization [39]. Here, we extend this proposal to decision tasks with sequential dependencies, which require simultaneous learning and dynamic integration of both the distribution of immediately available rewards and the underlying contingencies that dictate future outcomes. This form of relational or category learning has long been associated with distinct cognitive processes and neural substrates from those thought to underlie reward-guided decisions [152], including the foraging decisions we investigate here [108]. However, a network of neural regions overlapping those supporting relational learning are more recently thought to play a role in deliberative, goal-directed decisions [30, 198].

If foragers are learning a model of the environment and using it to make decisions for reward, this suggests that they may be doing something like model-based reinforcement learning (RL). In related theoretical work, patch leaving problems have been cast as a multi-armed bandit problem from RL. Which actions are treated as the "arms" is determined by the nature of the environment. In environments where the next patch is unknown to the foragers, the two arms become staying in the current patch and leaving for a new patch. In environments in which the forager does have control over which patch to travel to next, the arms can become the individual patches themselves. Casting patch leaving as an RL problem allows for the use of RL's optimal solutions as benchmarks for behavior. Application of these optimal solutions in foraging have been found to capture search patterns [188, 134], choice of lower valued options [103], and risk aversion [139]. In contrast to this work and our own, Constantino & Daw [40] found human foragers' choices to be better explained by a MVT model augmented with a learning rule than a standard reinforcement learning model. However, importantly, their task environment was homogeneous and the RL model tested was model-free (temporal-difference learning). Thus, the difference in results could be attributed to differences in task environments and class of models considered. A key way our model deviates from a model-based RL approach is that prospective prediction is only applied in computing the value of staying while the value of leaving is similar to MVT's threshold for leaving – albeit discounted proportionally to the agent's internal uncertainty over their representation's accuracy. In the former respect, our model parallels the framework discussed by Kolling & Akam [107] to explain humans sensitivity to the gradient of reward rate change during foraging observed by Wittman et al [206]. Given that computing the optimal exit threshold under a pure model-based strategy would be highly computationally expensive, Kolling & Akam [107] suggest pairing model-based patch evaluation with a model-free, MVT-like exit threshold. Under their proposal, the agent leaves once the local patch's average predicted reward rate over n time steps in the future falls below the global reward rate. We build on, formally test, and extend this proposal by explicitly computing the representational uncertainty at each trial and adjusting planning horizon accordingly.

While learning a model of the environment is beneficial, it is also challenging and computationally costly. With limited experience and computational noise, an inaccurate model of the environment may be inferred. An inaccurate model, however, can be counteracted by adapting certain computations. In this way, lowering the temporal discounting factor acts as a form of regularization or variance reduction [150, 96, 63, 197, 5]. Empirical work has found humans appear to do something like this in standard intertemporal choice tasks. Gershman & Bhui [70] found evidence that individuals rationally set their temporal discounting as a function of the imprecision or uncertainty of their internal representations. Here, we found that humans while foraging act similarly, overharvesting to a greater extent at points of peak uncertainty. While temporal discounting has been proposed as a mechanism of overharvesting previously [25, 40, 34], the discounting factor is usually treated as a fixed, subject-level parameter, inferred from choice. Thus, it provides no mechanism for how the factor is set let alone dynamically adjusted with experience. In contrast, our model proposes a mechanism through which the discounting factor is rationally set in response to both the external and internal environment. To further test the model, future work could examine the model's prediction that overharvesting should increase as the environment's stochasticity (observation noise) increases. In the current task environment, noise comes from the variance of the generative decay rate distributions. An additional source of noise could be from the reward itself. After the decay rate has been applied to the previously received reward, white Gaussian noise could be added to the product. As a result, the distribution of observed decay rates would have higher variance than the generating decay rate distributions. This reward generation process should elicit greater uncertainty for the forager than the current reward generation process, and consequently, greater overharvesting.

Finally, our observation that humans adjust their planning horizons dynamically in response to state-space uncertainty may have practical applications in multiple fields. In psychiatry, foraging has been proposed as a translational framework for understanding how altered decision-making mechanisms contribute to psychiatric disorders [3]. An existing body of work has examined how planning and temporal discounting are impacted in a range of disorders from substance use and compulsion disorders [7, 74] to depression [155] to schizophrenia [92, 43]. This wide range has led some to suggest that these abilities may be a useful trans-diagnostic symptom and a potential target for treatment [6]. However, it remains unclear *why* they are altered in these disorders. Our findings may provide further insight by way of directing attention towards identifying differences in structure learning and uncertainty adaptation. How uncertainty is estimated and negotiated has been found to be altered in several mood and affective disorders [10, 154], theoretical work has suggested that symptoms of bipolar disorder and schizophrenia may be explained through altered structure learning [156], and finally, in further support, compulsivity has been empirically associated with impaired structure learning [170]. Our model suggests a rationale for why theses phenotypes co-occur in these disorders. Alternatively, myopic behavior may not reflect differences in abilities but rather in environment. Individuals diagnosed with these disorders, rather, may more frequently have to negotiate volatile environments. As a result, their structure learning and uncertainty estimation are adapted for these environments. Potential treatments, rather than targeting planning or temporal discounting, could address its possible upstream cause of uncertainty – increasing the individual's perceived familiarity with the current context or increasing their self-perceived ability to act efficaciously in it. Another application could be in the field of sustainable resource management, where it has recently been shown that, in common pool resource settings (e.g. waterways, grazing fields, fisheries), the distribution of individual participants' planning horizons strongly determines whether resources are sustainably managed [14]. Here, we show that discount factor, set as a rational response to uncertainty about environmental structure, directly impacts the degree to which an individual tends to (over)harvest their locally available resources. The present work suggests that policymakers and institution designers interested in producing sustainable resource management outcomes should focus on reducing uncertainty – about the contingencies of their actions, and the distribution of rewards that may result – for individuals directly affected by resource availability, thus allowing them to rationally respond with an increased planning horizon and improved outcomes for all participants.

Chapter 3

Age-related differences in structure learning explain differences in exploration during a patch foraging task

3.1 Introduction

Childhood and adolescence are characterized as periods of heightened exploration and learning. However, this exploration can often come at the cost of forsaking immediate reward. How individuals negotiate the tension between exploration and exploitation varies with their developmental stage: children and adolescents explore more extensively than adults [26, 143, 121, 192] but do so in a less strategic way [75, 55, 129, 168, 185]. Past work has predominantly taken the approach of using reinforcement learning algorithms to quantify developmental differences in exploration as a function of value learning or action selection. But, by solely focusing on the valuation or selection processes, this work has neglected another potential source of difference — how the environment is internally represented.

The abilities underpinning mental model formation vary in their developmental trajectories. For instance, statistical learning emerges in infancy [163], while relational inference continues to be refined into young adulthood [166]. Due to their simplistic structure, conventional exploration tasks are ill-suited for characterizing how these differences in structure learning impact exploration. For example, in multi-armed bandit tasks, decision makers repeatedly choose from a set of options that differ in their probability of yielding reward. While these tasks do require some degree of structure learning [2], they offer few causal structures that the decision maker could reasonably entertain. The lack of complexity found in these tasks sharply contrasts with the rich structure found in natural environments. In response to this, some have called for the development of tasks that better capture this complexity, and as such, more accurately reflect the learning and decision contexts faced in the real world [205].

Patch foraging has been proposed as one more "naturalistic" alternative to multi-armed bandit tasks [132, 3]. Instead of choosing from a repeated set of options, decision makers decide between staying with a known patch of resources that cannot be returned to once left and searching for an alternative. In this decision context, understanding the environment's structure is critical. Because each patch is only visited once, the decision maker must infer properties of the current patch by generalizing from past patches. Moreover, without a specific alternative option in mind, the decision maker is required to estimate the environment's overall distribution of rewards to use as a "best guess" for the alternative. Marginal Value Theorem (MVT; [35]) provides the optimal decision policy within these tasks and the reference point by which exploration and exploitation are defined with respect to. Under the assumption of perfect knowledge of the environment, staying with a current patch longer than optimal is considered over-exploitation, or "overharvesting", while leaving sooner is considered exploratory.

In these patch foraging tasks, adults consistently have been found to overharvest [40, 117, 115, 173, 108]. Similarly, adolescents have also been found to overharvest [89], albeit to a lesser extent [122]. This prompts the question: if patch foraging tasks more closely mimic real world decision contexts, why do we consistently behave suboptimally in them? Of relevance to this question, rewards are homogeneously distributed in standard versions of the task. This facilitates rapid and effective generalization between patches, thus putting the decision maker in greater alignment with MVT's assumption of perfect knowledge of the environment. However, this simplified structure removes one of the critical features that makes this class of tasks more naturalistic. Potentially, the appearance of overharvesting in these simplified environments could reflect the use of decision strategies people tend to use in complex and uncertain real world environments, strategies that deftly handle this uncertainty and seek to resolve it. Related work has found that people alter their decision strategies to perceived changes in the environment, even when these changes are spurious. They attribute this to the structure of real world environments in which truly random events are rare [208].

In recent work in adults [88], we used a Bayesian structure learning model to demonstrate that overharvesting could be explained by a mismatch between foragers' prior expectation of the environment's structure and it's true underlying structure. Using the same computational model and task, we ask: can developmental differences in exploration, from middle childhood through young adulthood, be explained by differences in representation? The literature suggests two competing hypotheses. One one hand, from a very young age, children are prodigious structure learners, who can quickly extract temporal regularities from their environment [164]. Under this hypothesis, we would expect to find no age-related differences in how participants internally represent the environment and this should be accompanied by a lack of difference in exploration. On the other hand, there is a developmental dissociation between structure acquisition and its *use* during decision making. The use of structure knowledge is known to have a more protracted development [144, 47, 153, 37]. This suggests that younger participants' decisions should be guided by simpler representations of the environment, leading them to explore more than adults. Here, we sought to arbitrate between these two hypotheses.

3.2 Methods

3.2.1 Participants

252 participants between the ages of 8 and 25 completed the online study and were included in all analyses (mean age = 17.11 years, standard deviation age = 5.29, 128 females, 124 males). The target sample size was based on extrapolating from the prior adult study's sample size to evenly cover our age range. It well surpasses those used in prior studies examining value-guided learning and decision making in samples spanning the same age range [37, 144, 143]. An additional 45 participants completed the study but were excluded. Participants were sequentially excluded according to the following criteria: failed the instruction comprehension check quiz more than two times (n=4), had an average reaction time below 200 milliseconds (n=12), their average planet residence time fell 2 standard deviations above or below the group average (n=14) or they consistently used an extreme strategy, either fully depleting more than 75% of visited planets' gem mines (n=4) or leaving more than 75% of visited planets immediately after the initial dig (n=11). For completing the study, participants were compensated with $10 Amazon gift cards. Depending on task performance, they could also earn a bonus that ranged from $0 to $2.

Our final sample of participants included 70 children (8.08 - 12.94 years; mean age = 10.49, 36 females), 68 adolescents (13.07 - 17.94 years; mean age = 15.47, 35 females), and 114 adults (18 - 25.83 years; mean age = 22.14, 57 females). All participants reported normal or corrected-to-normal vision and no history of psychiatric or learning disorders.

Participants were recruited from the Hartley lab's participant database for which we solicit sign-ups via Facebook and Instagram ads, local science fairs and events, and fliers on New York University's campus. Prior to their participation in an online study, participants' identity and age were confirmed by the researchers.

3.2.2 Task

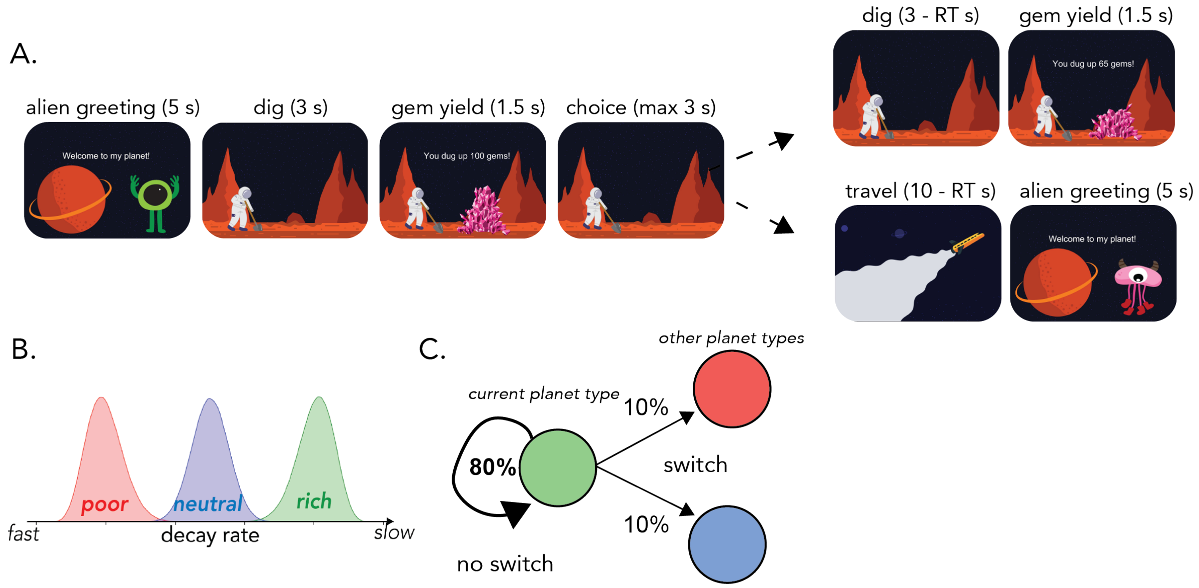

Figure 3.1: A. Serial stay-switch task. Across four six-minute blocks, participants mined for space gems on different planets. On each trial, they had to decide between staying to dig from a depleting gem mine or incurring a time cost to travel to a new planet. **B. Environment structure.** Planets varied in their richness or, more specifically, the rate at which they depleted with each dig. There were three planet types — poor, neutral, and rich — each with their own characteristic distribution over depletion rates. **C. Environment dynamics.** Planets of a similar type clustered together. A new planet had an 80% probability of being the same type as the prior planet ("no switch"). However, there was a 20% probability of transitioning or "switching" to a planet of a different type.

Participants completed a variant of a patch foraging task used in a prior adult study [88]. We adapted the task to be child-friendly by reducing its length, adding more extensive instructions, and increasing the maximum decision time. Within the task, participants traveled to different planets to mine for space gems (Fig. 3.1). They were told that the amount of gems they collected determined their bonus payment. On each planet, participants would perform an initial dig yielding a reward generated from a Normal distribution ($\mu = 100$, $\sigma = 5$). Then, they decided between staying on the same planet to dig from the depleting mine or incurring a time cost to travel to a new planet with a replenished mine. If they decided to stay, a short animation of their avatar digging was shown (3 sec, minus the decision time for that trial) followed by the gem yield (1.5 sec). If they decided to travel, a short animation of a rocket ship was displayed (10 sec minus decision time) followed by their landing and an alien greeting them (5.5 s). The duration of the animations varied according to the participant's decision time on trial. This ensured that they could not influence the environment's overall reward rate via the rapidity of their responses. If the participant failed to make a decision in the allotted time, a screen was displayed with a red 'X' and a message urging them to make their decisions more quickly. They were then given the opportunity to try choosing again. Participants repeated making stay-leave decisions until the conclusion of the block. Participants completed four blocks lasting 6 minutes each.

Planets in the task environment varied in their quality. They could belong to one of three types distinguished by their depletion rate per dig. Rich planets depleted the slowest on average (Beta distributed with parameters α=50, β=12, mean=0.8, sd=0.05), poor planets the fastest (α=13, β=51, mean=0.2, sd=0.05), and neutral planets fell in between (α=50, β=50, mean=0.5, sd=0.05). To mimic natural environments, planet quality was correlated in time. New planets had an 80% likelihood of being the same type as the last planet. The other 20% of trials were "switch" trials, on which the new planet was equally likely (10% each) to belong to one of the two remaining types. Critically, neither the environment's structure or dynamics were explicitly signaled to participants, requiring them to infer this information from the sequence of rewards they received. Because we were interested in identifying potential age-related variation in participants' representational biases, independent of any task demands, we structured the task environment such that use of one form of representation was not incentivized over the other. In simulation with the structure learning model defined below, agents using a single planet type representation and those using a multi-planet representation earned similar total rewards on average.

3.2.3 Analysis approach

Mixed effects models We used the “lme4” package for R [16] to fit mixed-effects models to our data. Except where noted, models included participant-level random intercepts and random slopes across within-participant fixed effects. In constructing our models, we began with the maximal model in order to minimize Type I error [15]. If the model failed to converge, we removed interactions between random slopes and then random slopes themselves until models converged. We set the number of model iterations to 10,000 and used the ‘bobyqa’ optimizer. All continuous variables (age, planet number, and reaction time) were z-scored prior to their inclusion in the models. Age was z-scored across participants while planet number and reaction times were z-scored within. Reaction times were also log-transformed. In the results section, we only discuss the results from the regression models that are relevant to our stated hypotheses. For the remaining results, see the appendix.

Marginal Value Theorem To quantify the extent to which individuals over- or under-harvested, we compared their planet (patch) resident times to the predictions of Marginal Value Theorem [35]. An MVT-optimal agent compares the current option’s expected immediate returns (V_{stay}) to the opportunity cost of choosing to engage with it over an alternative planet (V_{leave})

We take V_{stay} as the reward expected from digging again on the current planet. The MVT-optimal forager knows the richness of the planet they are on, and uses the true mean of its distribution to predict the upcoming depletion.

$$V_{stay} = r_t * \hat{d} \tag{3.1}$$

$$\hat{d} = \begin{cases} 0.2 & \text{if planet is poor} \\ 0.5 & \text{if planet is neutral} \\ 0.8 & \text{if planet is rich} \end{cases}$$

Where r_t is the reward received on the last dig, and $\hat{d}$ is the predicted depletion.

Because the forager is unaware of the quality of the next planet they will encounter, they estimate the rewards earned from one dig on an alternative planet by taking the global reward rate of the environment — the total rewards received (r_{total}) divided by the total time (t_{total}) spent foraging thus far — and taking the product of it and time required to dig up rewards (t_{dig}).

$$V_{leave} = \frac{r_{total}}{t_{total}} * t_{dig} \tag{3.2}$$

The forager compares these values and chooses greedily, taking the higher valued option.

Structure learning and adaptive discounting model Our model relaxes MVT's assumption of perfect knowledge of the environment's structure and dynamics. Doing so introduces two novel computations into the evaluation process.

Foragers do not know how many planet types there are, which type a given planet belongs to, nor the types' associated decay rate distributions. To model how this information could be rationally inferred, we use a Chinese Restaurant Process [4]. The distinguishing feature of the CRP is its prior which is composed of two parts: 1. the more planets already assigned to a type, the more likely that type is and 2. there remains some probability of a new type being created, proportional to the parameter α. The latter allows for the complexity of the foragers' representation to grow as experience in the environment is accumulated.

$$P(k) = \begin{cases} \frac{n_k}{N+\alpha} & \text{if k is old} \\ \\ \frac{\alpha}{N+\alpha} & \text{if k is new} \end{cases}$$

Where n_k is the number of planets assigned to cluster k, α is a clustering parameter, and N is the total number of planets encountered.

After observing one depletion on a planet, the forager can compute the posterior probability of a planet belonging to a type:

$$P(k|D) = \frac{P(D|k)P(k)}{\sum_{j=1}^{J} P(D|j)P(j)} \tag{3.3}$$

Where J is the number of clusters created up until the current planet, D is a vector of all the depletions observed on the current planet, and all probabilities are conditioned on prior cluster assignments of planets, $p_{1:N}$

Computing the posterior probability of a cluster type is computationally intractable. In order to approximate it, we use a particle filter with 200 particles. Each particle maintains a hypothetical set of planet type assignments, and how well these assignments explain the data determines the particle's weighting during the resampling process which occurs every time an agent leaves a planet. During resampling, a new particle pool is generated by weighted sampling with replacement from the old particle pool. Because of the weighting, the particles that best explain the data are more likely to be represented in the subsequent particle pool.

The forager's posterior probability over planet types can then be used to predict how much the gem yield will deplete on the next dig. To estimate this posterior, we use a Monte Carlo sampling procedure. A particle is sampled with probability proportional to its weights, then a planet type is sampled proportional to the posterior probability over types given by that particle, and finally, a decay rate is sampled from the distribution associated with the planet type. This procedure is repeated 1000 times and the decay rates are averaged over to produce the final prediction.

Each planet type's decay rate distribution is initialized as a Gaussian with μ=0.5 and σ=0.5. While the true decay rates are Beta distributed, the model assumes observations are normally distributed to allow for analytic updating with a Normal-Gamma prior.

A MVT-optimal forager does not discount future expected rewards in their computation of V_{leave}. This is reasonable under an assumption of perfect knowledge because their expectations are, by definition, correct. In relaxing this assumption, agents following our model do not have the same guarantee. The forager can never be fully certain that their representation accurately reflects the environment. Theoretical work from reinforcement learning has shown that under this form of uncertainty reducing the planning horizon can improve agents' performance [96]. We instantiated this ideal with an adaptive discount factor, $\gamma_{effective}$. Within our task reducing the planning horizon and decreasing the discounting rate produce indistinguishable behavior. $\gamma_{effective}$ is thus computed as a sum of an individual baseline discounting rate (γ_{base}), their current uncertainty (U), and the extent to which they modulate their discounting based on their uncertainty, (γ_{coef}). With the same Monte Carlo sampling procedure used to predict the next dig's gem yield, we define a multinomial distribution over the sampled clusters and take the entropy of it to get U.

$$\gamma_{effective} = \frac{1}{1 + e^{(-\gamma_{base} + \gamma_{coef} * U)}} \tag{3.4}$$

Action selection is modeled as a drift diffusion process (DDM; [157, 158]). Decisions are made via the accumulation of noisy information in favor of the considered actions, here, stay and leave. The value difference between the two actions (scaled by v) determines the rate at which the accumulated information proceeds toward a fixed decision threshold (a=1). With the DDM, we can jointly use choices and response times to inform our inference of participants' representation of the environment. Incorporating response times into the estimation of reinforcement learning model parameters has been shown to produce more robust fits to participant data [11, 62]. The structure-learning component of our model predicts decision uncertainty should peak at different points in the task depending on participants' representation, and thus response times can be particularly revealing of decision uncertainty [32].

We compared two versions of the model: one in which α was fixed at 0 and another in which α was fixed at 1. We allowed γ_{base}, γ_{coef} and v to be free parameters.

Alternative models We compare our structure learning models to the two primary models considered in Constantino & Daw [40]. The model that provided the best explanation of their data was, similar to our structure learning model, based on MVT. Critically, it differs from ours in its statelessness. Under their model, agents estimate V_{stay} and V_{leave} by averaging over all past experiences.

$$V_{stay} = \bar{D} * r_{t-1} \tag{3.5}$$

Where r_{t-1} is the reward received on the last dig, and D is the predicted depletion, which is an average over all past depletions.

The global reward rate, central to the computation of V_{leave}, is learned incrementally through trial and error. We allow η, the learning rate, to be a free parameter.

$$\begin{aligned} V_{leave} &= \rho * t_{dig} \\ \delta_t &= \frac{r_t}{\tau_t} - \rho_t \\ \rho_{t+1} &= \rho_t + (1 - (1 - \eta)^{\tau_t}) * \delta_t \end{aligned} \tag{3.6}$$

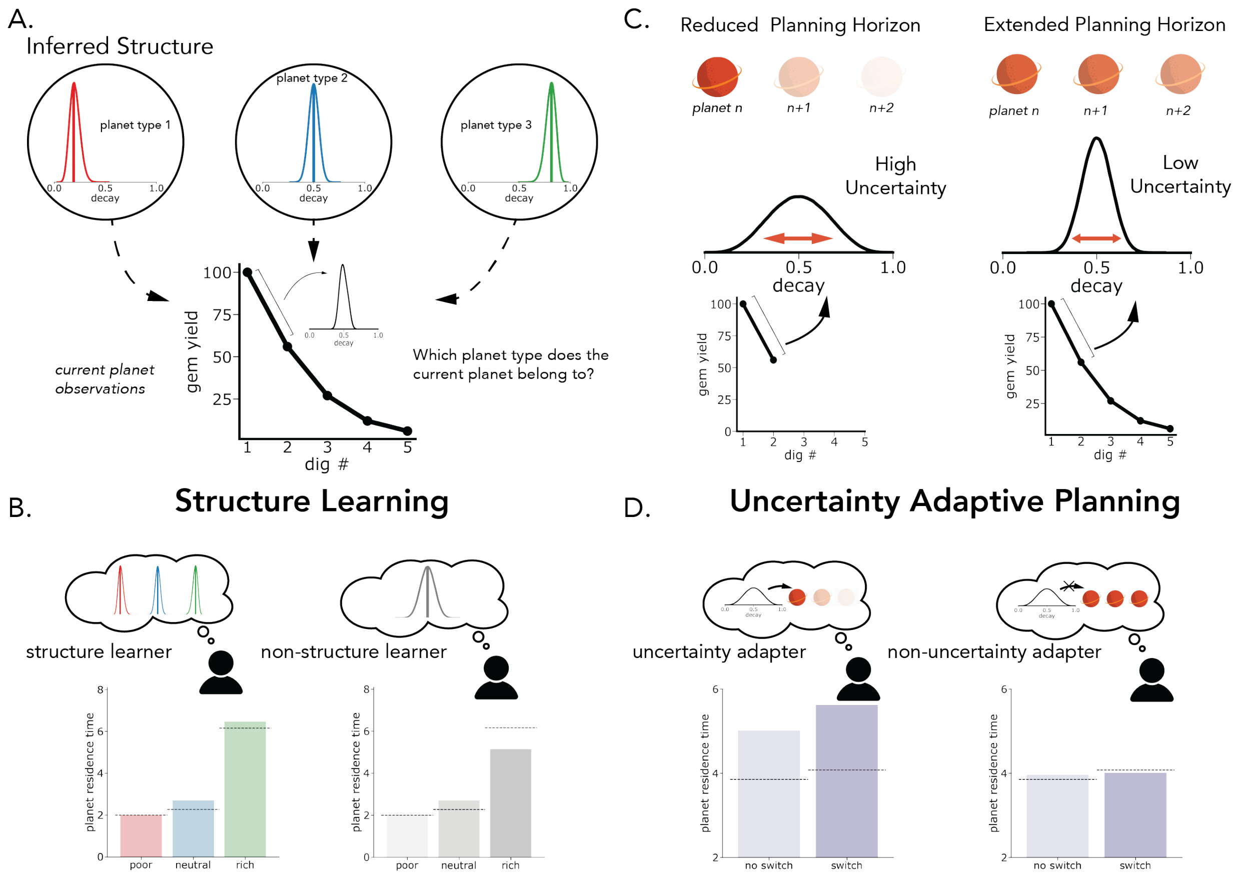

Figure 3.2: A. Structure learning computation. Based on the distribution of depletions observed on the current planet, the forager must infer the type that the current planet is most likely to belong to. Simultaneously, the forager must also infer the number of planet types present in the environment. This higher level inferential process is done according to a Chinese Restaurant Process. **B. Structure learning predictions.** The number of planet types the forager infers is dependent on the concentration parameter, α. Our model predicts distinct patterns of over- and under-harvesting on the various planet types depending on the number of planet types the forager infers. Bar heights indicate the agent simulated under the model's predicted planet residence time while the dotted lines indicate the MVT-optimal PRT. **C. Uncertainty adaptive planning computation.** Foragers adapt their planning horizon with respect to their uncertainty over the current planet's underlying type. When uncertain, they shouldn't plan too far in advance, but when certain, they should plan further into the future. **D. Uncertainty adaptive planning predictions.** If a forager adapts their planning horizon with their internal uncertainty, then their overharvesting should increase following relatively rare switches in planet type. If they do not adapt, then their planet residence times should not change in response to switches.

We also considered a model implementing a temporal-difference (TD) learning algorithm. TD agents learn a state-action value function, Q which estimates the cumulative future expected reward for taking an action in a given state. The state was determined by the reward received on the last dig, and the state space was created by considering the range of possible rewards the agent could receive (lower bound = 0, upper bound = 135) and segmenting it into logarithmically spaced bins. To implement this, we took the upper bound of a bin, b_{j+1}, for bin j, and found its lower bound, b_j by taking the difference between $log(b_{j+1})$ and $log(\bar{k})$, the mean decay rate across all planet types. To initialize the Q values, we set them to the cumulative reward the agent should earn based on the true global reward rate and their discounting factor, $\frac{rho_init}{1-\gamma}$. η and γ are free parameters

$$\begin{gathered}
V_{stay} = Q_t(s_t, dig) \\
V_{leave} = Q_t(leave) \\
D_{t-1} \sim Bernoulli(P(a_t)) \\
\delta_t = r_{t-1} + \gamma^{\tau_t}(D_t * Q_t(s_t) + (1 - D_t) * Q_t(leave)) - Q_t(s_{t-1}, a_{t-1}) \\
Q_t(s_{t-1}, a_{t-1}) = Q_t(s_{t-1}, a_{t-1}) + \eta * \delta_t
\end{gathered} \tag{3.7}$$

Departing from Constantino & Daw's [40]'s original models, we use a drift diffusion model for action selection in both the MVT and TD models to better equate them with our structure learning model. As in our model, the difference between V_{stay} and V_{leave} determines the drift rate, with the difference scaled by the free parameter v, and the threshold is fixed across all trials.

Model fitting For all models, we fit individual participants' stay-leave decisions and reaction times choice-by-choice, simulating the DDMs and calculating the analytic likelihood their choices using the approach and implementation introduced by Drugowitsch [53]. Each models' free parameters and bounds are listed in the appendix. We identified the parameter values that minimized the negative log posterior of participants' choices using Bayesian Adaptive Direct Search (BADS, [1]), an optimization algorithm designed to handle stochastic and computationally expensive functions. To increase the probability of finding the global minimum, we used 30 different starting positions generated from a Sobol sequence. Quasi-random search using Sobol sequences has been shown to be more effective than grid search or random search [21] while being more computationally efficient than Latin hypercube sampling [159]. We took the parameter values that produced the minimum negative log posterior across all starting points. For our primary model of interest, the structure learning model with $\alpha = 1$, we conducted parameter recoverability analyses (see appendix).

Amongst our considered models, the structure learning models' likelihoods are uniquely stochastic. This is because its posterior distribution over planet type assignments is approximated. To combat the noise induced by the approximation, we run the cluster assignment process 1000 times. For each run, we compute the log likelihood of each participant's choice, summed over them, and finally, added the log prior to get the log posterior. We marginalized over the 1000 runs, averaging over the log posteriors and negating it to compute the final value input to the optimization algorithm.

Model comparison We compared the models' ability to account for participant choices using 10-fold cross validation (see appendix for model recovery). For each participant, we split choices into 10 separate training and test datasets. For each fold, we identified the model parameters that best explained choices in the training dataset. We then averaged the final negative log posterior from each fold to compute the model's final cross-validation score used for comparison.

3.3 Results

3.3.1 Model-free analyses

Overharvesting increases with age We first sought to characterize the extent to which participants aligned with the MVT-optimal policy and how the extent of alignment varied with age. Using mixed-effects linear regression, we modeled the effect of age on the deviance of participants' planet residence time (PRT, quantified as the number of digs completed on a planet) from the MVT-optimal PRT. Participants showed a general tendency towards staying longer than optimal, replicating the widely observed phenomena of overharvesting (β_0=0.81, $p < .001$). This tendency strengthened with age (β_{age}=0.22, $p = .045$) with younger participants exploring more and consequently, acting more closely to the MVT-optimal policy.

Use of structure knowledge strengthens with age Next, we asked which environmental features modulated participants' overharvesting. Our Bayesian structure-learning model predicts that a forager's response to these features depends on how they represent the environment. Notably, rich planets should produce the most patent response differences – foragers representing multiple planet types should overharvest while those representing a single type should underharvest (Fig. 4.1 — model schematic, model predictions). To test the model's predictions, we ran a mixed-effects linear regression modeling the influence of planet richness, planet number, age, and their interactions on participants' deviance from MVT optimality. Participants overharvested the most on neutral planets and the least on poor (Fig. 3.3A; intercept: β=1.29, $p < .001$; poor planet: β=-0.63, $p < .001$; rich planet: β=-0.42, $p = .0018$). The extent of overharvesting decreased with experience in the environment, particularly so for rich planets (Fig. 3.3B; planet number: β=-0.24, $p < .001$; planet number x poor planet interaction: β=0.066, $p = .15$; planet number x rich planet interac-

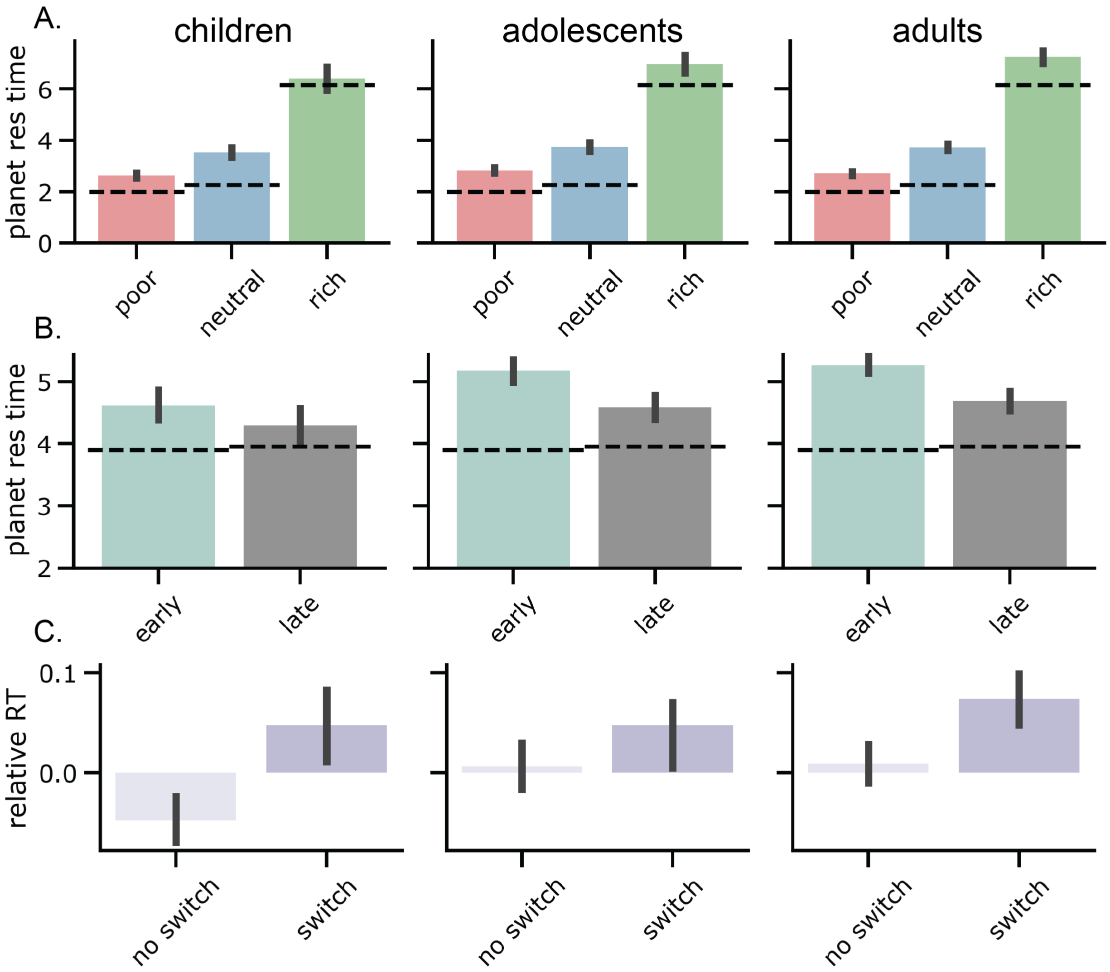

Figure 3.3: Signatures of Structure Learning. **A-B.** Planet resident times (PRT) relative to the MVT-optimal prescribed PRT. Bars indicate the age group mean, error bars indicate the mean ± the standard error of the mean, and the dotted line indicate the MVT-optimal PRT. **A.** Across all age groups, PRTs increased with planet richness. When compared to the MVT-optimal PRTs, participants overharvested across all planet types but the extent varied across the types. Neutral planets produced the most overharvesting and poor the least. Uniquely on rich planets, older participants overharvested to a greater extent than younger participants, aligning them more closely with the predictions of our structure learning model. There were no age-related differences on the other two planet types.**B.** Participants' in all age groups decreased the extent of their overharvesting with task experience (early - first two blocks, late - final two blocks). Because MVT-optimality assumes the forager knows the environment's true underlying structure, increasing alignment with MVT is suggestive of structure learning. **C.** On planet's in which there was a switch type, participants' took longer to respond on choices immediately following the first depletion, the first observation indicative of whether there was in fact a switch. Participants in all age groups' showed this decision slowing in response to switches.

tion: β=-0.26, $p < .001$). As an MVT optimal agent has full knowledge of the environment, participants' increasing compliance with MVT over the course of the task is consistent with incrementally learning the environment's underlying structure.

Examining effects of age and its interactions, we found that the signature of having learned a representation of environmental structure with multiple planet types increased with age. Uniquely on rich planets, overharvesting increased with age (age: β=0.059, $p = .47$; age x poor planet interaction: β=-0.045, $p = .39$; age x rich planet interaction: β=0.36, $p = .0078$).

As an additional, implicit measure of structure learning, we examined reaction times following switches in planet type. We only considered the second decision made on a planet, as these decisions directly followed the first observation that informs inference of the planet's type. If a participant has learned that there are multiple planet types which cluster together in time, then they should be surprised when the planet type changes. Thus, if participants are slower to make their next decision following a change in planet type, this provides evidence of sensitivity to environmental structure. To address this, we modeled the effect of a planet type switch, planet number, age, and their interactions on reaction time with participant-level random intercepts and random slopes for the switch regressor. Participants' reaction times increased following switches, indicating knowledge of the environment's structure and dynamics (Fig. 3.3C; switch point: β=0.049, $p = .037$). While reaction times grew faster over the course of the task (planet number: β=-0.049, $p < .001$), reaction time slowing at switch points did not lessen with task experience (switch x planet number interaction: β=0.014, $p = .55$).

In line with prior findings [47], participants' choices suggested age-related differences in structure learning while their reaction times indicated no such differences. Participants across our age range showed similar slowing at switch points (age x switch point interaction: $\beta = 0.0088$, $p = .71$). Collectively, our results suggest a dissociation between younger participants' knowledge of environmental structure and their choice behavior. Decision slowing at switch points suggests an at least implicit awareness that planets differ. Nevertheless, they do not integrate this knowledge into their decision making, as evidenced by their patterns of overharvesting.

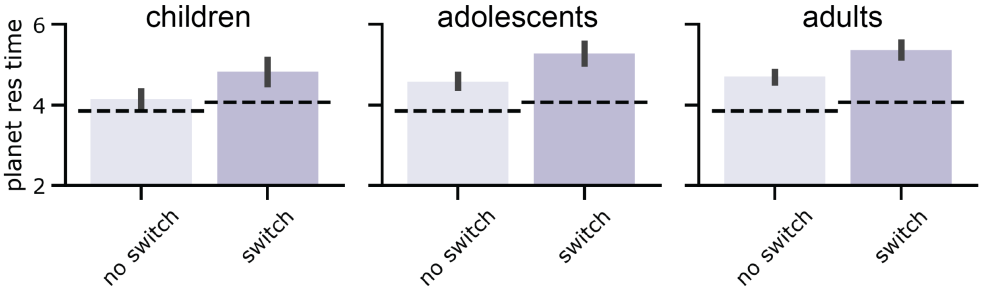

Figure 3.4: Signatures of Uncertainty Adaptive Planning. If participants were to engage in uncertainty adaptive planning, then they should shorten their planning horizons when they are uncertain about the environment, and consequently, stay longer with the current option. Consistent with this, participants in all age groups increased the extent of their overharvesting following a relatively rare switch in planet type.

Uncertainty adaptive planning emerges early in development Under our model's uncertainty-adaptive planning mechanism, foragers should adjust their planning horizon with respect to their uncertainty over the environment's structure (Fig 3.4). If participants adapt their planning in this way, we would expect their overharvesting to increase at planet type switch points. To test this, we examined the effect of switch point, planet number, age and their interactions on deviance from MVT optimality, including participant-level random intercepts and random slopes for planet number. Indeed, we found that participants overharvesting did increase at switch points (switch point: β=0.31, $p < .001$), and the extent of this marginally diminished with task experience (switch point x planet number interaction: β=-0.078, $p = .065$). We did not observe any age-related differences in overharvesting at switch points (age x switch point interaction: β=-0.0094, $p = .81$). However, for older participants, the impact of switch points on their overharvesting diminished with experience (age x switch point x planet number: β=-0.11, $p = .0082$), potentially further substantiating their learning of the environment's structure and dynamics.

3.3.2 Model-based analyses

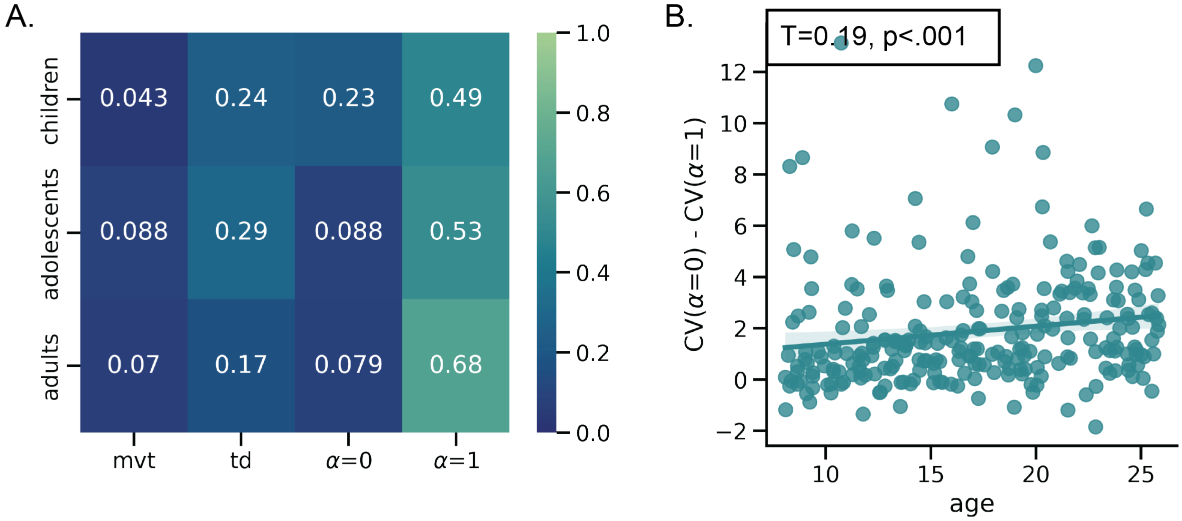

Figure 3.5: Model-based results. A. The proportion of participants in each age group best fit by a model based on cross-validation score. In all age groups, the structure learning model with α fixed at 1 provided the best fit for the greatest proportion of participants. However, this model's advantage was greatest in the adults. **B.** To quantify the relative fit of the structure learning model when α was fixed at 0 versus when it was fixed at 1, we took the difference of their cross validation scores. A positive value would indicate the participant was better fit by the $\alpha = 1$ model. Correlating the difference with age, we found that the extent the $\alpha = 1$ model better explained a participant's choices increased with age (Kendall's τ=0.18, $p < .001$).

Model comparison revealed that the $\alpha = 1$ model provided the best fit for the greatest proportion of participants in all age groups (children: 49%; adolescents: 53%; adults: 68%). Notably, however, a smaller proportion of children were best fit by this model. To explore this further, we took the difference in cross-validation scores between the $\alpha = 0$ and the $\alpha = 1$ models for each participant. A positive value would indicate the the participants' choices were better described by the $\alpha = 1$ model. We found that the difference in scores grew increasingly positive with age (Kendall's $\tau = .19$, $p < .001$), further substantiating an improvement in structure inference with age.

Because the $\alpha = 1$ model provided the best account of a plurality of participants in each age group, we next examined how its free parameters varied with age. We found that the baseline discounting factor decreased with age, γ_{base} (τ=-0.15,$p < .001$), suggesting that additional factors beyond structure learning contributed to younger participant's more exploratory tendencies. We also found that the scaling factor applied to drift rate increased with age (τ=0.20,$p < .001$). This suggests older participants accumulated decision-relevant information more quickly, potentially as a consequence of applying more attentional resources. We did not find a significant relationship between age and the parameter modulating discounting with uncertainty, γ_{coef} (tau=0.037,$p = 0.38$).

3.4 Discussion

Exploration has primarily been studied within cognitive science through the lens of identifying the algorithms explorers use, and how the use of different algorithms changes from childhood into adulthood. In this study, we asked if developmental shifts in exploration could be explained by differences in another potential contributor, structure inference. To address this question, participants completed a patch foraging task set within a richly structured environment, affording multiple possible representations to infer and plan over.

Across all age groups, participants adapted their exploration to the richness and volatility of the environment. Their choice patterns, under our model, were consistent with the predictions of our structure learning model, which attributes overharvesting to the process of ascertaining the environment's structure. Over the course of the task, participants' decisions became increasingly aligned with MVT-optimality, suggesting that they discovered the environment's true generative structure. This was further substantiated by their response times.

In line with prior work [88], we found that children and adolescents explored more than adults. This was particularly evident on the richest planets. The predictions of our model indicate that these differences in exploration stemmed from differences in representation.

More specifically, older participants' choices implied use of a more complex, and accurate representation of the environment. Younger participants' choices were consistent with use of a simplistic, single- planet-type representation, although their response times revealed awareness of different planet types. Taken together, these findings in younger participants point to a gap between knowledge and its influence on behavior.

In multi-armed bandit settings, prior work has sought to characterize developmental differences in exploration in terms of changes in algorithm: identifying differences in which decision variables are maintained (e.g. information's utility) and how variables are integrated during action selection (e.g. is information gain or reward gain more important?). Younger decision makers tend to broadly sample their options, directed by their own uncertainty while older decision makers strategically constrain their exploration [168, 149], seeking out the information with the highest future utility [185]. In spite of its lower efficiency, children's broad exploration strategies can confer greater flexibility and aid in the discovery of the environment's underlying structure [26, 192, 121].

In our patch foraging task, we found that younger participants could leverage uncertainty to guide their exploration like adults [143, 26]. But, we did not find that this uncertainty-guided behavior was necessarily coupled with the formation of a veridical representation of the environment. This could be due to a difference in task structure, and which actions are uncertainty-resolving within them. In a multi-armed bandit task, a small set of options with different values are repeatedly sampled from. A broader sampling strategy enables the learner to detect changes in the highest-valued option over time. In a patch foraging task in which patches are distinguished by their depletion rate, staying in a patch longer — often considered as the "exploit" action — can reduce uncertainty. The longer a forager stays in a patch, the more observations they gather that inform their inference of the patch's type. This, in turn, informs inference of the distribution of patch types across the environment. In such a task, the distinctions between "explore" and "exploit" are not as clear-cut.

We observed that younger participants relied on more simplistic representations to guide their exploration. These findings align with prior work demonstrating that structure learning ability emerges early in development but continues to strengthen and be refined into young adulthood. Even from infancy, we have the ability to extract temporal regularities to build structured, hierarchical representations that can be generalized to novel contexts [163, 201]. Conversely, abilities requiring more challenging inferential leaps, such as relational inference [166] and counterfactual updating [147], show a more protracted developmental trajectory.

Our model casts age-related differences in structure learning not as differences in ability, but rather as differences in prior experience. Its key parameter, alpha, controls the prior expectation of encountering a new planet type. If we define environment complexity in terms of the number of distinct planet types, alpha reflects the prior expectation of complexity. Given that children and adolescents often have less diverse and varied experiences than adults, it's reasonable that they would anticipate environments to be more simply structured. Similar structure learning models have been used to develop an account of how children come to understand that others have preferences different from their own [73]. They too model age-related improvements in understanding as an increase in alpha. Collectively, our results suggest that across development children construct more complex and flexible models that they deploy during decision making.

Although younger participants choices' were consistent with a simplistic, single- planet-type representation, they demonstrated decisional slowing following sudden, rare changes in planet richness, suggesting awareness of the environment's true multi-planet type structure. Notably, stay-leave decisions rely on using representations to form a prediction of future outcomes while slowed decision times are a reaction to unexpected change. A growing body of work has demonstrated a similar developmental dissociation between the acquisition and use of structure representations. For instance, in a task measuring planning over a mental model of the environment, children's choices showed less use of the model while their reaction times revealed knowledge of its probabilistic transition structure and even further, they were able to report this structure when asked [47, 144, 153] In another task, children were able to acquire the causal structure of the environment, but did not deploy this knowledge to guide their reinforcement learning computations [37]. Perhaps, while a more complex representation of the environment can be transiently invoked, only a simpler representation can be recruited and maintained to guide choice [135].

Effective representations of the environment provide the foundation upon which learning and decision making algorithms can act. However, the simplicity of environments in tasks commonly used to study exploration has limited our understanding of the role of structure inference in shaping individuals' exploration. By using a patch foraging task set within a richly structured environment, intended to more closely mimic natural environments, we have demonstrated how individuals' prior expectations of environmental structure influence their current structure inference. These findings pave the way for future research exploring the complementary roles that algorithms and representations play in shaping how individuals explore their environments across the lifespan.

Chapter 4

Interval timing as a computational pathway from early life adversity to affective disorders

The contents of this chapter were published in Harhen, N.C., Bornstein A.M. Interval timing as a computational pathway from early life adversity to affective disorders. *Topics in Cognitive Science* (2024).

4.1 Introduction

Across development, brain circuits adapt to reflect the environment's structure, preferentially encoding more frequent aspects of the world. The statistics of the early life environment tune sensory receptive fields, producing non-homogeneous sensitivity to perceptual stimuli and determining discrimination abilities in adulthood [195, 56]. Early consistency in these sensory inputs are crucial for the future functionality of involved circuits [119]. Similar developmental processes may take place in reward and memory systems, those underlying associative learning, implying that the consistency or predictability of associations in early life may shape the acquisition of associations later on [23].

Caregivers are primary contributors to the associative structure infants encounter. Associations may take the form of a caregiver's response to an action the infant preforms. These responses may vary in their valence and predictability. Valence influences whether the infant will repeat the action preceding the response, while predictability constrains the infant's learning to associate the two. Prior work has largely focused on the effect of valence on later child mental health outcomes [189, 137, 18, 84]. However, recent work has highlighted how early life unpredictability, or ELU, may also contribute [13]. Research done in animals has illustrated that offspring exposed to unpredictable caregiver signals show a reduction in motivation and the experience of pleasure, characteristics of the trans-diagnostic symptom, anhedonia [28]. Work in humans accords with these findings, showing relationships between experiences of early life unpredictability, reduced reward anticipation, and symptom severity in anhedonia, depression, and anxiety [85, 50, 130, 78, 186].

Here, we propose that the study of early-life unpredictability can be understood in part via its influence on the development of temporal representations (TRs) that serve as basis sets for associative learning more generally [98, 94]. TRs capture the intuition that the strength of learned associations is dependent on the time between events [12]. These tuning curves are similar to those found in sensory areas, but rather than being tuned to visual angle or auditory pitch, are sensitive to the temporal duration between related events.

We specifically examine how early-life unpredictability can, via its influence on the adaptation of temporal representations, result in an anhedonic phenotype. We extend a principled computational model of interval timing [124] to simulate how enhanced volatility during an early period of heightened plasticity can, with minimal assumptions, affect later predictions of reward during maturity. With this model, we formally demonstrate that early unpredictability in timing, and adaptation of temporal representations to this timing, can lead to the development of several defining characteristics of anhedonia – including slowed associative learning, reduced motivation, and a bias towards learning from negative events – in the absence of differences in the overall amount of reward. Our results reproduce empirical findings that unpredictability in early life experience can heighten susceptibility to poor mental health outcomes even after controlling for the childhood environment's overall resource availability [77].

While we show that a singular type of adversity can alone produce an anhedonic phenotype, in the real world, individuals are often subject to multiple adversities. Modeling the nature of these interactions and their combined effect on learning will be critical for characterizing the developmental trajectory of psychopathology. As a first step, we model how temporal unpredictability interacts with the environment's availability of reward, or richness, to shape later learning and expectations of reward. Under the common cumulative risk approach to conceptualizing and measuring early life adversity [61], these two adversities are assumed to have an additive effect on development: individuals facing both are predicted to have the most negative outcomes. Our model predicts that unpredictability always has a negative effect on associative learning, however, contrary to the cumulative risk prediction, this effect is most pronounced in richer environments. Both unpredictability and an abundance of rewards individually alter temporal representations to be more expansive or diffuse, producing the observed interaction. Our results highlight the potential value of computational psychiatric approaches to tackling the heterogeneity of early life adversity and making sense of its developmental consequences.

4.2 Isolating the contributions of one form of adversity, unpredictability

4.2.1 Methods

During the initial phase ("critical period"), agents' temporal representations were allowed to adapt to the environment's temporal statistics. Agents belonged to one of two groups, early life unpredictability (ELU) or control. The two groups were differentiated by the distributions their reward timings were sampled from, with the ELU agents' distribution having the same mean as the control agents' but a higher variance. In the second phase ("post critical period"), both groups received reward at the same time step on each rewarded trial and, critically, agents' temporal representations were not allowed to adapt to the novel environment's statistics.

The Temporal-Difference Learning Model

Temporal-Difference (TD) models aim to accurately estimate the value of world states, V, in terms of the future rewards they predict. Time is explicitly represented in these models with each time step identifying a world state.

$$V^* = E[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}] \tag{4.1}$$

where r_t is the reward received at the current time step, and γ is a parameter controlling how heavily future rewards are discounted. Future rewards are less influential on the estimation of V when γ is low. A TD agent learns V via an error driven learning rule — the difference, δ_t, between the reward received ($r_t + \gamma V_t$) and the previously predicted reward (V_{t-1}) is used to update the estimate of V at the next time step.

$$\delta_t = r_t + \gamma V_t - V_{t-1} \tag{4.2}$$

Microstimulus Representation of Time

All TD models explicitly represent time, but do so in various ways. Basic TD models use a complete-serial-compound (CSC) representation in which each time step is treated as independent from one another and agents are assumed to have perfect knowledge of when events occur. This representation prohibits temporal generalization, creating issues in environments where the time between cue and reward varies. The microstimulus representation addresses this problem by relaxing its temporal markers [124]. CSC's discrete markers are replaced with continuous "microstimuli" which allow for temporal uncertainty to be represented (Figure 4.1). A stimulus is assumed to leave behind a memory trace that decays with time. The trace is approximated by a set of Gaussian temporal basis functions uniformly distributed across the heights of the memory trace. This approximation produces a set of microstimuli increasing in their peak and width from the time of stimulus onset.

$$f(y, \mu, \sigma) = \frac{1}{\sqrt{2}\pi} e^{(-\frac{(y-\mu)^2}{2\sigma^2})} \tag{4.3}$$

A time step's value, V_t, is estimated as the weighted average of the microstimuli.

$$V_t = w_t^T x_t = \sum_{i=1}^{n} w_t(i) x_t(i) \tag{4.4}$$

V_t is compared to the reward received to compute an error term, δ_t that is used to adjust the weights on the microstimuli. Adjusting the weights updates the predicted value at the next time step.

$$w_{t+1} = w_t + \alpha \delta_t e_t \tag{4.5}$$

α is the learning rate controlling the time window over which experiences are integrated. e_t is a vector containing each stimulus's eligibility traces.

$$e_t = \gamma\lambda e_t + x_t \tag{4.6}$$

Following the stimulus, its eligibility trace decays at a rate determined by γ and λ. γ is a temporal discounting factor as it was for the TD model with a CSC representation, while λ controls the time window over which a stimulus can induce learning within a trial. For all simulations, we use the parameter settings from Ludvig et al, 2008 — $\alpha = 0.01$, $\gamma = 0.98$, $\lambda = 0.95$, $n = 50$, and $\sigma = 0.08$.

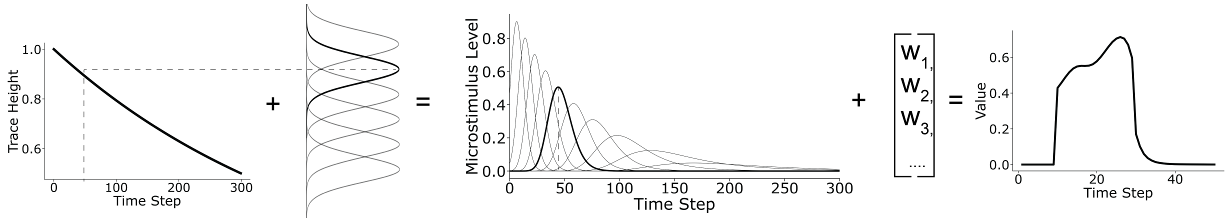

Figure 4.1: Stimulus encoding by microstimuli. From left to right, the memory trace produced by a stimulus is approximated with a set of temporal basis functions, whose centers vary such that they evenly cover the trace's possible heights. The decaying nature of the memory trace produces microstimuli that become shorter and wider the further their center is to the stimulus onset. The microstimuli are weighted and averaged to estimate the future expected reward following the stimulus. The weights can be adjusted with experience to support accurate predictions of reward.

Simulating Development

To model developmental changes in learning, we restrict the updating of microstimuli weights to the initial period which we treat as a "critical period" during which the temporal representations are tuned to support accurate estimation of V. This adaptation process is designed to mimic the observed tuning of sensory receptive fields during analogous sensitive periods of development [178]. During the second phase ("post critical period"), the weights are frozen, prohibiting representation adaptation, to simulate adulthood.

We simulated two groups of agents learning to associate a cue with reward across the two phases (Figure 4.2). One group of agents, the early life unpredictability (ELU) group experienced a volatile critical period environment in which the timing of reward was much more variable than the timing experienced by the control group. Critically, however, the average timing of reward and the average amount of reward received (i.e. same probability of reward on each trial) was matched between groups.

On each of the 1000 simulated trials during the critical period, a cue was always presented at 10 time steps and there was a 75% probability of a reward following it. If a cue was followed by reward on a trial, the timing of reward was sampled from a normal distribution with $\mu = 30$ and truncated at 10 and 70 time steps. σ varied between agents. For agents in the ELU group, σ was sampled from a zero-truncated normal distribution with $\mu_{hyper,elu}$ $= 10$ and $\sigma_{hyper,elu} = 3$. The control group experienced much less variability, with σ being sampled from a zero-truncated normal distribution with $\mu_{hyper,control} = 1$, $\sigma_{hyper,control} = 2$. We varied σ within groups to reflect the variation observed in real life samples, particularly early life adversity facing ones, and to ensure our results were robust to such variation.

In the second phase, the microstimuli weights were frozen ("post critical period"), allowing us to directly examine the influence of highly variable early-life experiences. The temporal statistics of this environment differed from the critical period's environment in two ways: 1. The reward was delivered at the same time step each trial for both groups of agents. 2. This time step was later (50 time steps) than the previous environment's average time of reward (30 time steps). By testing ELU agents' learning in novel environments that are more stable than the environment they "developed" in, we formalize the Mismatch Hypothesis of Early Life Adversity and Depression [167]. Under this hypothesis, depression and other mental illnesses are proposed to be the byproduct of a mismatch between the developmental environment to which neural systems are optimized for and the later adulthood environment. We were particularly interested in characterizing how an agent's early adaptation to unpredictability would affect their response to uncertainty in adulthood. Within the simulated task, uncertainty should rise once the mean time of reward has passed and reward has failed to be delivered. This is because it becomes unclear whether the reward is delayed or is omitted altogether on the trial. To produce this circumstance, we moved back the time step of reward in the novel, post critical period environment to examine how the ELU and control groups differ in their response to reward and its omission following a period uncertainty. All agents completed 2 trials. On both trials, the cue arrived at 10 time steps. On one trial, reward followed the cue at 50 time steps. On the other, reward was omitted. We simulated agents only on two trials because the weights were no longer updated. Thus, the prediction error response on every trial of the same time (rewarded vs. omitted) would be identical.

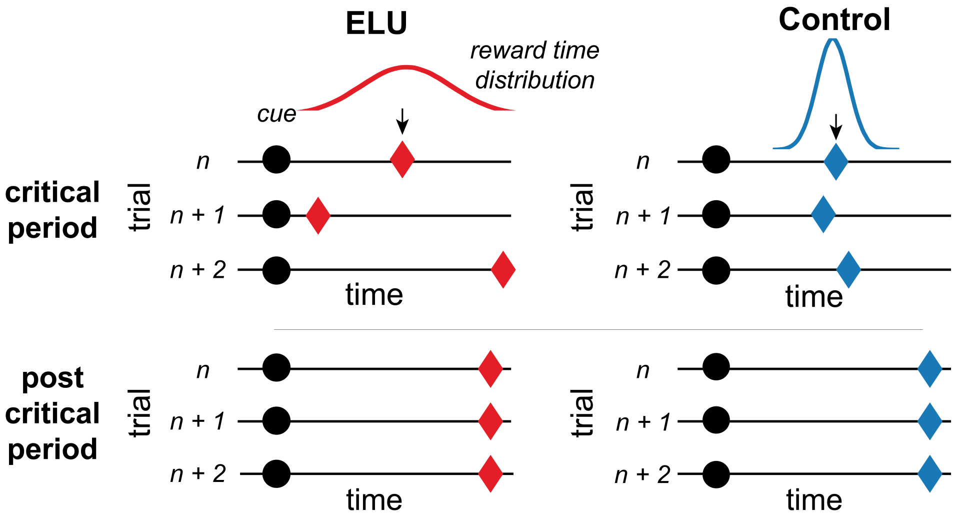

Figure 4.2: Simulated agents learned to associate a cue with reward in two different environments. The cue was partially reinforced — 75% of the time in the initial environment and 55% in the second. On rewarded trials, reward was delivered at a variable time step. Agents belonged to one of two groups, differing in the variability they experienced in the initial environment. The reward timings experienced by agents in the early life unpredictability (ELU) group were on average the same as those experienced by the control group. However, in the initial phase ("critical period"), they experienced more variably timed rewards trial to trial. In the second phase ("post critical period"), agents' weights were frozen, and all agents received reward at the same time step.

Statistical Analyses

Each simulated agent encountered a different sequence of reward timings during the initial critical period. Thus, a potential concern is that our results are largely driven by a subset of simulated agents. To assess the reliability of the relationship between prediction error magnitude and unpredictable experience, we performed a bootstrap analysis across agents within a group [106, 29]. For each group, we sampled agents with replacement until we reached the total number of agents, 100. We then computed the test statistic for a two sample t-test with the selected groups. We repeated this procedure 1,000 times to obtain a distribution of test statistics across shuffled permutations of the simulated groups. This re-sampling procedure provides a p-value that is the fraction of test statistic values with a different sign from the base effect size (the test statistic for the original two groups). We also computed the Cohen's d in order to evaluate the size of the difference between simulated populations. By convention, effect sizes greater than 0.80 are considered "Large", and thus reliable [38].

4.2.2 Results

Critical Period

First, we validated that the critical period environment shaped temporal representations by comparing the groups' microstimuli weights at the end of the critical period. For each agent, we computed a temporal imprecision measure by taking a weighted average of the microstimuli's standard deviations, with the weights being the same as those used to generate the value signal. Consistent with our prediction that temporal representations would adapt to reflect the statistics of their environment, we found that the ELU group relied on more broadly-tuned temporal representations relative to controls (Figure 4.3; $t(198) = 8.43$, $p <$

.001, Cohen's $d = 1.19$).

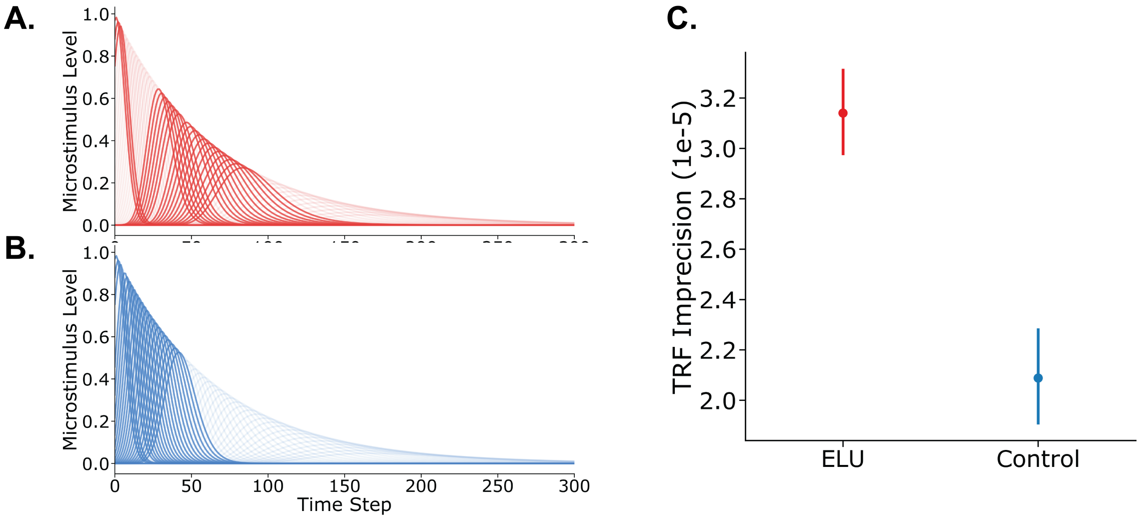

Figure 4.3: A-B. Positively weighted microstimuli. With experience, the ELU group grew to more heavily weigh delayed, imprecise microstimuli to account for the frequent delayed rewards. **C. Temporal Imprecision.** We computed a summary statistic of temporal representation (TR) imprecision by taking a weighted average of the standard deviations of the positively weighted microstimuli at the end of the critical period. ELU agents' temporal representations were, on average, more than twice as imprecise as control agents.

Early life unpredictability has been shown to produce slower learning from reward in adulthood [22, 50]. We next examined the model's ability to capture this. As a proxy for learning, we used a particular pattern of prediction error responses. If a cue has become associated with reward, then there should be large positive prediction error in response to the cue, a smaller positive prediction error at the time of reward, and a large negative prediction error when reward is omitted. To compare prediction errors between groups, we computed, across time within each trial, the prediction error extremum for each agent. On rewarded trials, the maximum prediction error magnitude following the cue was taken and on omission trials, the minimum was taken. We then took the average of these values across trials of the same type for each participant. We found that, on rewarded trials, the ELU group's positive prediction errors were larger than the control group's (Figure 4.4, 4.5; $t(198) = 12.59$, $p < .001$, Cohen's $d = 1.78$) but, were less negative on omission trials ($t(198) = 6.23$, $p < .001$, Cohen's $d = 0.88$). Despite both groups experiencing the same amount of reward, the ELU group showed slower learning under reinforcement. Collectively, these results demonstrate how impaired associative learning, as observed in anhedonia, can emerge from experienced temporal volatility alone during a period of plasticity.

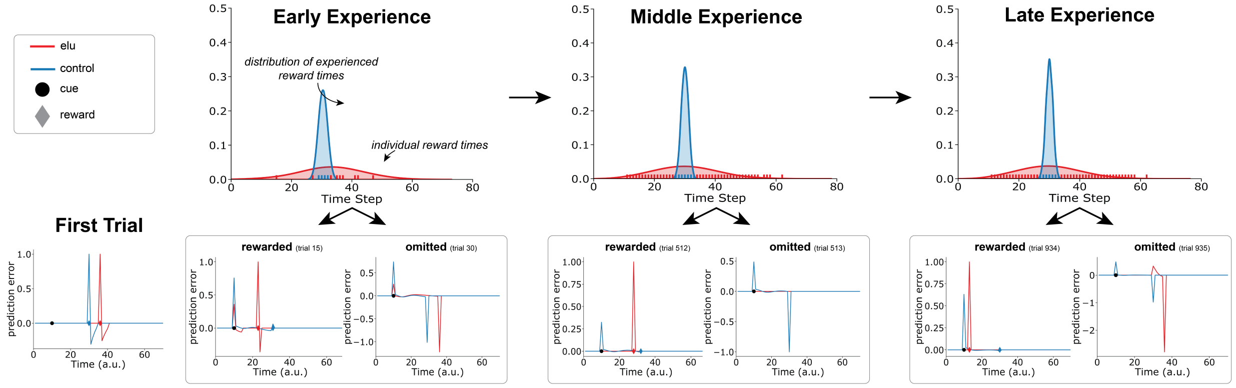

Figure 4.4: An example ELU and control agent's prediction errors (δ) from individual trials within the critical period. A cue always occurred at 10 time steps, while the reward's timing varied from trial to trial. Temporal variability was determined by which group an agent belonged to — an ELU agent experienced a much wider distribution of reward times. Reward elicited a strong positive prediction error from both agents on the first trial. Even very early on, the control agent demonstrated a positive prediction error in response to the cue, a weak positive prediction error at the time of reward, and a strong negative prediction error when reward was omitted, matching the pattern of responses expected for well-learned, consistent contingencies using this temporal-difference learning rule. This pattern held throughout the 1000 trial critical period. In contrast, even very late into the critical period, the ELU agent's prediction errors continuously moved around in time and were larger in magnitude, a consequence of their more volatile environment.

Early life unpredictability has also been shown to impair motivation [86], potentially stemming from reduced expectations of reward. Thus, we next compared the groups' expectations of future reward, as reflected by their value signals. When averaged across trials, control agents' value signals quickly increased in response to the cue (Figure 4.6; mean at 10 time steps = 0.43, sd = 0.022), gradually rose until the average time of reward (mean at 26 time steps = 0.71, sd = 0.075) after which the signal rapidly dropped off (mean at 32 time steps = 0.059, sd = 0.078). ELU agents' value signals similarly rose in response to the cue but peaked much earlier ($t(198) = -27.75$, $p < .001$, Cohen's $d = -3.92$) and fell more gradually (mean at 32 time steps = 0.29, sd = 0.045, $t(198) = 26.34$, $p < .001$, Cohen's $d = 3.73$). Importantly, ELU agents' expectations of reward were diminished at the time steps right be-

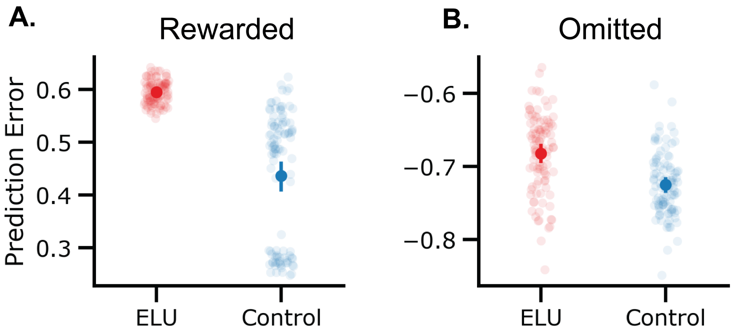

Figure 4.5: Critical period prediction error signals. Reward elicited larger positive prediction errors in ELU agents while reward omission produced weaker negative prediction errors, a pattern of responses suggesting ELU agents were slower in learning from reward.

fore when reward as most likely (mean at 26 time steps = 0.48, sd = 0.048, $t(198) = -25.87$, $p < .001$, Cohen's $d = -3.66$). These differences could have a particularly significant impact on decision making which requires deciding not only *which* option to take but also *when* to take it. Diminished expectations of reward should produce slower decision times, a characteristic found in anhedonia [54, 80, 202, 46]. ELU agents also showed greater variability in their value signals from trial to trial as revealed by taking the standard deviation of the time steps at which value signals peaked (ELU mean = 10.49; Control mean = 1.50; $p < .001$, Cohen's $d = 3.39$). This aligns with prior empirical work that found more variable ventral striatal activity following early life stress [85].

Post Critical Period

To simulate adulthood, in the second phase, we closed the "critical period" by preventing the updating of the microstimuli weights in the novel environment. Thus, their expectations of reward are carried over and fixed once the developmental period ends. In this environment, reward was delivered at a later time step than the average time of reward during the critical period. This induces an interval of uncertainty during which its unclear whether the reward is delayed or omitted. We examined how the expectations acquired in an unpredictable early life environment shape the prediction error response when this uncertainty is resolved.

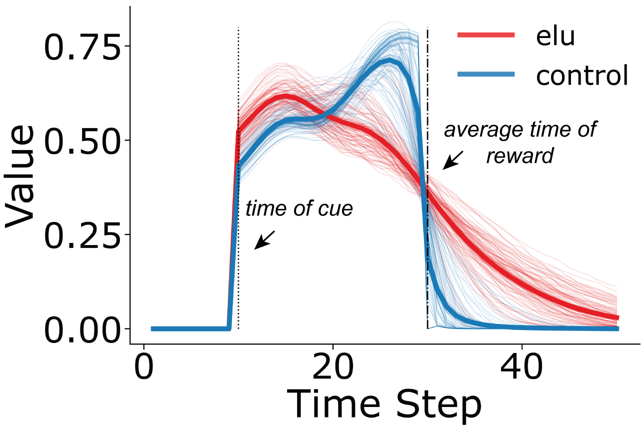

Figure 4.6: The value signal, V, averaged over all critical period trials. Individual agents' value signals are depicted by the thin lines. The thicker lines depict the group averages. Control agents' expectations of future reward quickly rose following the cue and steadily increased until the average time of reward, after which their expectations quickly dropped. ELU agents' expectations of reward similarly rose in response to the cue but subsequently decreased at a gradual rate rather than increasing. Notably, ELU agents had higher expectations of reward at later time steps compared to controls — a consequence of having experienced more delayed rewards which required relying on more diffuse, later peaking microstimuli. When aggregated across trials, ELU agents' expectations were more spread out. This is both because they relied on more diffuse microstimuli and because their value signals fluctuated from trial to trial in response to variably timed rewards.

Because ELU agents experienced rewards at more variable time steps, they grew to have a higher expectation that reward could arrive at later time steps (Figure 4.7A). This affects their response to the cue and reward. Control agents have a strong positive prediction error immediately after the cue is presented because they have learned well that the cue predicts reward (Figure 4.7B). ELU agents instead have a weaker and delayed response to the cue because of their weaker association between the cue and reward. Control agents experience a slightly negative prediction error when reward is not delivered at the most expected time step (Figure 4.7C). But, when reward ultimately arrives at a later time step, they show a large positive prediction error, a consequence of their low expectations of reward this late in the trial. ELU agents had relatively higher expectations of reward at the time step when reward was delivered, thus they showed relatively blunted positive prediction errors ($t(198) = -2.25$, $p < .001$, Cohen's $d = -0.32$). The same expectations produced amplified negative predictions error when reward was omitted ($t(198) = -12.29$, $p < .001$, Cohen's

$d = -1.74$). In other words, their higher expectations allowed them to experience greater disappointment.

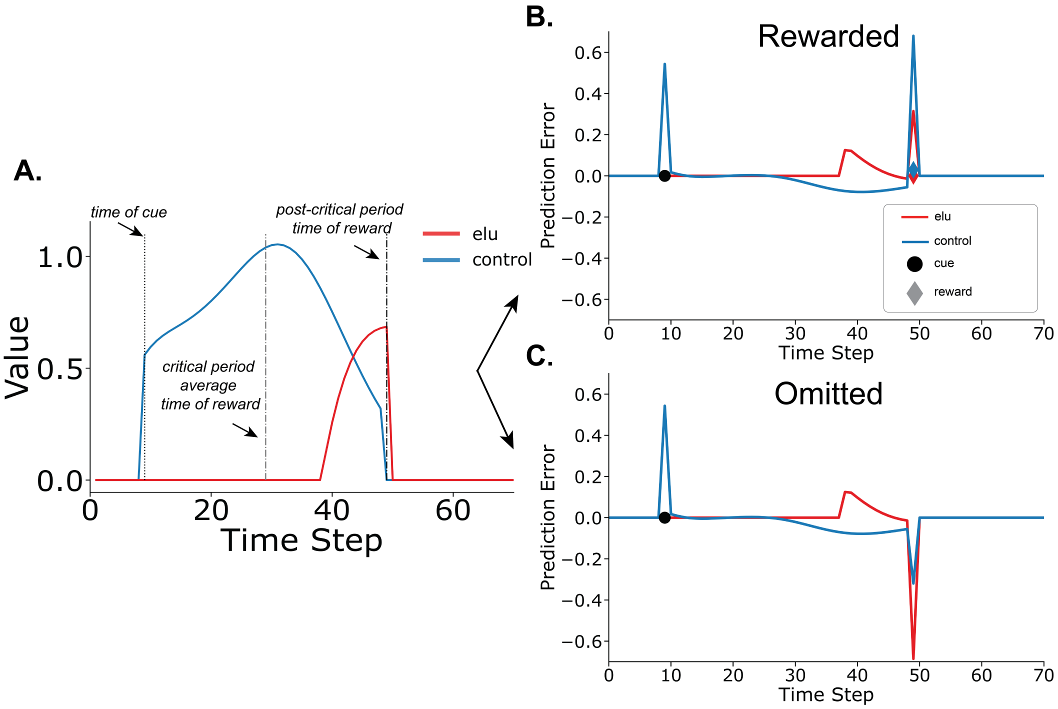

Figure 4.7: A. Representative agents' value signals. The value signal, taken from the end of the critical period, reflects the individual agent's expectation of future reward following the cue. These expectations are "frozen" and determine the agent's response to reward and its omission. **B. Example prediction error signals for a single rewarded trial.** The ELU agent's expectation of future reward only begins to rise at 40 time steps whereas the control agent's rises immediately at 10 time steps in response to the cue. Accordingly, the ELU agent demonstrates a weaker and delayed response to the cue. When reward is delivered at 50 time steps instead of its average previous time, 30 time steps, the control agent shows a more positive prediction error than the ELU agent. Again, this is a result of their expectations. The control agent does not expect the reward to arrive this late in the trial, and thus, is surprised when it does. The ELU agent, having experienced more delayed rewards, is less surprised. **C. Example prediction error signals for a single omission trial.** The ELU agent's greater expectation of reward at later time steps also produces a larger negative prediction error when reward is omitted.

We next examined how early life unpredictability affected agents' response to rewards of varying magnitudes. When given a reward of the same magnitude as those received during the critical period, control agents responded with larger positive prediction errors (Figure 4.8; $\beta_{elu} = -0.51, p < .001$). As the reward magnitude increases, diverging from those previously experienced, both groups show increasingly large prediction errors ($\beta_{magnitude} = 0.55, p <$

.001). The ELU agents do so at a faster rate than control agents, demonstrating larger prediction errors than controls in response to higher magnitude rewards ($\beta_{elu*magnitude} = 0.43, p < .001$). When coupled with their blunted response to the cue, ELU agents appear to be hyposensitive to rewards in anticipation but hypersensitive to them in consumption. This pattern has been observed in a monetary incentive delay task designed to distinguish between reward anticipation and consumption [27]. More generally, it concords with widespread findings that early life adversity impairs cue-reward learning [50, 191, 22, 49] while increasing sensitivity to dopamine-releasing drugs [110, 199, 42, 146, 111, 112, 209].

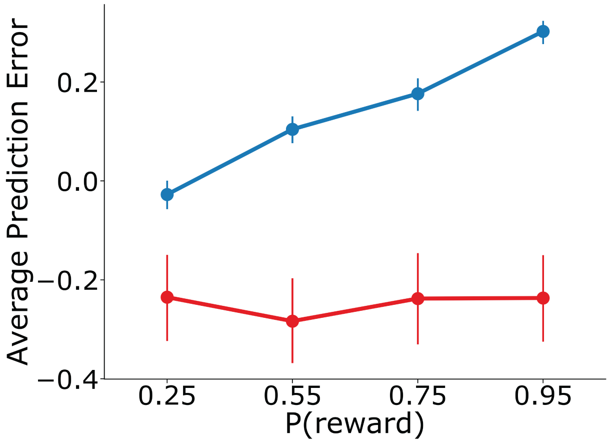

Figure 4.8: Sensitivity to increasing rewards. We varied the magnitude of rewards delivered during the second phase. As the magnitude of rewards increased, both groups showed larger positive prediction errors on rewarded trials. ELU agents were more sensitive to changes in reward magnitude – their prediction errors increased to a great extent in response to larger rewards. At the lowest reward magnitude, which was the magnitude experienced during the critical period, the control group experienced larger positive prediction errors than the ELU. This pattern reversed at larger magnitudes with ELU agents demonstrating hypersensitivity to rewards. Error bars are 95% bootstrapped confidence intervals.

Prediction error magnitude determines the extent to which an agent learns or updates their expectations. Because valence asymmetries in learning have been proposed to be clinically relevant [162, 151, 161], we next compared prediction error magnitude on the rewarded and omission trials to probe for such asymmetries. We computed an asymmetry index for each agent as follows:

$$index = \frac{|PE_{+}| - |PE_{-}|}{|PE_{+}| + |PE_{-}|} \tag{4.7}$$

ELU agents' asymmetry indices were overall negative (Figure 4.9; $t(99) = -5.62$, $p < .001$, Cohen's $d = -0.79$) while the control agents' were positive ($t(99) = 8.49$, $p < .001$, Cohen's $d = 1.20$). Because prediction error magnitude enhances learning and memory, this suggests that negative events would have an outsized influence on ELU agents, making their value estimates overly pessimistic while control agents' are overly optimistic [171]. Our model provides a mechanism through which both of these biases could emerge under minimal assumptions.

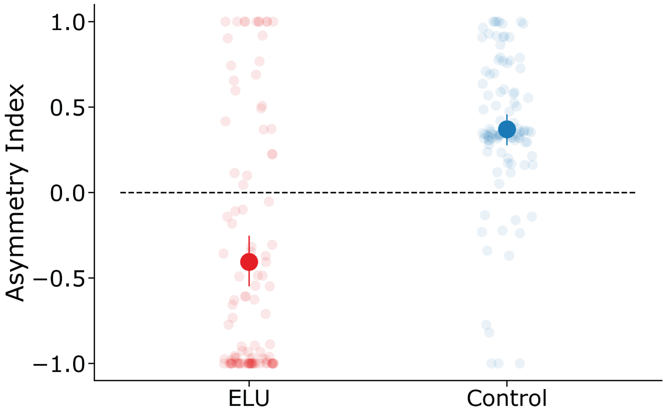

Figure 4.9: Learning Asymmetry Indices. The ELU group showed a negativity bias, experiencing more extreme prediction errors on the omission trial than the rewarded trial. In contrast, the control group demonstrated a positivity bias, experiencing larger prediction errors on the rewarded trial. Error bars are 95% bootstrapped confidence intervals.

4.3 Interactions between multiple forms of adversity

4.3.1 Methods

Critical Period

To examine the interaction between multiple forms of early life adversity — temporal unpredictability and low reward availability, we additionally manipulated the richness of the critical period environment and observed its effect on both groups. This allowed us to test the assumptions of the cumulative risk conceptualization of early life adversity which assumes an additive effect of adversities on developmental outcomes. We simulated groups of the ELU and control agents in environments with 25, 55, 75, and 95% probability of reward. As in previous simulations, the time of reward delivery was sampled from a normal distribution with $\mu = 30$ time steps and truncated at 10 and 70 time steps, and the distribution's σ differed between groups — ELU agents' σ were sampled from a zero-truncated normal distribution with $\mu_{hyper,elu} = 10$ and $\sigma_{hyper,elu} = 3$ and controls' were sampled from a zero-truncated normal distribution with $\mu_{hyper,control} = 1$, $\sigma_{hyper,control} = 2$.

Post Critical Period

In the novel environment during the second phase, the cue was presented at 10 time steps on each trial. They experienced one rewarded and one omission trial. On the rewarded trial, reward was delivered at 50 time steps. As before, we only include two trials because the weights are no longer updated, thus, the response on each trial of the same type would be identical.

4.3.2 Results

Critical Period

As before, we assume that the smaller positive prediction errors are in response to reward and the larger negative prediction errors are in response to its omission then the more strongly an agent has learned to associate a cue with reward. Under this assumption, both temporal unpredictability and low reward availability were found to slow associative learning. On rewarded trials, positive prediction errors were larger for ELU agents and both groups' prediction errors became weaker with environment richness (Figure 4.10A;

$\beta_{elu} = 0.057, p < .001, \beta_{rich} = -0.90, p < .001$). On omission trials, negative prediction errors were stronger for control agents and with increasing environment richness (Figure 4.10B; $\beta_{elu} = -0.022, p = .015$, $\beta_{rich} = -0.98, p < .001$). The two dimensions interacted, with the difference between groups increasing as environment richness increased ($\beta_{elu*rich} = 0.15, p < .001$). In particular, the effects of unpredictability on learning were only observed in richer environments, with no main effect of group but an interaction effect between group and richness ($\beta_{elu} = -0.015, p = .11, \beta_{elu*rich} = 0.093, p < .001$). Taken together, our results reveal that the effect of temporal unpredictability is most fully felt when reward is abundant, a consequence of both dimensions increasing the imprecision of temporal representations ($\beta_{elu} = 1.04e-05, p < .0001, \beta_{rich} = 1.47e-05, p < .001, \beta_{elu*rich} = -1.91e-07, p = .95$). When rewards are both unpredictably timed and abundant, it increases the range of timings an agent's representation must accommodate.

The value signal reveals a similar impact of the environment's temporal unpredictability and overall richness on learning. The value signal correspondingly increased as richness increased (Figure 4.10C; $\beta_{rich} = 0.32, p < .001$). Yet, only when the environment is sufficiently rich can unpredictability exerts its blunting effect on the signal ($\beta_{elu} = 0.0041, p = .52, \beta_{elu*rich} = 0.035, p < .001$).

Post Critical Period

In the post critical period phase, we found the same complex relationship between the environment's temporal unpredictability and richness, in which greater reward availability allows unpredictability to exert its influence. Across all environments, control agents maintained a bias towards learning from reward over its omission as indicated by positive asymmetry indices (Figure 4.10D; 25% - $t(99) = 21.88$, $p < .001$, Cohen's $d = 3.09$; 55% - $t(99) = 15.79$, $p < .001$, Cohen's $d = 2.23$; 75% - $t(99) = 8.49$, $p < .001$, Cohen's $d = 1.20$; 95% - $t(99) = 8.62$, $p < .001$, Cohen's $d = 1.22$). The valence of ELU agents' biases, in

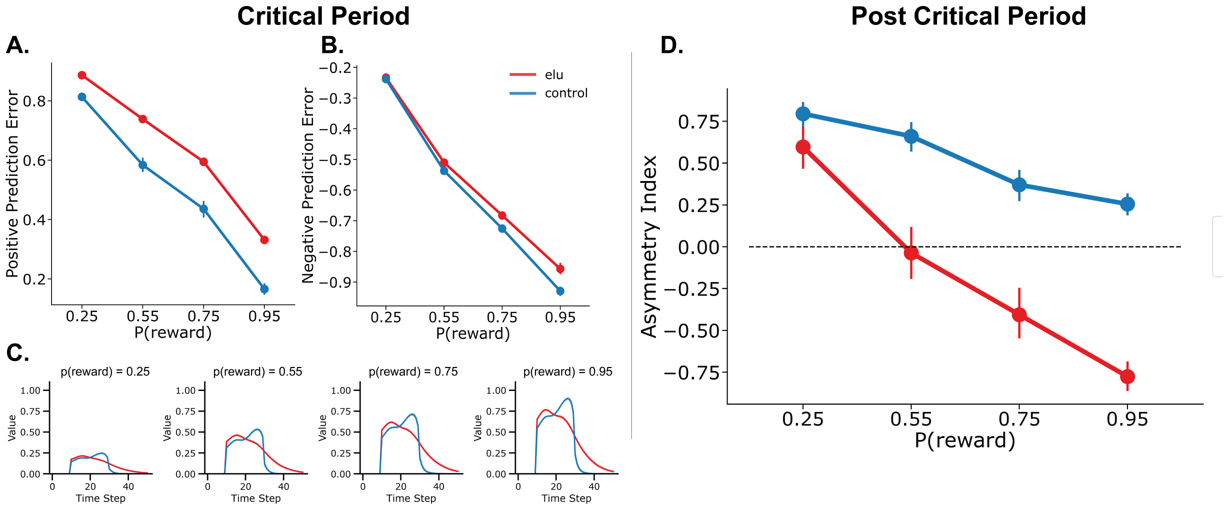

Figure 4.10: Varying critical period environment richness to examine the impact of multiple adversities. A. Critical period prediction errors in response to reward. Positive prediction error magnitude was modulated by the environment's richness (probability of reward) and its temporal unpredictability (ELU vs. Control), with richness attenuating magnitude and unpredictability amplifying it. **B.Critical period prediction errors in response to reward omission.** Negative prediction error magnitude was amplified by richness and attenuated by unpredictability. This pattern of responding suggest richness supports associative learning while unpredictability impairs it. **C. Value Signal.** Mirroring the reward statistics of their environment, agents' expectation of future reward increased accordingly with the overall richness of the environments. Notably, group differences were emphasized by richness. **D. Post critical period asymmetry indices.** Control agents demonstrated a consistent positivity bias that diminished the richer the environment. ELU agents showed a positivity bias only in the poorest environment and a negativity bias in richer environments. Error bars are 95% bootstrapped confidence intervals.

contrast, was dependent on the richness of the critical period environment. ELU agents who experienced the sparsest rewards during the critical period exhibited a positivity bias, similar to control agents although weaker (Figure 4.10D; 25% - $t(99) = 9.098$, $p < .001$ Cohen's $d = 1.28$). Those who experienced a less sparse environment showed no bias (55% - $t(99) = -0.46$, $p = 0.64$, Cohen's $d = -0.065$), and those who experienced an environment abundant with rewards exhibited a negativity bias (75% - $t(99) = -6.60$, $p < .001$, Cohen's $d = -0.79$; 95% - $t(99) = -17.72$, $p < .001$, Cohen's $d = -2.51$). This pattern of results is a byproduct of the reward expectations built up during the critical period. ELU agents whose representations are adapted for richer environments have a stronger prior expectation that reward will have a delayed arrival rather than being omitted altogether. Thus, when reward is omitted on a trial, they experience a particularly large negative prediction error. Our simulations contradict the predictions that would be made under the cumulative risk approach which assumes an additive effect of adversities.

4.4 Discussion

Here, we propose a novel computational link between early life unpredictability and the emergence of anhedonia — the optimization of temporal representations to the early life environment. By simply assuming that temporal representations are adapted to the statistics of the early life environment, several behaviors associated with anhedonia emerge — impaired learning from reinforcement, reduced anticipation of reward, and a greater response to the omission of events.

These findings are consistent with behavioral outcomes observed in the laboratory and clinical settings. One representative set of such findings is of an asymmetric attentional bias in anhedonia. If we treat the omission of reward as a negatively valenced event and the presence of reward as a positive event, this suggests a negative attentional bias in the ELU group and positive bias in the controls, reproducing empirical findings [51, 64]. Larger negative prediction errors may not only affect attention in the moment but also have longer lasting consequences via memory. Surprising events, like prediction errors, are known to be more easily retrieved from memory [162, 180]. This provides a mechanism by which singular negative events can have an outsized influence on expectations and consequently, shape mood over the longer term [57]. Frequent large negative prediction errors could produce the persistent negative mood that characterizes anhedonia [50]. We found that the development of this negativity bias was critically dependent on the overall richness of the environment. To experience a pronounced negative prediction error when reward was omitted, agents needed to have a strong expectation that reward would come but a weak expectation of when that would be. Only in environments rich with variously timed rewards did such expectations emerge.

Our results contradict the assumptions and predictions of the cumulative risk conceptualization of early life adversity [61]. The cumulative risk approach has been crucial in establishing the robust association between negative events early in life and a wide array of negative outcomes later in development. However, aggregating over heterogeneous experiences may obscure the mechanisms linking such experiences to later psychopathology [182, 128]. One proposed alternative are dimensional models which identify influential features of the early life environment on development and seek to characterize how these features exert their influence. Supporting the dimensional approach, recent work has found divergent associations between measures of threat and deprivation in the early life environment with later developmental outcomes including amygdala reactivity to threat, aversive learning, cognitive control, and pubertal timing [114, 126, 131, 160, 175, 193, 194]. However, adopters of these approaches have been criticized for an unprincipled choice of dimensions, particularly lacking neurobiological grounding [182]. Given the potential relevance of reward systems to psychopathology, it may be valuable to look at the statistical properties of the environment known to influence associative learning as potential candidate dimensions.

Thus far in our interpretation of the results, we've treated the cue-paired outcome as reward. However, the model is agnostic to the valence of the outcome — allowing for different interpretations where the outcome is treated as neutral or aversive. Different valences will suggest different behavioral phenotypes. Treating the outcome as aversive, like a shock, the ELU group's prolonged expectation of a negative outcome's appearance could be interpreted as sustained hypervigilance (perhaps akin to a form of "paranoia"), a symptom of anxiety. Treating the outcome as neutral, impairments in associative learning become more general impairments in relational learning. This may explain memory deficits and alterations in hippocampal structure in ELU individuals [81, 133] and anhedonia's associated memory deficits. Prior work has suggested that anhedonia is characterized not only by the inability to experience pleasure in the moment but also the inability to recall past and anticipate future pleasurable experiences [51].

Here, we've only considered the mechanism under Pavlovian learning conditions. However, it also suggests differences in ELU individuals' instrumental learning and action selection. The inability to accurately predict the timing of future outcomes diminishes an individual's perceived controllability of the environment, which has also been implicated in psychiatric disorders such as anxiety [24].

Hidden-state inference models capture a similar idea as the microstimulus model at a different level of analysis [190]. Often, the true state of the world is unknown or hidden and must be inferred from observations. This inference process is in part driven by prediction errors [162], and by extension is more difficult in volatile environments. As a result, ELU individuals may infer fewer states in the world (or, analogously, more states in an environment where negative prediction errors predominate) and group their experiences accordingly as a result of this early volatility. We have previously shown that this assumption of reduced sensitivity with a hidden-state inference model can produce reduced exploration in a foraging task [88], a behavior found in ELU populations [123], and may also explain why individuals who experience early life unpredictability are at higher risk of developing substance use disorders and relapsing following treatment [87].

Our model is predicated on the assumption that prediction error learning can serve as a mechanism of environmental adaptation across multiple timescales — within a task and across development. Embodying an extreme form of sensitive period, adulthood is conceptualized as a period in which learning has altogether ceased. Future work could examine the effect of more realistic, relaxed constraints on learning in adulthood – in which developmental experience lays the groundwork for the architecture of neural systems which later adulthood experience can modify and reorganize [66, 102]. Under this scenario, the prior biases instilled by the developmental environment should have their greatest influence in few shot or one shot learning experiences. When current experience underdetermines what an agent should expect or do, past experience should largely influence the conclusion an agent reaches, with early life experience having a particularly privileged role [82]. Such inductive biases facilitate learning in environments aligned with these biases and frustrate it in misaligned environments. If the influence of the developmental environment on expectations and choice is greatest in environments in which the agent has limited experience, this has implications for when symptoms for disorders like anxiety and substance use disorder should worsen [172, 31].

Our results highlight the key role time plays in shaping reinforcement learning and consequently its impact on behaviors associated with mental illness. The model's ability to produce varied phenotypes from the same computations suggests that the model's implications extend beyond anhedonia. Potentially it provides a common origin for a number of psychiatric disorders, offering a potential explanation for high co-morbidity rates [95, 104, 113]. Further research is needed to empirically test the model's behavioral predictions, namely, for early life unpredictability's impact on interval timing, and interval timing's relationship with psychiatric disorders. Finally, our results offer a demonstration of the value of computational modeling to understanding the development of psychopathology. By drawing on a reinforcement learning framework, we can formalize the changing relationship between the agent and their environment across development, produce testable predictions of how the environment shapes the latent computations underlying clinically relevant behaviors, like learning, and propose mechanistic links between altered computations and the later emergence of psychiatric symptoms.

Chapter 5

Conclusion

5.0.1 General Discussion

In this dissertation, we reevaluated three classic behavioral biases: over exploitation, over exploration, and impaired reward learning. By definition, these biases stand at odds with the predictions of standard optimal learning and decision making models. And yet, many humans systematically engage in them. In each chapter, rather than question the validity of human behavior, we questioned the validity of the optimal model's assumptions. Often optimal models make simplifying assumptions, either about the nature of the environment or the agent's knowledge, to enhance tractability. But by doing so, these models incidentally eschew a key feature of the learning and decision making problems humans face in the real world — uncertainty. Our work proposes alternative sets of assumptions aimed at preserving the forms of uncertainty ever-present in natural environments, and demonstrates how classic biases emerge from these assumptions. We find that each bias reflects a deft handling of uncertainty, an ability made possible by learning processes that unfold across multiple timescales.

In Chapters 2 and 3, we considered patch foraging, a form of sequential decision making problem. We investigated instances in which choices deviated from Marginal Value Theorem's (MVT) prescribed behavior – ranging from under to over exploration. Implicitly, MVT assumes the forager possesses perfect knowledge of their environment. However, uncertainty is ubiquitous in real-world environments, those that decision makers are adapted to. To ameliorate this mismatch in assumption, we proposed a rational structure learning and uncertainty-adaptive planning model to augment MVT. Rather than presume perfect knowledge, our model specifies how a decision maker could learn the structure of their environment, ultimately progressing towards the MVT optimal policy and make reasonable decisions even with limited experience. A major assumption underpins our model: people deploy the sophisticated learning and decision making strategies that they've learned work well in complex real world environments even when faced with much simpler task environments.

In Chapter 2, our model provided a superior explanation of participants' overharvesting relative to MVT and alternative models. In Chapter 3, we applied the same model to understanding why exploration is heightened during development. Strikingly, the same structure inference process that captured adults' over exploitation also captured children and adolescents' over exploration. By specifying different structural priors varying in their complexity, the same model was capable of producing divergent behaviors. Model fitting revealed that children and adolescents' exploratory behavior was best explained by a structural prior biased towards simpler environments. Adults, in contrast, explored less, acting as if they had a structural prior biased towards complex environments. Potentially, the age-related differences in structural priors reflect differences in real world experience. Throughout development, we gain increasing experience with the complexity, variation, and nuances of the world. Our model suggests that this has meaningful consequences for how we explore novel environments: we draw on our experiences in past environments to guide how we mitigate uncertainty in the present environment.

In Chapter 4, we aimed to characterize the relationship between early life adversity, in the form of unpredictability, and impaired reward learning. Prior work has identified early life unpredictability as a common antecedent to alterations in reward learning and its associated neural circuitry. Often, these alterations are often cast as "disruptions" or "aberrations," but, here, we suggest that they are the result of rational adaptation to an unpredictable environment. To demonstrate this, we simulated a standard reinforcement learning model undergoing development. We simplified development into two periods — a period of heightened plasticity followed by a period devoid of it. During the plasticity period, agents' representations could flexibly adapt to the structure of the environment. We compared two groups of simulated agents with different early life environments. During the early period of plasticity, one group experienced much more unpredictably timed rewards than the other. When plasticity ceased, both groups were placed in a novel environment with predictable rewards. In the adulthood environment, the unpredictability-exposed group exhibited impaired reward learning which our model attributes to a discrepancy between the developmental and adulthood environments. As in Chapters 2 and 3, our bias of interest emerges from the process of experience in prior environments shaping behavior in the current environment.

In summary, for each behavior, we examined how the decision problems presented by the current task environment differed from those encountered in previous environments. This led us to conclude that all three behaviors occur under greater uncertainty than is assumed by standard optimal models. For that reason, we advanced rational structure learning models as more realistic alternatives. Remarkably, these rational learning models successfully reproduced the observed suboptimal behavior from a minimal set of assumptions. As a result, there is further justification to reevaluate the suboptimality of these behaviors.

Our approach follows in the tradition of Anderson's rational analysis [8]. A major contribution of this dissertation has been to expose the brittleness of the objective functions, or goals, specified by the optimal models. We attribute the brittleness to their underlying assumptions. Often, these assumptions are made for mathematical or conceptual simplicity, and are justified by asserting that they do not change the core problem faced by the agent. However, through recognizing how complex, dynamic, and varied real-world environments are, it becomes evident that through this simplification, the decision problem has been misrepresented.

The ways we act and adapt should mirror the world. Our learning should be robust to uncertainty, our decision making should aim to resolve it, and, our developmental mechanisms should support adaptation to a wide array of environmental inputs.

5.0.2 Open Questions and Future Work

Rational models often begin with the assumption that cognitive mechanisms are *already* well-adapted to the environment. Yet, in the case of decision making, adaptation is an ongoing, lifelong process. Our work has examined how rational learning and adaptation mechanisms unfold and interact across multiple timescales, cooperating to enable rapid and flexible decision making in new situations. Through modeling these interactions, we widen the scope of rational approaches. Now, they can be applied to mechanisms that continually adapt to a range of environments throughout the lifespan. This, in turn, allows us to ask questions such as: why does the human developmental trajectory take the form it does?

In the following discussion, we sketch out how rational analysis can be applied to the study of developmental mechanisms by breaking down this broad question into two more specific ones: how should developmental stages be sequenced, and how long should sensitive periods of plasticity last? Unique to its application to development, rational analysis requires the agent to adapt to *multiple* environments. Even when the external environment remains constant, the internal environment—such as abilities and learning objectives—will inevitably change. We examine the empirical research relevant to this issue, its connection to our two key questions, and its implications for rational modeling.

How should developmental stages be sequenced?

Each development stage is defined by a unique internal environment, characterized by the agent's abilities and learning scope. How do these features determine a stage's ideal placement within the overall developmental trajectory?

As individuals progress through developmental stages, their abilities mature, allowing them to process and reason about increasing amounts of environmental input. Theoretical work in psychology and machine learning suggests that early maturational constraints may confer benefits for learning [58, 183]. Real world environments are highly complex and noisy, posing a challenging learning context even for adults. However, younger children's limited attentional and memory capacities simplify the learning context by reducing the amount of information they need to process. The gradual lifting of constraints across development creates a form of curriculum learning, where increasingly difficult problems build upon simpler ones, which serve as scaffolds [20]. Both theoretical and empirical studies have shown that such scaffolding accelerates learning and increases the generalizability of the acquired knowledge to new contexts [48, 148, 181, 116]. To apply a rational modeling approach, one could simulate various developmental curricula, evaluating their learning efficiency and robustness, and compare the "winning" curricula to the structure of human developmental trajectories.

As learning abilities improve, the scope or expansiveness of learning decreases. From infancy into young adulthood, we shift from constructing broad causal models of the world [187, 125] to refining models of specific, novel environments [47, 73, 79]. These changes in learning scope may explain the developmental differences in exploration observed in Chapter 3. Our findings indicate that children and adolescents broadly sample their environments, while adults deeply focus on specific regions, often overexploiting them, possibly to reduce uncertainty. Adults' strategies may reflect deep exploration, in which options are prioritized if they offer immediate and future information gains, promoting temporally-extended exploration [145]. Integrating formalizations of exploration with curriculum learning approaches would allow for developmental change to be defined in terms of learning ability and scope, overall enriching the modeling of developmental change. Scope adds a critical, yet underexplored, constraint on the sequencing of stages in curriculum learning. Often, an agent's learning scope or objective is assumed to remain constant throughout the learning period. This has critical implications for a rational analysis of development. Allowing the scope to vary along with abilities could alter the optimal developmental trajectory.

How long should sensitive periods last?

Across development, individuals vary in their sensitivity to the environment due to systematic changes in plasticity. The timing of these sensitive periods is not fixed; their onset and closure are modulated by environmental cues such as reward deprivation and unpredictability [33, 69, 210, 68]. In Chapter 4, we opted not to model the relationship between environmental unpredictability and sensitive period length for mathematical simplicity. However, future research could formalize this relationship using bounded optimality approaches, which treat computational limitations as factors equally crucial in shaping the mind as the environment is [177, 120, 72, 118]. In this application, plasticity is the primary constraint of interest. Plasticity allows for flexibility and adaptation to the environment but comes at the expense of efficiency. This reveals another problem a developmental mechanism must optimize for — determining the optimal duration of plasticity given the trade off between flexibility and efficiency and the structure of the environment.

Finally, because agents must adapt to multiple environments throughout their lifespan, there is no guarantee that the childhood and adulthood environments will be similar, let alone the same. How should developmental mechanisms accommodate this uncertainty [59]? What forms of plasticity should be retained into adulthood? To address these developmental questions through rational analysis, one could draw on meta-learning approaches [142, 65] to model how agents adapt to a range of environments, balancing flexibility and efficiency in the face of potential uncertainty. In meta-learning frameworks developed in machine learning, alter the learning objective. Rather than optimize for performance in a single environment, performance is optimized across many environments [93, 200]. Changing the learning objective in this way enhances performance even in environments dissimilar to the ones agents were originally trained in [97]. Importantly, meta-learning models specify a wide range of meta-learning objectives [36, 19]. Future work could evaluate how well different classes of meta-learning objectives capture the empirically observed results of a match or mismatch between the childhood and adulthood environment.

5.0.3 Summary

The results presented in this dissertation suggest that studying how multi-scale learning and decision making processes unfold and interact opens up novel opportunities to reexamine the rationality of certain behaviors. Additionally, it extends the application of rational frameworks to temporally extended processes, including development. Rational frameworks excel at unifying and organizing seemingly disjoint bodies of work, offering parsimonious accounts. Leveraging these tools will enable us to build on an already extensive literature on cognitive development and offer principles that explain how the mind is shaped by the numerous environments it encounters across development.

Bibliography

[1] L. Acerbi and W. Ji. Practical bayesian optimization for model fitting with bayesian adaptive direct search. *Adv. Neural Inf. Process. Syst.*, pages 1836–1846, May 2017.

[2] D. E. Acuña and P. Schrater. Structure learning in human sequential decision-making. *PLoS Comput. Biol.*, 6(12):e1001003, Dec. 2010.

[3] M. A. Addicott, J. M. Pearson, M. M. Sweitzer, D. L. Barack, and M. L. Platt. A primer on foraging and the Explore/Exploit Trade-Off for psychiatry research. *Neuropsychopharmacology*, 42(10):1931–1939, Sept. 2017.

[4] D. J. Aldous. Exchangeability and related topics. In D. J. Aldous, I. A. Ibragimov, and J. Jacod, editors, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198. Springer Berlin Heidelberg, 1985.

[5] R. Amit, R. Meir, and K. Ciosek. Discount factor as a regularizer in reinforcement learning. *CoRR*, abs/2007.02040, 2020.

[6] M. Amlung, E. Marsden, K. Holshausen, V. Morris, H. Patel, L. Vedelago, K. R. Naish, D. D. Reed, and R. E. McCabe. Delay discounting as a transdiagnostic process in psychiatric disorders: A meta-analysis. *JAMA Psychiatry*, 76(11):1176–1186, Nov. 2019.

[7] M. Amlung, L. Vedelago, J. Acker, I. Balodis, and J. MacKillop. Steep delay discounting and addictive behavior: a meta-analysis of continuous associations. *Addiction*, 112(1):51–62, Jan. 2017.

[8] J. R. Anderson. The adaptive nature of human categorization. *Psychol. Rev.*, 98(3):409–429, 1991.

[9] C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Ann. Stat.*, 2(6):1152–1174, 1974.

[10] J. Aylward, V. Valton, W.-Y. Ahn, R. L. Bond, P. Dayan, J. P. Roiser, and O. J. Robinson. Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nat Hum Behav*, 3(10):1116–1123, Oct. 2019.

[11] I. C. Ballard and S. M. McClure. Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *J. Neurosci. Methods*, 317:37–44, Apr. 2019.

[12] P. D. Balsam, M. R. Drew, and C. R. Gallistel. Time and associative learning. *Comp. Cogn. Behav. Rev.*, 5:1–22, 2010.

[13] T. Z. Baram, E. P. Davis, A. Obenaus, C. A. Sandman, S. L. Small, A. Solodkin, and H. Stern. Fragmentation and unpredictability of early-life experience in mental disorders. *Am. J. Psychiatry*, 169(9):907–915, Sept. 2012.

[14] W. Barfuss, J. F. Donges, V. V. Vasconcelos, J. Kurths, and S. A. Levin. Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. *Proceedings of the National Academy of Sciences*, 117(23):12915–12922, 2020.

[15] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.*, 68(3), Apr. 2013.

[16] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen. Parsimonious mixed models. June 2015.

[17] T. E. J. Behrens, M. W. Woolrich, M. E. Walton, and M. F. S. Rushworth. Learning the value of information in an uncertain world. *Nat. Neurosci.*, 10(9):1214–1221, Sept. 2007.

[18] J. Belsky and R. M. P. Fearon. Early attachment security, subsequent maternal sensitivity, and later child development: does continuity in development depend upon continuity of caregiving? *Attach. Hum. Dev.*, 4(3):361–387, Dec. 2002.

[19] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. A Meta-Transfer objective for learning to disentangle causal mechanisms. Jan. 2019.

[20] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, June 2009. Association for Computing Machinery.

[21] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. `https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf`, 2012. Accessed: 2021-5-6.

[22] R. M. Birn, B. J. Roeber, and S. D. Pollak. Early childhood stress exposure, reward pathways, and adult decision making. *Proc. Natl. Acad. Sci. U. S. A.*, 114(51):13549–13554, Dec. 2017.

[23] M. T. Birnie, C. L. Kooiker, A. K. Short, J. L. Bolton, Y. Chen, and T. Z. Baram. Plasticity of the reward circuitry after Early-Life adversity: Mechanisms and significance. *Biol. Psychiatry*, 87(10):875–884, May 2020.

[24] S. J. Bishop and C. Gagne. Anxiety, depression, and decision making: A computational perspective. *Annu. Rev. Neurosci.*, 41:371–388, July 2018.

[25] T. C. Blanchard and B. Y. Hayden. Monkeys are more patient in a foraging task than in a standard intertemporal choice task. *PLoS One*, 10(2):e0117057, Feb. 2015.

[26] N. J. Blanco and V. M. Sloutsky. Exploration, exploitation, and development: Developmental shifts in decision-making. *Child Dev.*, Feb. 2024.

[27] R. Boecker, N. E. Holz, A. F. Buchmann, D. Blomeyer, M. M. Plichta, I. Wolf, S. Baumeister, A. Meyer-Lindenberg, T. Banaschewski, D. Brandeis, and M. Laucht. Impact of early life adversity on reward processing in young adults: EEG-fMRI results from a prospective study over 25 years. *PLoS One*, 9(8):e104185, Aug. 2014.

[28] J. L. Bolton, J. Molet, L. Regev, Y. Chen, N. Rismanchi, E. Haddad, D. Z. Yang, A. Obenaus, and T. Z. Baram. Anhedonia following Early-Life adversity involves aberrant interaction of reward and anxiety circuits and is reversed by partial silencing of amygdala Corticotropin-Releasing hormone gene. *Biol. Psychiatry*, 83(2):137–147, Jan. 2018.

[29] A. M. Bornstein, M. Aly, S. F. Feng, N. B. Turk-Browne, K. A. Norman, and J. D. Cohen. Associative memory retrieval modulates upcoming perceptual decisions. Sept. 2023.

[30] A. M. Bornstein and N. D. Daw. Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. *PLoS computational biology*, 9(12):e1003387, 2013.

[31] A. M. Bornstein and H. Pickard. "chasing the first high": memory sampling in drug choice. *Neuropsychopharmacology*, 45(6):907–915, May 2020.

[32] J. R. Busemeyer. Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *J. Exp. Psychol. Learn. Mem. Cogn.*, 11(3):538–564, July 1985.

[33] B. L. Callaghan and N. Tottenham. The stress acceleration hypothesis: Effects of early-life adversity on emotion circuits and behavior. *Curr Opin Behav Sci*, 7:76–81, Feb. 2016.

[34] E. C. Carter and A. D. Redish. Rats value time differently on equivalent foraging and delay-discounting tasks. *J. Exp. Psychol. Gen.*, 145(9):1093–1101, Sept. 2016.

[35] E. L. Charnov. Optimal foraging, the marginal value theorem. *Theor. Popul. Biol.*, 9(2):129–136, Apr. 1976.

[36] Y. Chen, Y. Zhang, Y. Bian, H. Yang, K. Ma, B. Xie, T. Liu, B. Han, and J. Cheng. Learning causally invariant representations for Out-of-Distribution generalization on graphs. Feb. 2022.

[37] A. O. Cohen, K. Nussenbaum, H. M. Dorfman, S. J. Gershman, and C. A. Hartley. The rational use of causal inference to guide reinforcement learning strengthens with age. *NPJ Sci Learn*, 5:16, Oct. 2020.

[38] J. Cohen. Statistical power analysis. *Current directions in psychological science*, 1(3):98–101, 1992.

[39] A. G. E. Collins and M. J. Frank. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.*, 120(1):190–229, Jan. 2013.

[40] S. M. Constantino and N. D. Daw. Learning the opportunity cost of time in a patch-foraging task. *Cogn. Affect. Behav. Neurosci.*, 15(4):837–853, Dec. 2015.

[41] P. H. Crowley, D. R. DeVries, and A. Sih. Inadvertent errors and error-constrained optimization: fallible foraging by bluegill sunfish. *Behav. Ecol. Sociobiol.*, 27(2):135–144, Aug. 1990.

[42] F. C. Cruz, I. M. Quadros, C. d. S. Planeta, and K. A. Miczek. Maternal separation stress in male mice: long-term increases in alcohol intake. *Psychopharmacology*, 201(3):459–468, Dec. 2008.

[43] A. J. Culbreth, A. Westbrook, N. D. Daw, M. Botvinick, and D. M. Barch. Reduced model-based decision-making in schizophrenia. *J. Abnorm. Psychol.*, 125(6):777–787, Aug. 2016.

[44] I. C. Cuthil, P. Haccou, and A. Kacelnik. Starlings (sturnus vulgaris) exploiting patches: response to long-term changes in travel time. *Behav. Ecol.*, 5(1):81–90, Mar. 1994.

[45] I. C. Cuthill, A. Kacelnik, J. R. Krebs, P. Haccou, and Y. Iwasa. Starlings exploiting patches: the effect of recent experience on foraging decisions. *Anim. Behav.*, 40(4):625–640, Oct. 1990.

[46] C. V. Day, J. M. Gatt, A. Etkin, C. DeBattista, A. F. Schatzberg, and L. M. Williams. Cognitive and emotional biomarkers of melancholic depression: An iSPOT-D report. *J. Affect. Disord.*, 176:141–150, May 2015.

[47] J. H. Decker, A. R. Otto, N. D. Daw, and C. A. Hartley. From creatures of habit to Goal-Directed learners: Tracking the developmental emergence of Model-Based reinforcement learning. *Psychol. Sci.*, 27(6):848–858, June 2016.

[48] R. B. Dekker, F. Otto, and C. Summerfield. Curriculum learning for human compositional generalization. *Proc. Natl. Acad. Sci. U. S. A.*, 119(41):e2205582119, Oct. 2022.

[49] M. J. Dennison, M. L. Rosen, K. A. Sambrook, J. L. Jenness, M. A. Sheridan, and K. A. McLaughlin. Differential associations of distinct forms of childhood adversity with neurobehavioral measures of reward processing: A developmental pathway to depression. *Child Dev.*, 90(1):e96–e113, Jan. 2019.

[50] D. G. Dillon, A. J. Holmes, J. L. Birk, N. Brooks, K. Lyons-Ruth, and D. A. Pizzagalli. Childhood adversity is associated with left basal ganglia dysfunction during reward anticipation in adulthood. *Biol. Psychiatry*, 66(3):206–213, Aug. 2009.

[51] D. G. Dillon and D. A. Pizzagalli. Mechanisms of memory disruption in depression. *Trends Neurosci.*, 41(3):137–149, Mar. 2018.

[52] H. M. Dorfman, R. Bhui, B. L. Hughes, and S. J. Gershman. Causal inference about good and bad outcomes. *Psychol. Sci.*, 30(4):516–525, Apr. 2019.

[53] J. Drugowitsch. Fast and accurate monte carlo sampling of first-passage times from wiener diffusion models. *Sci. Rep.*, 6:20490, Feb. 2016.

[54] S. Dubal, A. Pierson, and R. Jouvent. Focused attention in anhedonia: a P3 study. *Psychophysiology*, 37(5):711–714, Sept. 2000.

[55] M. Dubois, A. Bowler, M. E. Moses-Payne, J. Habicht, R. Moran, N. Steinbeis, and T. U. Hauser. Exploration heuristics decrease during youth. *Cogn. Affect. Behav. Neurosci.*, 22(5):969–983, Oct. 2022.

[56] A. Efrati and Y. Gutfreund. Early life exposure to noise alters the representation of auditory localization cues in the auditory space map of the barn owl. *J. Neurophysiol.*, 105(5):2522–2535, May 2011.

[57] E. Eldar, R. B. Rutledge, R. J. Dolan, and Y. Niv. Mood as representation of momentum. *Trends Cogn. Sci.*, 20(1):15–24, Jan. 2016.

[58] J. L. Elman. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99, July 1993.

[59] T. W. Fawcett and W. E. Frankenhuis. Adaptive explanations for sensitive windows in development. *Front. Zool.*, 12 Suppl 1(Suppl 1):S3, Aug. 2015.

[60] P. Fearnhead. Particle filters for mixture models with an unknown number of components. *Stat. Comput.*, 14(1):11–21, Jan. 2004.

[61] V. J. Felitti. The relation between adverse childhood experiences and adult health: Turning gold into lead. *Perm. J.*, 6(1):44–47, 2002.

[62] L. Fontanesi, S. Gluth, M. S. Spektor, and J. Rieskamp. A reinforcement learning diffusion decision model for value-based decisions. *Psychon. Bull. Rev.*, 26(4):1099–1121, Aug. 2019.

[63] V. Francois-Lavet, G. Rabusseau, J. Pineau, D. Ernst, and R. Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *J. Artif. Intell. Res.*, 65:1–30, May 2019.

[64] M. J. Frank. *Dynamic dopamine modulation of striato-cortical circuits in cognition: Converging neuropsychological, psychopharmacological and computational studies.* PhD thesis, University of Colorado at Boulder, Ann Arbor, United States, 2004.

[65] W. E. Frankenhuis and A. Gopnik. Early adversity and the development of explore-exploit tradeoffs. *Trends Cogn. Sci.*, May 2023.

[66] A. Galván. Neural plasticity of development and learning. *Hum. Brain Mapp.*, 31(6):879–890, June 2010.

[67] N. Garrett and N. D. Daw. Biased belief updating and suboptimal choice in foraging decisions. *Nat. Commun.*, 11(1):3417, July 2020.

[68] D. G. Gee and E. M. Cohodes. Caregiving influences on development: A sensitive period for biological embedding of predictability and safety cues. *Curr. Dir. Psychol. Sci.*, 30(5):376–383, Oct. 2021.

[69] D. G. Gee, L. J. Gabard-Durnam, J. Flannery, B. Goff, K. L. Humphreys, E. H. Telzer, T. A. Hare, S. Y. Bookheimer, and N. Tottenham. Early developmental emergence of human amygdala-prefrontal connectivity after maternal deprivation. *Proc. Natl. Acad. Sci. U. S. A.*, 110(39):15638–15643, Sept. 2013.

[70] S. J. Gershman and R. Bhui. Rationally inattentive intertemporal choice. *Nat. Commun.*, 11(1):3365, July 2020.

[71] S. J. Gershman, D. M. Blei, and Y. Niv. Context, learning, and extinction. *Psychol. Rev.*, 117(1):197–209, Jan. 2010.

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

[73] S. J. Gershman, H. T. Pouncy, and H. Gweon. Learning the structure of social influence. *Cogn. Sci.*, 41 Suppl 3:545–575, Apr. 2017.

[74] C. M. Gillan, M. Kosinski, R. Whelan, E. A. Phelps, and N. D. Daw. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife*, 5, Mar. 2016.

[75] A. P. Giron, S. Ciranka, E. Schulz, W. van den Bos, A. Ruggeri, B. Meder, and C. M. Wu. Developmental changes in exploration resemble stochastic optimization. *Nat Hum Behav*, Aug. 2023.

[76] P. W. Glimcher, M. C. Dorris, and H. M. Bayer. Physiological utility theory and the neuroeconomics of choice. *Games Econ. Behav.*, 52(2):213–256, Aug. 2005.

[77] L. M. Glynn, H. S. Stern, M. A. Howland, V. B. Risbrough, D. G. Baker, C. M. Nievergelt, T. Z. Baram, and E. P. Davis. Measuring novel antecedents of mental illness: the questionnaire of unpredictability in childhood. *Neuropsychopharmacology*, 44(5):876–882, Apr. 2019.

[78] B. Goff, D. G. Gee, E. H. Telzer, K. L. Humphreys, L. Gabard-Durnam, J. Flannery, and N. Tottenham. Reduced nucleus accumbens reactivity and adolescent depression following early-life stress. *Neuroscience*, 249:129–138, Sept. 2013.

[79] N. Goldway, E. Eldar, G. Shoval, and C. A. Hartley. Computational mechanisms of addiction and anxiety: a developmental perspective. *Biol. Psychiatry*, Feb. 2023.

[80] J. K. Gollan, H. T. Pane, M. S. McCloskey, and E. F. Coccaro. Identifying differences in biased affective information processing in major depression. *Psychiatry Res.*, 159(1-2):18–24, May 2008.

[81] S. J. Granger, L. M. Glynn, C. A. Sandman, S. L. Small, A. Obenaus, D. B. Keator, T. Z. Baram, H. Stern, M. A. Yassa, and E. P. Davis. Aberrant maturation of the uncinate fasciculus follows exposure to unpredictable patterns of maternal signals. *J. Neurosci.*, 41(6):1242–1250, Feb. 2021.

[82] T. L. Griffiths, N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.*, 14(8):357–364, Aug. 2010.

[83] T. L. Griffiths, D. J. Navarro, and A. N. Sanborn. A more rational model of categorization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 28(28), 2006.

[84] A. A. Hane, H. A. Henderson, B. C. Reeb-Sutherland, and N. A. Fox. Ordinary variations in human maternal caregiving in infancy and biobehavioral development in early childhood: A follow-up study. *Dev. Psychobiol.*, 52(6):558–567, Sept. 2010.

[85] J. L. Hanson, D. Albert, A.-M. R. Iselin, J. M. Carré, K. A. Dodge, and A. R. Hariri. Cumulative stress in childhood is associated with blunted reward-related brain activity in adulthood. *Soc. Cogn. Affect. Neurosci.*, 11(3):405–412, Mar. 2016.

[86] J. L. Hanson, A. V. Williams, D. A. Bangasser, and C. J. Peña. Impact of early life stress on reward circuit function and regulation. *Front. Psychiatry*, 12:1799, 2021.

[87] N. Harhen, T. Z. Baram, M. A. Yassa, and A. M. Bornstein. Formalizing the relationship between early life adversity and addiction vulnerability: The role of memory sampling. *Biol. Psychiatry*, 89(9):S189, May 2021.

[88] N. C. Harhen and A. M. Bornstein. Overharvesting in human patch foraging reflects rational structure learning and adaptive planning. *Proc. Natl. Acad. Sci. U. S. A.*, 120(13):e2216524120, Mar. 2023.

[89] M. B. Harms, Y. Xu, C. S. Green, K. Woodard, R. Wilson, and S. D. Pollak. The structure and development of explore-exploit decision making. *Cogn. Psychol.*, 150:101650, Mar. 2024.

[90] B. Y. Hayden. Time discounting and time preference in animals: a critical review. *Psychonomic bulletin & review*, 23(1):39–53, 2016.

[91] B. Y. Hayden, J. M. Pearson, and M. L. Platt. Neuronal basis of sequential foraging decisions in a patchy environment. *Nat. Neurosci.*, 14(7):933–939, June 2011.

[92] E. A. Heerey, B. M. Robinson, R. P. McMahon, and J. M. Gold. Delay discounting in schizophrenia. *Cogn. Neuropsychiatry*, 12(3):213–221, May 2007.

[93] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-Learning in neural networks: A survey. Apr. 2020.

[94] M. W. Howard, C. J. MacDonald, Z. Tiganj, K. H. Shankar, Q. Du, M. E. Hasselmo, and H. Eichenbaum. A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *J. Neurosci.*, 34(13):4692–4707, Mar. 2014.

[95] F. Jacobi, H.-U. Wittchen, C. Holting, M. Höfler, H. Pfister, N. Müller, and R. Lieb. Prevalence, co-morbidity and correlates of mental disorders in the general population: results from the german health interview and examination survey (GHS). *Psychol. Med.*, 34(4):597–611, May 2004.

[96] N. Jiang, A. Kulesza, S. Singh, and R. Lewis. The dependence of effective planning horizon on model accuracy. `https://nanjiang.cs.illinois.edu/files/gamma-AAMAS-final.pdf`. Accessed: 2022-2-18.

[97] P. Jiang, K. Xin, Z. Wang, and C. Li. Invariant meta learning for Out-of-Distribution generalization. Jan. 2023.

[98] D. Z. Jin, N. Fujii, and A. M. Graybiel. Neural representation of time in cortico-basal ganglia circuits. *Proc. Natl. Acad. Sci. U. S. A.*, 106(45):19156–19161, Nov. 2009.

[99] A. Kacelnik and I. A. Todd. Psychological mechanisms and the marginal value theorem: effect of variability in travel time on patch exploitation. *Anim. Behav.*, 43(2):313–322, Feb. 1992.

[100] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[101] G. A. Kane, A. M. Bornstein, A. Shenhav, R. C. Wilson, N. D. Daw, and J. D. Cohen. Rats exhibit similar biases in foraging and intertemporal choice tasks. *Elife*, 8, Sept. 2019.

[102] B. A. Karmiloff-Smith. Beyond modularity: A developmental perspective on cognitive science. *Eur. J. Disord. Commun.*, 29(1):95–105, Jan. 1994.

[103] T. Keasar, E. Rashkovich, D. Cohen, and A. Shmida. Bees in two-armed bandit situations: foraging choices and possible decision mechanisms. *Behav. Ecol.*, 13(6):757–765, Nov. 2002.

[104] R. C. Kessler, W. T. Chiu, O. Demler, K. R. Merikangas, and E. E. Walters. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. *Arch. Gen. Psychiatry*, 62(6):617–627, June 2005.

[105] Z. P. Kilpatrick, J. D. Davidson, and A. El Hady. Uncertainty drives deviations in normative foraging decision strategies. Apr. 2021.

[106] G. Kim, J. A. Lewis-Peacock, K. A. Norman, and N. B. Turk-Browne. Pruning of memories by context-based prediction error. *Proc. Natl. Acad. Sci. U. S. A.*, 111(24):8997–9002, June 2014.

[107] N. Kolling and T. Akam. (reinforcement?) learning to forage optimally. *Curr. Opin. Neurobiol.*, 46:162–169, Oct. 2017.

[108] N. Kolling, T. E. J. Behrens, R. B. Mars, and M. F. S. Rushworth. Neural mechanisms of foraging. *Science*, 336(6077):95–98, Apr. 2012.

[109] W. Kool, S. J. Gershman, and F. A. Cushman. Cost-Benefit arbitration between multiple Reinforcement-Learning systems. *Psychol. Sci.*, 28(9):1321–1333, Sept. 2017.

[110] T. A. Kosten, M. J. Miserendino, and P. Kehoe. Enhanced acquisition of cocaine self-administration in adult rats with neonatal isolation stress experience. *Brain Res.*, 875(1-2):44–50, Sept. 2000.

[111] T. A. Kosten, H. Sanchez, X. Y. Zhang, and P. Kehoe. Neonatal isolation enhances acquisition of cocaine self-administration and food responding in female rats. *Behav. Brain Res.*, 151(1-2):137–149, May 2004.

[112] T. A. Kosten, X. Y. Zhang, and P. Kehoe. Heightened cocaine and food self-administration in female rats with neonatal isolation experience. *Neuropsychopharmacology*, 31(1):70–76, Jan. 2006.

[113] R. F. Krueger, Y. E. Chentsova-Dutton, K. E. Markon, D. Goldberg, and J. Ormel. A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. *J. Abnorm. Psychol.*, 112(3):437–447, Aug. 2003.

[114] H. K. Lambert, K. M. King, K. C. Monahan, and K. A. McLaughlin. Differential associations of threat and deprivation with emotion regulation and cognitive control in adolescence. *Dev. Psychopathol.*, 29(3):929–940, Aug. 2017.

[115] C. Le Heron, N. Kolling, O. Plant, A. Kienast, R. Janska, Y.-S. Ang, S. Fallon, M. Husain, and M. A. J. Apps. Dopamine modulates dynamic Decision-Making during foraging. *J. Neurosci.*, 40(27):5273–5282, July 2020.

[116] J. H. Lee, S. S. Mannelli, and A. Saxe. Why do animals need shaping? a theory of task composition and curriculum learning. Feb. 2024.

[117] J. K. Lenow, S. M. Constantino, N. D. Daw, and E. A. Phelps. Chronic and acute stress promote overexploitation in serial decision making. *J. Neurosci.*, 37(23):5681–5689, June 2017.

[118] R. L. Lewis, A. Howes, and S. Singh. Computational rationality: linking mechanism and behavior through bounded utility maximization. *Top. Cogn. Sci.*, 6(2):279–311, Apr. 2014.

[119] Y. Li, D. Fitzpatrick, and L. E. White. The development of direction selectivity in ferret visual cortex requires early visual experience. *Nat. Neurosci.*, 9(5):676–681, May 2006.

[120] F. Lieder and T. L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.*, 43:e1, Feb. 2019.

[121] E. G. Liquin and A. Gopnik. Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, 218:104940, Jan. 2022.

[122] A. Lloyd, R. McKay, C. L. Sebastian, and J. H. Balsters. Are adolescents more optimal decision-makers in novel environments? examining the benefits of heightened exploration in a patch foraging paradigm. *Dev. Sci.*, 24(4):e13075, July 2021.

[123] A. Lloyd, R. T. McKay, and N. Furl. Individuals with adverse childhood experiences explore less and underweight reward feedback. *Proc. Natl. Acad. Sci. U. S. A.*, 119(4), Jan. 2022.

[124] E. A. Ludvig, R. S. Sutton, and E. J. Kehoe. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.*, 20(12):3034–3054, Dec. 2008.

[125] Y. Luo, L. Kaufman, and R. Baillargeon. Young infants' reasoning about physical events involving inert and self-propelled objects. *Cogn. Psychol.*, 58(4):441–486, June 2009.

[126] L. Machlin, A. B. Miller, J. Snyder, K. A. McLaughlin, and M. A. Sheridan. Differential associations of deprivation and threat with cognitive control and fear conditioning in early childhood. *Front. Behav. Neurosci.*, 13:80, May 2019.

[127] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, July 2010.

[128] K. A. McLaughlin, M. A. Sheridan, K. L. Humphreys, J. Belsky, and B. J. Ellis. The value of dimensional models of early experience: Thinking clearly about concepts and categories. *Perspect. Psychol. Sci.*, 16(6):1463–1472, Nov. 2021.

[129] B. Meder, C. M. Wu, E. Schulz, and A. Ruggeri. Development of directed and random exploration in children. *Dev. Sci.*, 24(4):e13095, July 2021.

[130] M. A. Mehta, E. Gore-Langton, N. Golembo, E. Colvert, S. C. R. Williams, and E. Sonuga-Barke. Hyporesponsive reward anticipation in the basal ganglia following severe institutional deprivation early in life. *J. Cogn. Neurosci.*, 22(10):2316–2325, Oct. 2010.

[131] A. B. Miller, L. Machlin, K. A. McLaughlin, and M. A. Sheridan. Deprivation and psychopathology in the fragile families study: A 15-year longitudinal investigation. *J. Child Psychol. Psychiatry*, 62(4):382–391, Apr. 2021.

[132] D. Mobbs, P. C. Trimmer, D. T. Blumstein, and P. Dayan. Foraging for foundations in decision neuroscience: insights from ethology. *Nat. Rev. Neurosci.*, 19(7):419–427, July 2018.

[133] J. Molet, P. M. Maras, E. Kinney-Lang, N. G. Harris, F. Rashid, A. S. Ivy, A. Solodkin, A. Obenaus, and T. Z. Baram. MRI uncovers disrupted hippocampal microstructure that underlies memory impairments after early-life adversity. *Hippocampus*, 26(12):1618–1632, Dec. 2016.

[134] J. Morimoto. Foraging decisions as multi-armed bandit problems: Applying reinforcement learning algorithms to foraging data. *J. Theor. Biol.*, 467:48–56, Apr. 2019.

[135] Y. Munakata, H. R. Snyder, and C. H. Chatham. Developing cognitive control: Three key transitions. *Curr. Dir. Psychol. Sci.*, 21(2):71–77, Apr. 2012.

[136] M. R. Nassar, R. C. Wilson, B. Heasly, and J. I. Gold. An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.*, 30(37):12366–12378, Sept. 2010.

[137] NICHD Early Care Research Network. Infant-mother attachment classification: risk and protection in relation to changing maternal caregiving quality. *Dev. Psychol.*, 42(1):38–58, Jan. 2006.

[138] J. Nicholas, N. D. Daw, and D. Shohamy. Uncertainty alters the balance between incremental learning and episodic memory. *Elife*, 11, Dec. 2022.

[139] Y. Niv, D. Joel, I. Meilijson, and E. Ruppin. Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adapt. Behav.*, 10(1):5–24, Jan. 2002.

[140] P. Nonacs. State dependent behavior and the marginal value theorem. *Behav. Ecol.*, 12(1):71–83, Jan. 2001.

[141] E. H. Norton, S. M. Fleming, N. D. Daw, and M. S. Landy. Suboptimal criterion learning in static and dynamic environments. *PLoS Comput. Biol.*, 13(1):e1005304, Jan. 2017.

[142] K. Nussenbaum and C. A. Hartley. Understanding the development of reward learning through the lens of meta-learning. *Nature Reviews Psychology*, pages 1–15, Apr. 2024.

[143] K. Nussenbaum, R. E. Martin, S. Maulhardt, Y. J. Yang, G. Bizzell-Hatcher, N. S. Bhatt, M. Koenig, G. M. Rosenbaum, J. P. O'Doherty, J. Cockburn, and C. A. Hartley. Novelty and uncertainty differentially drive exploration across development. *Elife*, 12, Aug. 2023.

[144] K. Nussenbaum, M. Scheuplein, C. V. Phaneuf, M. D. Evans, and C. A. Hartley. Moving developmental research online: Comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra Psychol.*, 6(1), Nov. 2020.

[145] I. Osband, B. Van Roy, D. J. Russo, and Z. Wen. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

[146] T. A. Paine, S. Brainard, E. Keppler, R. Poyle, E. Sai-Hardebeck, V. Schwob, and C. Tannous-Taylor. Juvenile stress increases cocaine-induced impulsivity in female rats. *Behav. Brain Res.*, 414:113488, Sept. 2021.

[147] S. Palminteri, E. J. Kilford, G. Coricelli, and S.-J. Blakemore. The computational development of reinforcement learning during adolescence. *PLoS Comput. Biol.*, 12(6):e1004953, June 2016.

[148] P. I. Pavlov. Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Ann Neurosci*, 17(3):136–141, July 2010.

[149] M. Pelz and C. Kidd. The elaboration of exploratory play. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 375(1803):20190503, July 2020.

[150] M. Petrik and B. Scherrer. Biasing approximate dynamic programming with a lower discount factor. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[151] A. C. Pike and O. J. Robinson. Reinforcement learning in patients with mood and anxiety disorders vs control individuals: A systematic review and meta-analysis. *JAMA Psychiatry*, 79(4):313–322, Apr. 2022.

[152] R. A. Poldrack, J. Clark, E. a. Paré-Blagoev, D. Shohamy, J. Creso Moyano, C. Myers, and M. A. Gluck. Interactive memory systems in the human brain. *Nature*, 414(6863):546–550, 2001.

[153] T. C. S. Potter, N. V. Bryce, and C. A. Hartley. Cognitive components underpinning the development of model-based learning. *Dev. Cogn. Neurosci.*, 25:272–280, June 2017.

[154] E. Pulcu and M. Browning. The misestimation of uncertainty in affective disorders. *Trends Cogn. Sci.*, 23(10):865–875, Oct. 2019.

[155] E. Pulcu, P. D. Trotter, E. J. Thomas, M. McFarquhar, G. Juhasz, B. J. Sahakian, J. F. W. Deakin, R. Zahn, I. M. Anderson, and R. Elliott. Temporal discounting in major depressive disorder. *Psychol. Med.*, 44(9):1825–1834, July 2014.

[156] A. Radulescu and Y. Niv. State representation in mental illness. *Curr. Opin. Neurobiol.*, 55:160–166, Apr. 2019.

[157] R. Ratcliff. A theory of memory retrieval. *Psychol. Rev.*, 85(2):59–108, Mar. 1978.

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

[159] M. Renardy, L. R. Joslyn, J. A. Millar, and D. E. Kirschner. To sobol or not to sobol? the effects of sampling schemes in systems biology applications. *Math. Biosci.*, 337:108593, July 2021.

[160] M. L. Rosen, M. P. Hagen, L. A. Lurie, Z. E. Miles, M. A. Sheridan, A. N. Meltzoff, and K. A. McLaughlin. Cognitive stimulation as a mechanism linking socioeconomic status with executive function: A longitudinal investigation. *Child Dev.*, 91(4):e762–e779, July 2020.

[161] N. Rouhani and Y. Niv. Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, 236(8):2425–2435, Aug. 2019.

[162] N. Rouhani, K. A. Norman, Y. Niv, and A. M. Bornstein. Reward prediction errors create event boundaries in memory. *Cognition*, 203:104269, Oct. 2020.

[163] J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, Dec. 1996.

[164] J. R. Saffran and N. Z. Kirkham. Infant statistical learning. *Annu. Rev. Psychol.*, 69:181–203, Jan. 2018.

[165] A. N. Sanborn, T. L. Griffiths, and D. J. Navarro. Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.*, 117(4):1144–1167, Oct. 2010.

[166] M. L. Schlichting, K. F. Guarino, A. C. Schapiro, N. B. Turk-Browne, and A. R. Preston. Hippocampal structure predicts statistical learning and associative inference abilities during development. *J. Cogn. Neurosci.*, 29(1):37–51, Jan. 2017.

[167] M. V. Schmidt. Animal models for depression and the mismatch hypothesis of disease. *Psychoneuroendocrinology*, 36(3):330–338, Apr. 2011.

[168] E. Schulz, C. M. Wu, A. Ruggeri, and B. Meder. Searching for rewards like a child means less generalization and more directed exploration. *Psychol. Sci.*, 30(11):1561–1572, Nov. 2019.

[169] R. Schurr, D. Reznik, H. Hillman, R. Bhui, and S. J. Gershman. Dynamic computational phenotyping of human cognition. June 2023.

[170] T. X. F. Seow, E. Benoit, C. Dempsey, M. Jennings, A. Maxwell, R. O'Connell, and C. M. Gillan. Model-Based planning deficits in compulsivity are linked to faulty neural representations of task structure. *J. Neurosci.*, 41(30):6539–6550, July 2021.

[171] T. Sharot. The optimism bias. *Curr. Biol.*, 21(23):R941–5, Dec. 2011.

[172] P. B. Sharp, G. A. Miller, R. J. Dolan, and E. Eldar. Towards formal models of psychopathological traits that explain symptom trajectories. *BMC Med.*, 18(1):264, Sept. 2020.

[173] A. Shenhav, M. A. Straccia, J. D. Cohen, and M. M. Botvinick. Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nat. Neurosci.*, 17(9):1249–1254, Sept. 2014.

[174] R. N. Shepard. Mind sights: Original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art. 228, 1990.

[175] M. A. Sheridan, M. Peverill, A. S. Finn, and K. A. McLaughlin. Dimensions of childhood adversity have distinct associations with neural systems underlying executive functioning. *Dev. Psychopathol.*, 29(5):1777–1794, Dec. 2017.

[176] Y. S. Shin and S. DuBrow. Structuring memory through Inference-Based event segmentation. *Top. Cogn. Sci.*, 13(1):106–127, Jan. 2021.

[177] H. A. Simon. Rational choice and the structure of the environment. *Psychol. Rev.*, 63(2):129–138, Mar. 1956.

[178] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, 24:1193–1216, 2001.

[179] C. R. Sims, H. Neth, R. A. Jacobs, and W. D. Gray. Melioration as rational choice: sequential decision making in uncertain environments. *Psychol. Rev.*, 120(1):139–154, Jan. 2013.

[180] A. H. Sinclair and M. D. Barense. Surprise and destabilize: prediction error influences episodic memory reconsolidation. *Learn. Mem.*, 25(8):369–381, Aug. 2018.

[181] B. F. Skinner. *The Behavior of Organisms: An Experimental Analysis.* B. F. Skinner Foundation, Dec. 2019.

[182] K. E. Smith and S. D. Pollak. Rethinking concepts and categories for understanding the neurodevelopmental effects of childhood adversity. *Perspect. Psychol. Sci.*, 16(1):67–93, Jan. 2021.

[183] L. B. Smith, S. Jayaraman, E. Clerkin, and C. Yu. The developing infant creates a curriculum for statistical learning. *Trends Cogn. Sci.*, 22(4):325–336, Apr. 2018.

[184] I. M. Sobol. Distribution of points in a cube and approximate evaluation of integrals. *Zh. Vych. Mat. Mat. Fiz.*, 7:784–802, 1967.

[185] L. H. Somerville, S. F. Sasse, M. C. Garrad, A. T. Drysdale, N. Abi Akar, C. Insel, and R. C. Wilson. Charting the expansion of strategic exploratory behavior during adolescence. *J. Exp. Psychol. Gen.*, 146(2):155–164, Feb. 2017.

[186] A. D. Spadoni, M. Vinograd, B. Cuccurazzu, K. Torres, L. M. Glynn, E. P. Davis, T. Z. Baram, D. G. Baker, C. M. Nievergelt, and V. B. Risbrough. Contribution of early-life unpredictability to neuropsychiatric symptom patterns in adulthood. *Depress. Anxiety*, 39(10-11):706–717, Oct. 2022.

[187] E. S. Spelke and K. D. Kinzler. Core knowledge. *Dev. Sci.*, 10(1):89–96, Jan. 2007.

[188] V. Srivastava, P. Reverdy, and N. E. Leonard. On optimal foraging and multi-armed bandits. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 494–499, Oct. 2013.

[189] L. A. Sroufe. Attachment and development: a prospective, longitudinal study from birth to adulthood. *Attach. Hum. Dev.*, 7(4):349–367, Dec. 2005.

[190] C. K. Starkweather, B. M. Babayan, N. Uchida, and S. J. Gershman. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.*, 20(4):581–589, Apr. 2017.

[191] S. A. Stuart, J. K. Hinchcliffe, and E. S. J. Robinson. Evidence that neuropsychological deficits following early life adversity may underlie vulnerability to depression. *Neuropsychopharmacology*, 44(9):1623–1630, Aug. 2019.

[192] E. Sumner, A. X. Li, A. Perfors, B. Hayes, D. Navarro, and B. W. Sarnecka. The exploration advantage: Children's instinct to explore allows them to find information that adults miss. Sept. 2019.

[193] J. A. Sumner, N. L. Colich, M. Uddin, D. Armstrong, and K. A. McLaughlin. Early experiences of threat, but not deprivation, are associated with accelerated biological aging in children and adolescents. *Biol. Psychiatry*, 85(3):268–278, Feb. 2019.

[194] Y. Sun, J. Fang, Y. Wan, P. Su, and F. Tao. Association of Early-Life adversity with measures of accelerated biological aging among children in china. *JAMA Netw Open*, 3(9):e2013588, Sept. 2020.

[195] S. Tanaka, J. Ribot, K. Imamura, and T. Tani. Orientation-restricted continuous visual exposure induces marked reorganization of orientation maps in early life. *Neuroimage*, 30(2):462–477, Apr. 2006.

[196] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, Sept. 1974.

[197] H. van Seijen, M. Fatemi, and A. Tavakoli. Using a logarithmic mapping to enable lower discount factors in reinforcement learning. *CoRR*, abs/1906.00572, 2019.

[198] O. M. Vikbladh, M. R. Meager, J. King, K. Blackmon, O. Devinsky, D. Shohamy, N. Burgess, and N. D. Daw. Hippocampal contributions to model-based planning and spatial memory. *Neuron*, 102(3):683–693, 2019.

[199] A. G. P. Wakeford, E. L. Morin, S. N. Bramlett, B. R. Howell, K. M. McCormack, J. S. Meyer, M. A. Nader, M. M. Sanchez, and L. L. Howell. Effects of early life stress on cocaine self-administration in post-pubertal male and female rhesus macaques. *Psychopharmacology*, 236(9):2785–2796, Sept. 2019.

[200] J. X. Wang. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38:90–95, Apr. 2021.

[201] D. M. Werchan, A. G. E. Collins, M. J. Frank, and D. Amso. 8-month-old infants spontaneously learn and generalize hierarchical rules. *Psychol. Sci.*, 26(6):805–815, June 2015.

[202] D. A. White, J. Myerson, and S. Hale. How cognitive is psychomotor slowing in depression? evidence from a meta-analysis. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.*, 4(3):166–174, Sept. 1997.

[203] A. M. Wikenheiser, D. W. Stephens, and A. D. Redish. Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task. *Proc. Natl. Acad. Sci. U. S. A.*, 110(20):8308–8313, May 2013.

[204] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen. Humans use directed and random exploration to solve the explore-exploit dilemma. *J. Exp. Psychol. Gen.*, 143(6):2074–2081, Dec. 2014.

[205] T. Wise, K. Emery, and A. Radulescu. Naturalistic reinforcement learning. *Trends Cogn. Sci.*, Sept. 2023.

[206] M. K. Wittmann, N. Kolling, R. Akaishi, B. K. H. Chau, J. W. Brown, N. Nelissen, and M. F. S. Rushworth. Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nat. Commun.*, 7:12327, Aug. 2016.

[207] V. Wyart and E. Koechlin. Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, 11:109–115, Oct. 2016.

[208] A. J. Yu and J. D. Cohen. Sequential effects: Superstition or rational behavior? *Adv. Neural Inf. Process. Syst.*, 21:1873–1880, 2008.

[209] X. Y. Zhang, H. Sanchez, P. Kehoe, and T. A. Kosten. Neonatal isolation enhances maintenance but not reinstatement of cocaine self-administration in adult male rats. *Psychopharmacology*, 177(4):391–399, Feb. 2005.

[210] J. Zhu, C. M. Anderson, K. Ohashi, A. Khan, and M. H. Teicher. Potential sensitive period effects of maltreatment on amygdala, hippocampal and cortical response to threat. *Mol. Psychiatry*, Mar. 2023.

Appendix A

Supplementary information for Chapter 3

A.1 Methods

A.1.1 Parameter priors

Priors were beta-distributed with parameters, $\alpha = 1.1, \beta = 1.1$. Parameters whose bounds fell outside of the range between 0 and 1 were transformed to fit between those bounds using a logistic transformation.

A.1.2 Model recovery

For each model, we simulated data from 200 participants by uniformly drawing parameters between the bounds defined in Table SA.1. We then fit each simulated participant's dataset to all four models. Our models demonstrated good recoverability. The model used to gen-

Model	Parameter	Bounds
Structure learning and adaptive discounting	α	0 or 1
	γ_{base}	0,10
	γ_{coef}	-1,1
	v	0.01,1
Marginal value theorem	η	0,1
	v	0.01,1
Temporal-difference learning	η	0,1
	γ	0,1
	v	0.01,1

Table A.1: Bounds for parameters in each model.

erate the data was identified as the best-fitting model for the majority ($\geq 75\%$) simulated participants (Fig A.1A).

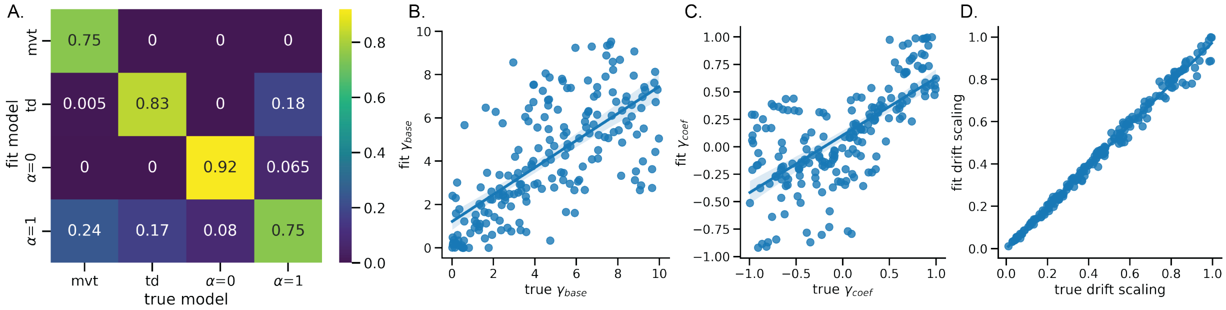

Figure A.1: A. Model recovery results The majority of the 200 simulated participants were best fit by the same model that was used to generate their data. **B-D. Parameter recovery results** Correlations between simulated and recovered parameter values for the $\alpha = 1$ structure learning model ranged from .50 to .95.

A.1.3 Parameter recovery

To ensure parameters were reliably recoverable from the data, we simulated data from 200 participants using our primary model of interest, the structure learning with $\alpha = 1$. Parameters other than α were uniformly drawn from bounds defined in Table SA.1. We examined the correlation between the "true" generating parameters, and the parameters our fitting procedure identified as providing the best account of the data. The strength of the correlation between the "true" simulated parameter value and the recovered parameter value varied.

According to standardized thresholds for intraclass correlation coefficients (used similarly for assessing parameter recoverability in **(author?)** [169]), the discounting parameters, γ_{base} and γ_{coef}, showed moderate recoverability (Fig A.1BC, γ_{base}: τ=0.53, $p < .001$; γ_{coef}: τ=0.50, $p < .001$) while the parameter the drift rate scaling factor showed high recoverability (Fig A.1D, τ=0.95, $p < .001$).

A.2 Results

In the tables that follow, we present the full results from our mixed-effects regression models.

Parameter	β	p-value
intercept	1.30	< .001
age (z-scored)	0.059	.47
poor galaxy	-0.63	< .001
rich galaxy	-0.42	.0018
planet number	-0.24	< .001
age x poor galaxy	-0.045	.39
age x rich galaxy	0.36	.0078
age x planet number	-0.060	.24
poor galaxy x planet number	0.067	.15
rich galaxy x planet number	-0.26	< .001
age x poor galaxy x planet number	0.017	.71
age x rich galaxy x planet number	-0.058	.34

Table A.2: Results from a mixed effects model regressing planet type, planet number, and age on the difference between the participants' actual planet residence time and the MVT-optimal residence time. We did not find any interaction between age, planet number, and richness on overharvesting.

Parameter	β	p-value
intercept	-0.0086	0.51
age (z-scored)	-0.0022	.87
switch point	0.049	.038
planet number	-0.049	$< .001$
age x switch point	0.0083	.71
age x planet number	-0.024	.047
switch point x planet number	0.014	.55
age x switch point x planet number	-0.019	.44

Table A.3: Results from a mixed effects model regressing presence of a switch in planet type, planet number, and age on reaction times (z-scored within participant and log-transformed). We also did not find any baseline differences in reaction time nor interaction between age, switch point, and planet number.

Parameter	β	p-value
intercept	0.71	$< .001$
age (z-scored)	0.22	.043
switch point	0.31	$< .001$
planet number	-0.31	$< .001$
age x switch point	0.0094	.81
age x planet number	0.018	.69
switch point x planet number	-0.078	.065
age x switch point x planet number	-0.11	.0082

Table A.4: Results from a mixed effects model regressing presence of a switch in planet type, planet number, and age on on the difference between the participants' actual planet residence time and the MVT-optimal residence time. In the absence of a switch point, overharvesting similarly occurred as did its decrease with experience.