Graduate School of Neural Information Processing
University of Tübingen

Allocate Your Resources Wisely!
Rate–Distortion Theory in Reinforcement Learning

Essay Rotation Report

Ali Gholamzadeh

The study was supervised by

Mani Hamid and Dr. Charley Wu

Cluster of Excellence in Machine Learning

Duration of the lab rotation: Oct 09 – Nov 24, 2023
Deadline for submission: 11 Dec 2023

Abstract

In this review paper, we explored the intersection of rate-distortion theory and reinforcement learning, examining how cognitive constraints influence decision-making. The concept of policy compression was introduced, which assesses the trade-off between decision-making efficiency and the use of cognitive resources. Our research provides new insights into behavioral patterns such as perseveration and cognitive limitations observed in schizophrenia. By proposing innovative models, we enhance the understanding of human and animal decision-making in complex environments, focusing on optimizing rewards while managing cognitive resources efficiently.

1. Introduction

The effectiveness of decision-making is contingent upon the extent of cognitive resources available, since these resources are integral to making well-informed choices (Lai & Gershman, 2023). The concept of "bounded rationality" frames our understanding that the quality of human decisions is constrained by the availability of cognitive resources (Rubinstein, 1998; Gigerenzer & Selten, 2002). This idea is further refined in the concept of "resource rationality," which interprets bounded rationality through a computational perspective, focusing on the physical constraints of machines running decision-making algorithms (Lieder & Griffiths, 2019; Bhui et al., 2021). This approach underscores the significance of the capacity to store and transmit information in decision-making processes.

Information theory offers a framework to comprehend the boundaries of storage and transmission capabilities. It defines the capacity of a channel as the highest possible mutual information between the channel's inputs and outputs (often referred to as the rate), calculated across all feasible input distributions. Within the realm of decision making, these inputs can be considered as states (which are visible to the decision-maker) and the outputs as the resultant actions. The mapping from states to actions constitutes the decision-maker's policy. The agent's policy complexity is determined by how much the states and actions are interrelated, known as mutual information. The maximum complexity the agent can manage is set by the highest level of complexity its action selection process can handle. This upper limit defines the most sophisticated policy the agent can achieve. Fundamentally, the capacity of this channel influences how dependent the action selection is on the state: while more adaptable state-action mappings can lead to better performance, the channel's capacity restricts the agent's proficiency in differentiating between states, necessitating a representation of states with less detail. This scenario sets up a balancing act between the channel's capacity and the effectiveness of task performance. When the capacity is limited, the agent has to streamline or "compress" their action policies, meaning the state-dependence is reduced, which in turn can lead to a decrease in performance Lai & Gershman, 2021.

Recent research employing the concept of policy compression has shed light on various decision-making phenomena, such as perseveration (Gershman, 2020) cognitive impairments in schizophrenia (Gershman & Lai, 2021), navigational strategies in mice (Amir et al., 2020), and the undermatching effect (Bari & Gershman, 2022), among others. These investigations, however, primarily analyzed data retrospectively from previously published studies. Much of this work has centered on understanding the impact of set size on performance, which typically decreases as the number of distinct stimuli to be remembered increases (Gershman, 2020; Gershman & Lai, 2021). Additionally, policy compression has been used to provide standard explanations for perseveration, a tendency to repeat the same action policy across different states, regardless of the reward outcome (Gershman, 2020). This body of research has significantly contributed to our understanding of how cognitive constraints influence learning and decision-making processes.

This paper merges key insights from information theory and reinforcement learning to provide an exploration of how decision-making processes and action selection are influenced by cognitive constraints. It aims to offer a summary of existing works on the adaptive strategies humans employ to navigate complex environments, balancing the demands of reward maximization and cognitive resource management. Additionally, the paper introduces and tests new models, further enriching the existing framework with new perspectives and insights.

2. Material & Methods

2_1. Information Theory:

Information theory is a multidisciplinary scientific field that intersects mathematics and engineering. It originated with Claude Shannon's groundbreaking work, titled "A Mathematical Theory of Communication" (Shannon, 1948), and expanded the following year with Warren Weaver's introductory essay in "The Mathematical Theory of Communication" (Shannon & Weaver, 1949). The primary goal of information theory is to quantify and understand the transmission, processing, and utilization of information. This theory provides a mathematical framework for analyzing communication systems and understanding the fundamental limits on compressing and reliably transmitting data.

A fundamental concept in information theory is the ”bit”, a unit of information measurement. Contrary to common belief, a bit is not limited to binary digits (1s and 0s) but can also represent information conveyed in analog forms, such as the graded signal from a retinal photoreceptor. Central to information theory are the notions of “random variables” and “entropy”. A random variable, denoted as *X*, can assume a range of values, each with a specific probability. For example, a fair dice roll is a random variable with six possible outcomes, each having a probability of $\frac{1}{6}$. Entropy, quantified using the formula:

$$H(x) = \sum_i P(x_i)\, \log P(x_i) \qquad \textit{Eq. (1)}$$

measures the average surprise or uncertainty in the outcomes of a random variable. (Sims, 2016)

This equation represents the surprise of each outcome, weighted by its probability, defining entropy in information theory. For a binary random variable with two equally likely outcomes, such as a fair coin flip, its entropy is 1 bit if the logarithm is base 2, and 1 nat (equivalent to 1.44 bits) when using the natural logarithm. Entropy quantifies the information inherent in a random variable. In the context of a communication channel (as shown in Fig. 1), the input is samples from this random variable *x*, and the output is another random variable y. The channel's function is to map the input to the output through a conditional probability distribution, *P(y|x)*.

To evaluate the information transmitted by a channel, we comprehend that the information it conveys cannot exceed that of its source. If the channel is prone to noise or errors, the information output might be less than the source's entropy. Therefore, both the information source and the channel's characteristics are crucial in assessing information communication. This relationship is mathematically expressed through the mutual information between the channel's input and output:

$$I(x:y) = H(x) - H(x|y) = \sum_i \sum_j p(y_j|x_i) P(x_i)\, log \frac{p(y_j|x_i)}{p(y_j)} \qquad \textit{Eq. (2)}$$

Mutual information quantifies the decrease in uncertainty about the channel input upon observing its output. A high-information channel significantly lowers uncertainty about the input signal. Measured in bits or nats like entropy, mutual information determines the information rate of a channel and its source. This rate refers to the average number of bits transmitted per symbol from the source alphabet (bits per symbol), not necessarily time (bits per second). For instance, with outcomes from a 6-sided die roll communicated through a channel, the channel's information rate, as defined by mutual information, is the average bits conveyed per roll, considering the reduction in uncertainty across the sequence of outcomes.

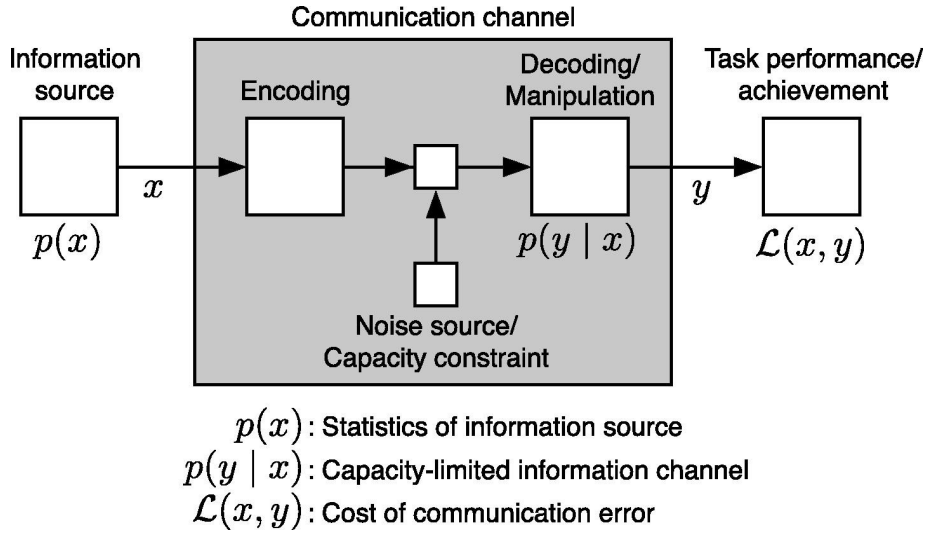


Figure 1. Rate-distortion theory is based on key concepts where an information source is characterized by a probability distribution across its set of alphabets. When samples from this source are transmitted over a channel with noise or limited capacity, it leads to a conditional probability distribution of the output from the channel. Additionally, a cost function is used to determine the impact of errors in the communication process (Sims, 2016)

The mutual information is influenced by the properties of information source, *P(x)*, and the channel's behavior, *P(y|x).* Using a constant channel, *P(y|x)*, to transmit signals from various source distributions will yield different mutual information levels, higher for some sources and lower for others. A channel's capacity is the maximum mutual information across all potential information sources *p(x)*.

2_2. Rate-Distortion Theory

Defining a perfect communication channel might seem straightforward — it's a channel where the output always matches the input, like having a flawless visual memory. Perfect communication is often infeasible, due to several factors: the channel's capacity may be less than the source's entropy, the channel's output alphabet might have fewer symbols than the input, or inherent limitations in the channel could reduce its accuracy. In such situations, the aim of error-free communication is unrealistic.

Instead, the focus shifts to minimizing communication errors. This necessitates defining a cost function, *L(x, y),* which calculates the cost when an input signal *x* is transmitted as value *y*. Channels that closely replicate the input in their output will incur a lower cost under this measure. The distortion, or the average cost for a specific information source and channel, is then calculated based on this function.

$$D = L(x_i, y_j) = \sum_i \sum_j L(x_i, y_j) p(y_j|x_i) P(x_i) \qquad \textit{Eq. (3)}$$

When a specific cost function and information source are defined, the optimal communication channel is the one that minimizes distortion. Generally, channels with greater capacity are more effective at reducing costs compared to those with lower capacity. However, since every physical channel, denoted as *Q*, has a finite capacity, this leads to an optimization challenge within certain constraints. This issue is the central focus of rate-distortion theory (Berger, 1971 and Shannon, 1959). The problem can be formally described as follows:

$$Q^* = arg\,min\, D_{(Q)}\ s.t.\ I_{(Q)}(x, y) \leq C \qquad \textit{Eq. (4)}$$

The equation describes that the optimal information channel, labeled Q^*, for a given information source is the one that minimizes channel distortion, with the condition that the mutual information between the source and channel doesn't exceed a specified capacity constraint (*C*). The “argmin” operator identifies the channel configuration that results in the least distortion. The distortion

$D_{(Q)}$ and mutual information $I_{(Q)}$ are calculated based on a specific channel *Q* and its conditional probability distribution *P(y|x),* with optimization over all potential channels.

Shannon's noisy-channel coding theorem (Shannon & Weaver, 1949) asserts that if a channel's capacity exceeds the information rate of the source, it is feasible to achieve an extremely low-error rate. On the other hand, if the channel's capacity is less than the information rate, errors are inevitable. Rate-distortion theory (RDT) highlights a fundamental tradeoff: reducing distortion (average cost) necessitates increasing the information transmitted by the channel. This balance is depicted in a rate-distortion curve(Fig. 2), showing the minimum channel capacity required for a certain performance level.

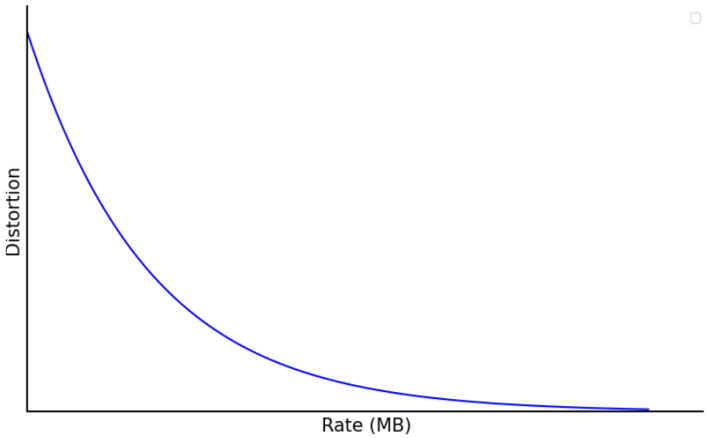


Figure 2. Common Rate-Distortion Curve

2_3. Rate-Distortion Theory (RDT) in Reinforcement Learning (RL):

Rate distortion theory bridges the gap between information theory and statistical decision theory, offering insights particularly relevant to cognitive psychologists (Sims, 2016). This explanation follows the somewhat unconventional terminology used by Parush et al., 2011, and Gershman, 2020, to better connect with the concepts introduced earlier. The theory assumes an agent either learns or has direct access to a value function *Q(s,a)*, which predicts the expected reward in a given state *s,* after taking an action *a*. Each state occurs with a probability *P(s)*, and actions are chosen based on a policy *π(a|s)*.

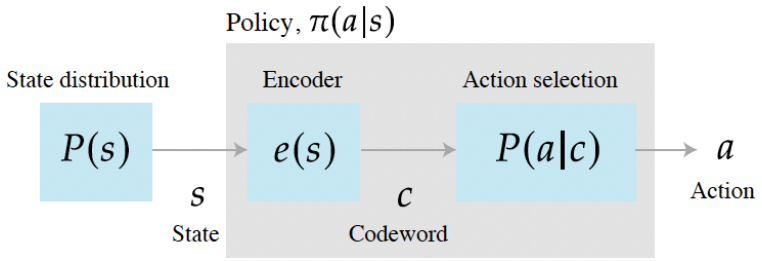


Figure 3. The conceptualization of the policy as a communication channel begins with a distribution *P(s)* over various states (s). Each state is then transformed into a memory-encoded codeword through an encoder, *e(s)*, resulting in codeword *c*. This codeword is subsequently linked to an action a based on the probability *P(a|c)*. The combination of encoding and action selection culminates in the formation of the policy *π(a|s)*, which effectively maps states to corresponding actions. (Lai & Gershman, 2021)

In terms of information theory, the distribution of states is treated as the source, and the decision-making strategy operates like a noisy channel(Fig. 3). This channel transforms the states (messages) into an internal representation (codewords), which then inform the resulting actions (output signals). The essential rate, or the average length of these codewords necessary to implement a strategy with negligible error, is effectively the mutual information between the states and the

actions. This perspective offers a novel way to understand decision-making dynamics through the lens of information theory (Gershman, 2020 and Lai & Gershman, 2021).

The concept of compression in this context is measured by the description length of a state, essentially the length of its corresponding codeword. Codewords are translatable into binary strings of 0s and 1s, known as bits, which allows for the comparison of their lengths in a standard unit of bits. Channel capacity imposes a limit on the average length of these codewords. Importantly, the minimum bit count required for transmitting the state's identity without errors is determined by the mutual information between states and actions $I(S:A)$.

Mutual information serves as a measure of the probabilistic relationship between states and actions, and is thus termed as policy complexity. Essentially, the more a policy relies on specific states (like the increased complexity in driving when navigating based on current location), the higher its complexity. In contrast, if a policy is uniformly applied across all states without variation, its complexity is reduced to the lowest level, resulting in mutual information being zero (Gershman, 2020).

Compression is crucial in real-world scenarios with numerous states and actions, such as the variety of driving directions in a city, because systems with limited resources cannot capture every possibility in an extensive look-up table. This contradicts the earlier belief that look-up tables are computationally inexpensive, as they may be easy to process but require a lot of data storage, indicating high policy complexity (Kool, Cushman, & Gershman, 2018). Rate distortion theory has been applied in psychology to illustrate memory confusability, emphasizing the impracticality of storing every detail in a look-up table (Sims et al., 2012).

The channel design problem focuses on how the brain can optimally encode and decode information to minimize distortion or maximize reward, given certain capacity limitations. To achieve optimality in lossless transmission, the encoder should conform to the Shannon bound, where the average description length is equal to the information rate. In a scenario with no channel noise, the inputs are clearly distinguishable from the outputs, resulting in a conditional entropy *H(S|A)* of zero. Therefore, according to Eq. (2), policy complexity equates to the source entropy, *H(S)*, and the shortest average description length also matches this source entropy.

Error-free coding under noiseless conditions can be achieved through entropy coding algorithms. Entropy coding is a method of encoding data based on the frequency or probability of occurrence of various data elements. The principle is to use shorter codes for more frequent elements and longer codes for less frequent ones, thereby reducing the overall length of the message. Huffman coding, a specific type of entropy coding developed (by Huffman, 1952), exemplifies this approach. It constructs a binary tree where each leaf node represents a data element (such as a character in a text file), and the path from the root to a leaf node forms the binary code for that element. The tree is designed so that the most common elements have the shortest paths (and thus the shortest codes), optimizing the average code length based on the statistical frequency of each element. This efficient coding reduces the size of the encoded data, which is particularly useful in data compression and transmission.

The concept of error-free transmission is unsuitable for action selection because the brain inherently makes errors, as recognized by Von Neumann, 1958. It operates as a low-precision, noise-affected system, making completely redundancy-free compression methods like Huffman coding unrealistic. In systems prone to errors like the brain, redundancy in bits is essential for correcting transmission errors, as noted by Bhui & Gershman, 2018. However, when capacity is limited and insufficient bits are available to rectify every error, the question arises: how should the brain efficiently allocate these limited bits?

The foundational idea in applying rate-distortion theory to action selection is the prioritization of transmitting significant information. This involves a distortion function that assesses the cost of specific actions in given states, aiming to minimize expected distortion while maintaining an information rate constraint. In action selection scenarios, it often makes more sense to consider the reward *Q(s, a)*, essentially the inverse of distortion. The application of rate-distortion theory to action selection has been explored in depth by several researchers, including Fox, Pakman, & Tishby, 2016; Grau-Moya, Leibfried, & Vrancx, 2018; Lerch & Sims, 2018; Parush, Tishby, & Bergman, 2011; Still & Precup, 2012; and Tishby & Polani, 2011 ;Lai & Gershman, 2023. This summary, however, is a

simplified version of those ideas, focusing on the essentials rather than delving into technical details. Interested Readers are encouraged to consult these papers for detailed information.

2_4. Mathematical Formulation:

In this approach, we aim to frame the Reinforcement Learning (RL) problem through the lens of Rate-Distortion Theory (RDT). In RL, the agent's objective is to perform actions that maximize its reward. The level of data compression in this context is directly influenced by the potential reward that can be obtained, given the complexity of the policy being employed. The average reward in this scenario can be represented as follows:

$$V^{\pi} = \sum_{s} P(s) \sum_{a} \pi(a|s)\, Q(s,a) \qquad \textit{Eq. (5)}$$

We can now formulate the optimization problem:

$$\pi^{*} = argmax\ V^{\pi}\, s.t.\ I_{\pi}(x,y) \leq C \qquad \textit{Eq. (6)}$$

where C represents the channel capacity, which is the highest level of policy complexity that can be achieved.

This representation considers the balance between achieving high rewards and managing policy complexity, a core principle in both RL and RDT. In essence, it's about finding the optimal strategy that yields the highest rewards with an acceptable level of complexity or data compression. This optimization problem, which implicitly includes additional constraints (non-negative action probabilities that sum to 1), can be reformulated as a Lagrangian:

$$\pi^{*} = argmax\ \beta V^{\pi} - I_{\pi}(x,y) - \sum_{s} \lambda(s) \left(\sum_{a} \pi(a|s) - 1 \right) \qquad \textit{Eq. (7)}$$

with Lagrange multipliers *β* and *λ(s)*. The solution to this problem has the following form(Parush et al., 2011; Still & Precup, 2012; Tishby & Polani, 2011):

$$\pi^{*}(a|s) \propto exp[\beta\, Q(s,a) + log\, P^{*}(a)] \qquad \textit{Eq. (8)}$$

$$log\, P^{*}(a) = log \sum_{s} P(s)\, \pi^{*}(a|s) \qquad \textit{Eq. (9)}$$

Several familiar elements emerge in this context. The optimal policy resembles a softmax function, widely used in reinforcement learning for simulating both artificial and biological agents. The Lagrange multiplier *β* functions as the "inverse temperature" parameter, influencing the exploration-exploitation balance by controlling the policy's stochasticity, as discussed by Sutton & Barto, 2018. A low *β* leads to a nearly uniform policy, while a higher *β* results in a policy increasingly focused on the highest-value action. However, it is important to note that the derivation of this optimal policy does not explicitly involve exploration (Still & Precup, 2012).

The second term within the softmax function represents a type of perseveration, indicating a preference for actions that are frequently selected in various states. Policies with low complexity compress the state information, leading to an inability to differentiate between policies for distinct states. Consequently, the optimal policy tends to align closely with the overall action distribution, disregarding the specific state involved.

To calculate the optimal policy, there's an inherent circularity to consider: the perseveration term is dependent on the optimal policy itself. The Blahut-Arimoto algorithm (Arimoto, 1972; Blahut, 1972), developed by Arimoto in 1972 and Blahut in the same year, effectively utilizes this circularity. It iterates between computing *P(a)* and *π(a|s)*, gradually converging to the optimal values. This iterative process allows for the construction of a reward-complexity curve by optimizing across various *β* values. This curve illustrates the most efficient policy under a specific resource constraint, achieving maximum reward with minimal perseveration. The curve's concave nature signifies that *β* increases

monotonically with the complexity of the policy, reflecting a systematic relationship between resource allocation, policy complexity, and the resulting rewards.

In real-world applications, the Blahut-Arimoto algorithm faces significant challenges. One primary issue is the requirement for marginalizing over the entire state space to compute *P(a)*, which becomes computationally prohibitive in large or continuous state spaces. Additionally, the algorithm assumes access to the true *Q(s,a)* values – the expected rewards for specific actions in certain states. However, in many practical scenarios, agents do not have prior knowledge of these values and must instead learn them through trial and error, a reality often encountered in dynamic environments.

To address these limitations, a more feasible approach based on reinforcement learning principles has been proposed, as highlighted in Gershman & Lai, 2021. This alternative algorithm incrementally modifies the policy according to reward feedback. Crucially, it incorporates a penalty for complex policies, aligning more realistically with situations where agents learn and adapt from their experiences. While this method eventually converges to the optimal policy, its gradual nature means it often deviates temporarily from the ideal trade-off curve. This aspect of the algorithm is particularly valuable, as it provides a framework to explain and understand empirical deviations from the optimal curve, reflecting the learning and adaptation processes that real-world agents undergo (Lai & Gershman, 2021).

Additionally, considering that the original studies did not conduct model comparisons or assess the proposed model against the conventional standard RL models, we have examined a range of standard RL models and other plausible alternative models. Detailed descriptions of each of these models will be presented in the subsequent section.

2_5. Process Models for Learning Under Constraints:

The Blahut-Arimoto algorithm, while effective, becomes impractical as it necessitates having the knowledge of the state-action value function. To address this challenge, here we introduce a more practical model, based on the principles of reinforcement learning (RL).

2_5_1. RDT-Derived Actor-Critic learning(AC_RDT):

Transitioning to a more cognitively realistic model, as proposed by Gershman & Lai, 2021, the Lagrangian optimization problem, as described in Eq. (7), can be reinterpreted as an expectation over states space:

$$\pi^{*} = argmax\, \mathbb{E}\left[\beta r - log\frac{\pi(a|s)}{P(a)}\right] \qquad \textit{Eq. (10)}$$

This perspective facilitates the development of an "actor-critic" learning approach, rooted in the stochastic policy gradient algorithm (Sutton and Barto, 2018). This method focuses on optimizing Eq. (10) by computing the gradient of the average reward relative to the policy parameters. Initially, a parameterized policy, known as the "actor," is defined:

$$\pi_{\theta}(a|s) \propto exp(\beta\, \theta_{sa} + log\, P\ (a)) \qquad \textit{Eq. (11)}$$

Here, θ represents the policy parameters, mirroring the optimal parameterization in Eq (8). The inverse temperature β is indirectly indicative of the channel capacity, showing a monotonic relationship with policy complexity, which reaches its peak at the channel capacity. The process of updating the policy parameters involves considering the observed reward *r* after executing action *a* in state *s*:

$$\Delta\theta_{sa} = \frac{\alpha_{\theta}}{t}\delta[1 - \pi_{\theta}(a|s)]\beta \qquad \textit{Eq. (12)}$$

The actor learning rate is denoted by α_{θ}, and t represents the trial number. Then,

$$\delta = \beta r - log\frac{\pi(a|s)}{P(a)} - \widehat{V}(s) \qquad \textit{Eq. (13)}$$

is the prediction error of the “critic” and $\widehat{V}(s)$ is an estimate of the expected cost-sensitive reward, which is updated as follows:

$$\Delta\widehat{V}(s) = \alpha_v \delta \qquad \textit{Eq. (14)}$$

With α_V representing the critic learning rate, the scaling of the actor learning rate (but not the critic's) by *1/t* serves dual purposes: it ensures convergence of the policy to an optimum, satisfying the Robbins-Munro conditions for stochastic approximation algorithms (Robbins & Monro, 1951), and it generally maintains a slower actor learning rate compared to the critic (Konda & Tsitsiklis, 2000). To complete the model, the marginal action probabilities are estimated using an exponential moving average:

$$\Delta P(a) = \alpha_p[\pi_\theta(a|s) - P(a)] \qquad \textit{Eq. (15)}$$

with learning rate α_P. We call this the "AC_RDT" model.

2_5_2. Alternative Models:

In addition to the model proposed by Gershman and Lai, 2021, we conducted tests on both the standard and various versions of cost-sensitive reinforcement learning models. The objective was to determine which model most accurately aligns with human data and provides a more precise explanation of behavior. Below are the specifications of each model we utilized:

Standard Q Learning (Q_Baseline): We examined a conventional reinforcement learning (RL) model of decision-making (Sutton R. S. & Barto A. G., 2018). In this model, an agent updates its action-value estimates for each state, *Q(s, a),* on every trial based on a delta rule (Rescorla R. A. & Wagner A. R., 1972):

$$\Delta Q(s,a) = \alpha_Q \delta \qquad \textit{Eq. (16)}$$

$$\delta = r - Q(s,a) \qquad \textit{Eq. (17)}$$

Here, δ represents the reward prediction error, α_Q is the learning rate, and *r* is the reward obtained in the current trial after executing action *a* in state *s*. These state-action values are then converted into choice probabilities using a softmax function with not perseveration term:

$$\pi(a|s) \propto exp(\beta Q(s,a)) \qquad \textit{Eq. (18)}$$

where β denotes the inverse temperature parameter. We refer to this as the "Q_Baseline" model.

RDT-Derived Q learning(Q_RDT): This model shares learning rules akin to the Q_baseline but with two significant distinctions. Firstly, it incorporates a cost term in the prediction error, reflecting the expense of utilizing cognitive resources, identical to the approach in the AC_RDT model.

$$\delta = \beta r - Q(s,a) - log\frac{\pi(a|s)}{P(a)} \qquad \textit{Eq. (19)}$$

Secondly, like the AC_RDT model, it includes a perseveration component within its policy parameterization.

$$\pi(a|s) \propto exp(\beta\, Q(s,a) + log\, P\,(a)) \qquad \textit{Eq. (20)}$$

Additionally, the estimation of marginal action probabilities is conducted in the same manner as in the AC_RDT model.

Cost-Sensitive Q learning(Q_Cost): This model is identical to the Q_RDT model, with the sole difference being the exclusion of the perseveration term in policy rule. Instead, it employs the basic softmax policy rule, like the Q_baseline model.

Disentangled Q learning(Q_tau): Disentangled Q Learning (Q_tau): This model operates on the premise that utilizing the same β parameter for both exploration rate and the reward-policy complexity trade-off is unnecessary. Therefore, it attempts to separate these elements, employing a different parameter (τ) specifically for the cost term penalty:

$$\delta = \tau r - Q(s,a) - log\frac{\pi(a|s)}{P(a)} \quad \textit{Eq. (21)}$$

This approach is referred to as the "Q_tau" model.

Basic Actor-Critic Learning (AC_Baseline): This model adheres to the conventional Actor-Critic method detailed by Sutton and Barto, 2018. Its architecture and learning guidelines are comparable to AC_RDT, but it excludes the cost factor in its prediction error:

$$\delta = r - \widehat{V}(s) \quad \textit{Eq. (22)}$$

Additionally, it implements a policy rule that, like Q_Baseline, does not include perseveration.

Cost-Sensitive Actor-Critic learning(AC_Cost): This model is identical to AC_RDT except for a key distinction: it utilizes the standard policy rule but does not include the perseveration term.

Disentangled Actor-Critic learning((AC_tau): This model follows the same learning and policy rules as AC_RDT, but distinguishes itself by segregating parameters for exploration and cognitive cost like the approach in the Q_tau model. It uses unique parameters for each of these aspects.

2.6 Dataset

We used the same dataset that was previously used by Gershman & Lai, 2021, which was initially presented in the work of Collins et al., 2014.. In this dataset, participants engaged in a reinforcement learning activity where the number of unique stimuli (representing different states) changed in each block of the experiment (as depicted in Fig. 4). During each trial, participants were shown a single stimulus, made a choice of action, and received consistent feedback regarding the reward. Each stimulus was linked to one specific action that would yield a reward. Participants completed 13 blocks in total, with the number of different stimuli in each block ranging from 2 to 6. Every stimulus was shown between 9 to 15 times in a block, determined by a performance criterion which required at least four correct responses in the last five presentations of each stimulus. Notably, no stimulus was repeated in subsequent blocks.

In the experiment, participants were divided into two groups: individuals diagnosed with schizophrenia and healthy control subjects. The group with schizophrenia, referred to here as SZ, included 49 participants (35 males and 14 females). This group was composed of individuals with a DSM-IV diagnosis, including 44 with schizophrenia and 5 with schizoaffective disorder. The healthy control group, referred to as HC, consisted of 36 individuals (25 males and 11 females). This group was demographically matched to the SZ group in several aspects, such as age, gender, race/ethnicity, and the educational background of their parents.

3. Results:

For each of the models we previously described, we explicitly fitted the model parameters to match the choice behavior of each individual, using maximum likelihood estimation. To evaluate how closely these fitted models mirrored actual data, we simulated each model for every participant, using identical stimuli to those presented to human subjects (Fig. 5a). In the process of selecting the most appropriate model among the various options, we utilized the Bayesian Information Criterion (BIC).

$$BIC = kln(n) - 2ln(\widehat{L}) \quad \textit{Eq. (23)}$$

The term $\widehat{L}$ represents the maximum value obtained from the likelihood function of the model. The variable n denotes the total number of trials conducted, and k stands for the quantity of parameters that the model estimates. The BIC is a statistical tool that compares models by balancing their complexity against their ability to explain the data. It helps in identifying models that achieve a good fit without overcomplicating the structure, thus penalizing models that are unnecessarily complex (Wit E. et al., 2012).

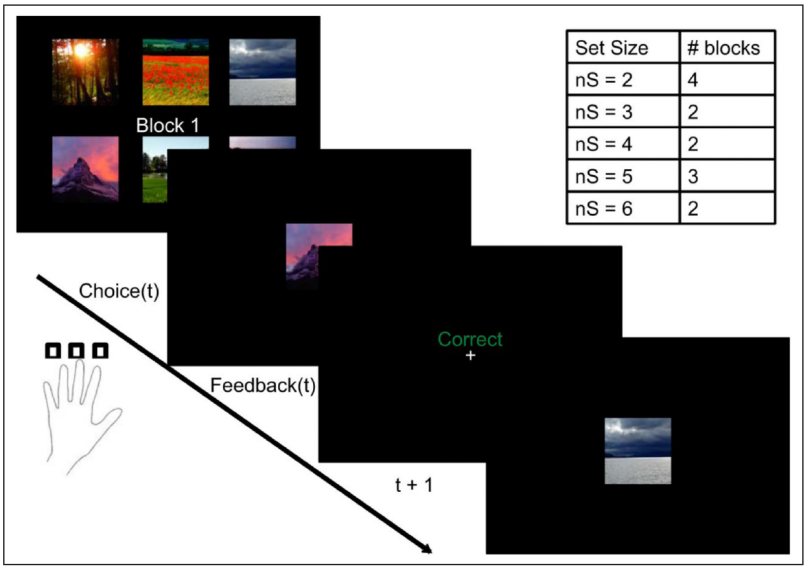
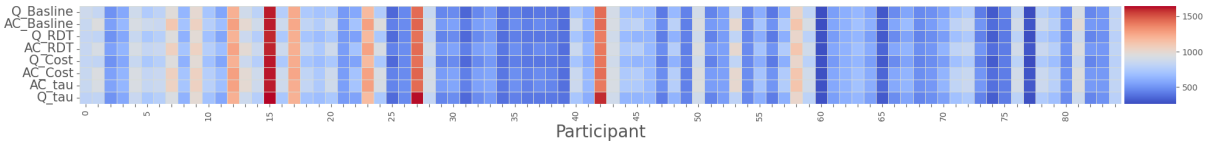
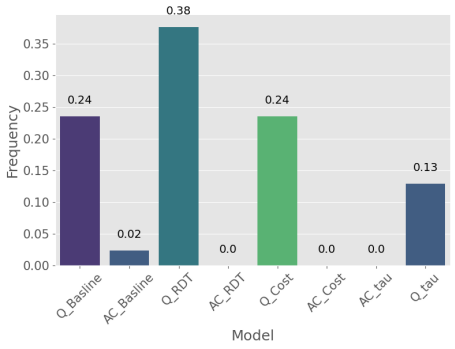


Figure 4. Overview of the Task. In each trial, participants were required to choose one out of three possible actions based on a given stimulus. Following their selection, they received definite feedback on the reward (correct/incorrect). The variety of stimuli (referred to as the set size) changed in different blocks (Collins et al., 2014).

For each participant's data, we selected the model with the lowest Bayesian Information Criterion (BIC) measure. In this scenario, the model most frequently chosen was Q_RDT, our recently adapted model, which overall showed better BIC values compared to others, as depicted in Fig. 5.b. This model includes a cost term for the use of cognitive resources and also an explicit perseveration term in its softmax policy rule. Notably, it also proved to be a better fit for the participant data compared to the original AC_RDT model, as detailed in a study by Gershman & Lai, 2021.



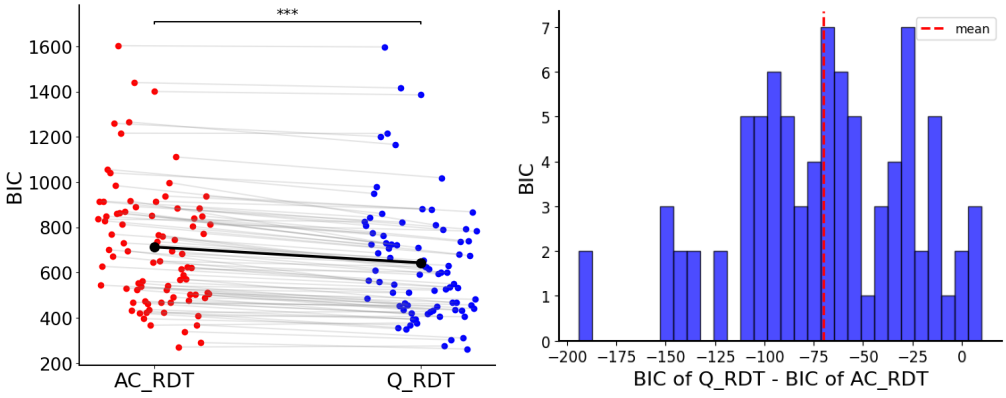
(a)



(b)

Figure 5. (a) BIC values for each fitted model for every participant. (b) Model Frequency

In the next step of our analysis, we compared the Q_RDT model with the proposed model Gershman & Lai, 2021, AC_RDT, which uses Actor-Critic Learning instead of action value learning (Fig. 6). Our objective was to determine if Q_RDT significantly outperforms AC_RDT in fitting the participants' data. To accomplish this, we used the non-parametric Wilcoxon signed-rank test, a method suited for comparing two related groups. This test was specifically applied to see if Q_RDT exhibits a notably lower BIC than AC_RDT. Our findings confirmed that there is a statistically significant difference in the BIC values of these two models (W=13.0, p-value=1.854e-15), indicating a better fit for the Q_RDT model.



(a) (b)

Figure 6. (a) Pairwise comparison of BIC for Q_RDT and AC_RDT across each individual. (b) Histogram depicting the differences in BIC values between the two models.

We examined how closely subjects align with the optimal reward-complexity trade-off curve, as shown in Fig. 7a–e. Key insights emerged from these comparisons. Firstly, there was a notable correlation between the optimal and empirical trade-off curves for both healthy controls (HC) and schizophrenia patients (SZ), reinforcing previous research by Gershman, 2020, that individuals tend to approach the optimal trade-off, especially those with high policy complexity.

Secondly, consistent with earlier findings (Collins, 2018; Collins & Frank, 2012, 2018), subjects earned less reward with larger set sizes, suggesting a limitation in reinforcement learning resources. This implies that a fixed policy complexity resource, when spread over more states, results in decreased precision per state. Thirdly, the average policy complexity didn't consistently change with varying set sizes for either group, as illustrated in Fig. 7f. This suggests a roughly constant resource constraint, supporting the idea that set size effects are due to reallocation of a fixed resource (Ma et al., 2014).

Lastly, the SZ group showed significantly lower policy complexity [mixed-effects ANOVA: $F(1,415) = 11.51$, $p < 0.001$], without interaction with set size ($p = 0.14$). This indicates that SZ subjects used fewer cognitive resources in the task.

In examining the data across all set sizes, it was evident that schizophrenia patients exhibited lower policy complexity compared to the healthy control group. This pattern is expected to be reflected in our model. Focusing on the best-fitting model, Q_RDT, we compared the fitted parameters between the two groups. The comparison, as shown in the figure, includes the fitted learning rate and inverse temperature for both healthy and schizophrenia groups. For the learning rate (Fig. 8.a), healthy subjects had, on average, higher rates (Mean = 0.249, Standard Deviation = 0.05) than schizophrenia patients (Mean = 0.219, Standard Deviation = 0.106), but the difference was not statistically significant (t-statistic = -1.513, p-value = 0.134). The analysis also included the inverse temperature parameter (Fig. 8.b), where the average for the healthy group was higher (Mean = 3.026, Standard Deviation = 0.438) compared to that of the schizophrenia patients (Mean = 2.77, Standard Deviation = 0.584), with this difference reaching statistical significance (t-statistic = -2.188, p-value = 0.031). Given that higher.

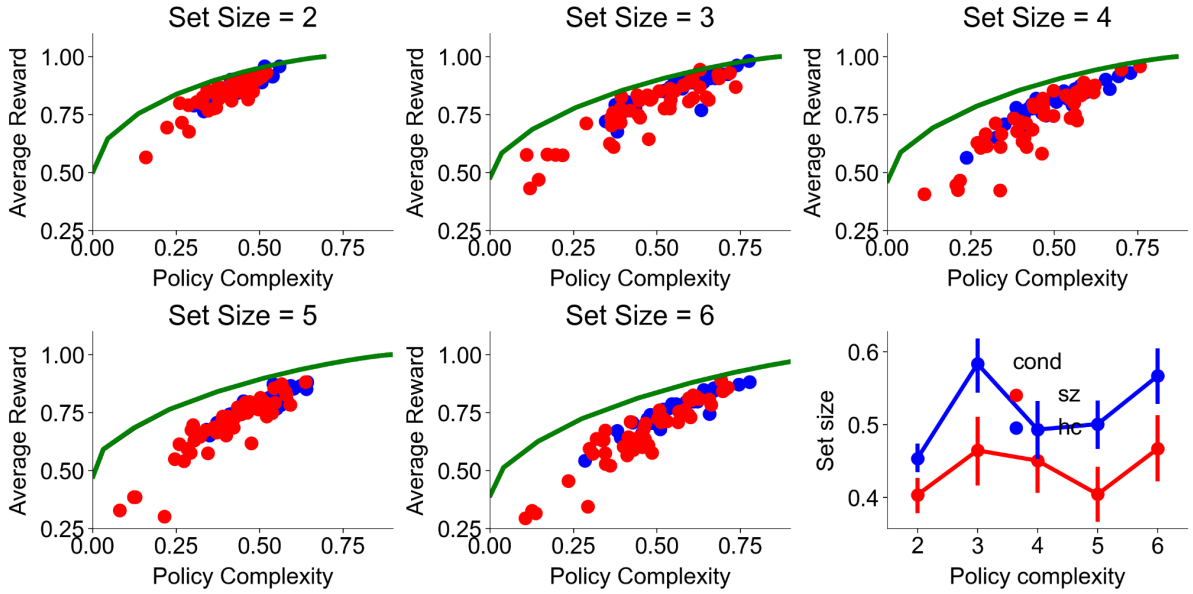


Figure 7. Analyzing the Balance Between Reward and Complexity. (A–E) Each graph depicts the ideal reward-complexity relationship (depicted as a solid line) for varying set sizes, along with the actual reward-complexity data (illustrated with circles) for individual subjects (HC = healthy controls; SZ = schizophrenia patients). Here, policy complexity is quantified in natural information units (nats), with a uniform action distribution equating to a policy complexity of 0. It's important to note that the optimal curve, assuming precise knowledge of deterministic reward outcomes, will invariably peak at an accuracy of 1. (F) Graph showing the relationship between policy complexity and set size, with error bars representing 95% confidence intervals

beta values indicate greater policy complexity, these findings are consistent with the empirical observation that schizophrenia patients tend to develop less complex policies

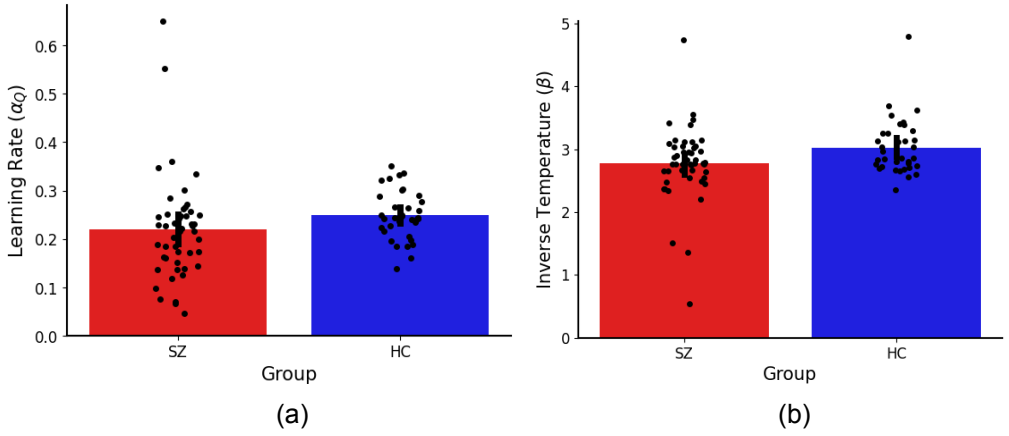


Figure 8. Comparison of Parameters Between Healthy(HC) and Schizophrenia(SZ) Groups: (a) Learning Rate (b) Inverse Temperature.

4. Behavioral signatures of policy compression:

In this part of the paper, we examine various behavioral patterns that may be explained through the concept of policy compression. This approach is more synthetic than discriminative. While different theories might account for each behavior individually, we propose that these phenomena can be collectively interpreted as manifestations of a singular fundamental principle.

4_1. Stochasticity:

The stochastic nature of human actions, even when faced with the same choices, has been a topic of discussion in various fields (Schulz & Gershman, 2019). One perspective from reinforcement learning suggests that stochasticity aids exploration, as agents must sample multiple actions to find the most rewarding one. This is often implemented via randomness in policies, such as the softmax equation. However, this view does not fully account for the stochastic nature of actions in well-known payoff situations (Mosteller & Nogee, 1951).

Rate-distortion theory offers an alternative explanation. It proposes that optimal stochastic behavior is inherent for capacity-limited agents functioning at the reward-complexity frontier, aligning with the softmax policy used in psychology and economics (Matejka & McKay, 2015). This theory connects the 'inverse temperature' in softmax policies to the reward-complexity trade-off, implying that more stochastic actions occur with lower policy complexity.

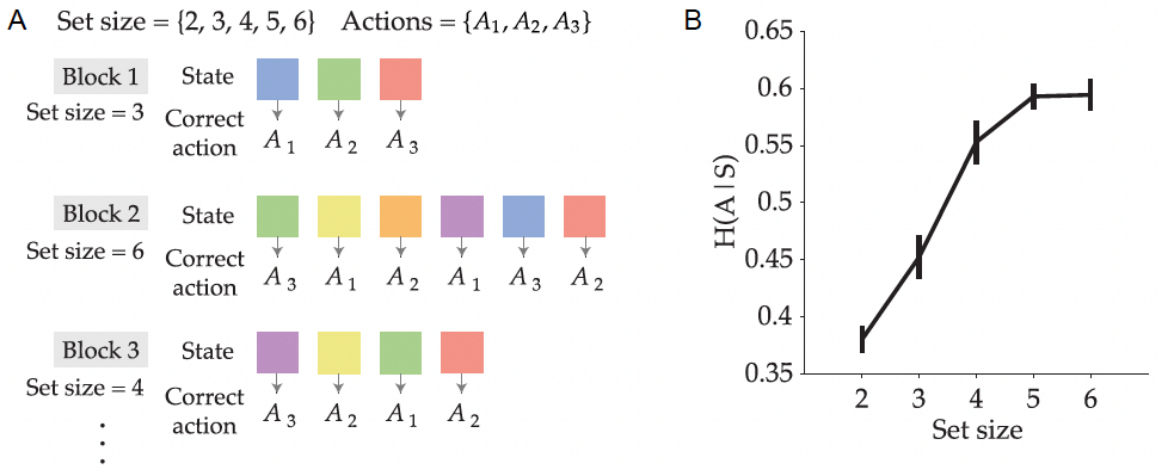


Figure 9. Stochasticity as a function of set size. (A) In the task developed by Collins and Frank, 2012, subjects saw a single stimulus on each trial and chose between 3 actions. Each stimulus had a single rewarded action. The number of states, or set size, varied between 2 and 6 across blocks. (B) Conditional entropy as a function of set size in data from healthy controls in Collins, Brown, Gold, Waltz, & Frank, 2014. Error bars show the standard error of the mean. (Lai & Gershman, 2021)

Further supporting this theory, experiments by Collins and Frank, 2012 using a contextual multi-armed bandit task (Fig. 9a) demonstrate that increasing cognitive load, thereby reducing policy complexity, results in more stochastic action selection. They observed that performance declined with an increase in the number of states or set size. This is consistent with the prediction that larger set sizes, which distribute policy complexity across more states, lead to greater stochasticity in action choices, as confirmed by the conditional entropy $H(A|S)$ measure (Fig. 9b).

4_2. Response time

Hick's law, established by Hick, 1952, is a fundamental concept in understanding response times in relation to the number of targets or choices. It states that the mean response time increases logarithmically with the number of targets. This law applies not only to target selection tasks but also to contextual multi armed bandit tasks, as demonstrated by Collins and others (Collins, 2018; Collins et al., 2014; Collins & Frank, 2012), with mean response time being a linear function of log set size (Fig. 10). Further analysis by McDougle & Collins, 2021 also supports this.

Hick's law is derived from information theory. In scenarios where each state is equally likely, the response time correlates with the number of bits required to decode the state, leading to the logarithmic relationship described in Hick's law. This relationship also holds when the number of states

is constant but their probability distribution changes, as indicated by studies like the one by Hyman, 1953. In line with this, in a work by Collins, 2018, it was found that policy complexity correlates with longer response times ($r = 0.35$, $P < 0.0001$).

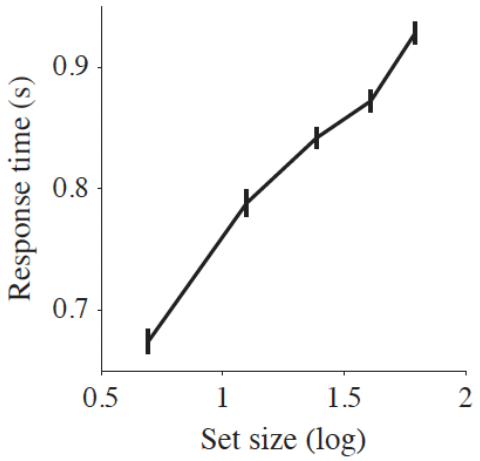


Figure 10. Mean Response Duration Relative to Logarithmic Set Size. This graph presents data from healthy participants in the study by Collins et al., 2014, who engaged in the stimulus-response activity illustrated in Fig.9 . The error bars represent the standard error of the mean (Lai & Gershman, 2021).

Proctor & Schneider, 2018, noted that Hick's law mainly applies when policies are retrieved from memory. For instance, response times increase with set size in tasks where targets are symbolically indicated but not when the target location corresponds directly to the stimulus location (Dassonville, Lewis, Foster, & Ashe, 1999). This supports the notion that memory demands in policy retrieval act as a bottleneck in action selection.

4_3. Navigation:

Navigation, as an essential aspect of goal-directed behavior, involves planning intricate action sequences to reach a desired location. In animal learning studies, tasks like the Morris Water Maze are employed to assess navigational skills, with performance traditionally measured by the time to reach a goal. However, this approach doesn't account for the complexity of the chosen route, which is critical as it pertains to how specific to a given state or location the trajectory is (Lai & Gershman, 2021).

Consider a practical scenario where you have to choose between two grocery stores: one nearby with several turns (indicative of high policy complexity) and the other farther but accessible via a straightforward highway route (demonstrating low policy complexity). This decision-making process underscores that the complexity of navigational policies, influenced by the specific requirements of each route, involves significant computational considerations.

Amir et al., 2020, explored this concept through the lens of the reward-complexity tradeoff. They analyzed the learning behavior of mice in the Morris Water Maze, focusing on the balance between the energy and time cost (value) of swimming trajectories and their complexity, which represents the computational effort needed for specific, goal-directed actions. They observed that trained mice developed more efficient, albeit complex, swimming paths, reflecting an increased awareness of their immediate surroundings to navigate effectively toward the goal.

Over the training period, the mice initially focused on maximizing the value of their routes, later shifting towards reducing the complexity of their paths. This shift indicated a progression along the optimal reward-complexity curve. Amir et al., 2020, quantified this learning process by fitting values of β to the daily data, noting a significant increase over the four days. The study concluded that as the mice became more familiar with the goal's location, their motor command precision improved, enabling them to navigate swiftly toward the platform from various starting points. This demonstrated a dynamic

balance in navigational learning, where mice fine-tuned their movements to optimize both the efficiency and complexity of their paths.

5. Discussion

In this review, we explored empirical support for the reward-complexity balance concept. Initially, we present a conceptual model that encapsulates the notion of policy compression – essentially, decreasing the cognitive cost by simplifying action policies. Next, we explored how this model helps to cohesively understand various behavioral patterns such as randomness in choices, reaction times, grouping actions, and navigational strategies.

We further developed models inspired by RDT and applied them to the previously analyzed dataset. Our findings revealed that our modified model, Q_RDT, provided a better fit to the data than the existing models on the same dataset. Given that deficits in working memory and cognitive effort are well-documented characteristics of schizophrenia (Culbreth, Moran, & Barch, 2018), it is logical to assume that these patients would display notable inefficiencies in their reward-complexity trade-off. Our proposed model effectively captured this phenomenon, indicating that schizophrenia patients had significantly lower inverse temperature parameters. This suggests a higher cognitive cost and reduced policy complexity in these patients.

The concept that various cognitive aspects are governed by resource-limited rationality has become increasingly accepted (Gershman et al., 2015; Lieder & Griffiths, 2019). Understanding the specific nature and impact of these resource limitations remains an active area of study. Our work adds to this discussion by utilizing rate distortion theory to address a key question in psychology: the reasons behind human and animal perseveration. Our explanation posits that perseveration emerges due to constraints on policy complexity. With a finite number of bits for policy encoding, a resource-efficient agent will naturally show perseveration. Our empirical evidence from two datasets demonstrates that varying policy complexity among subjects or in different task conditions leads to expected levels of reward based on the optimal reward-complexity balance. Our findings also align the behavioral pattern of perseveration with predictions from rate distortion theory, although we observed a consistent deviation from the ideal trade-off function among subjects with lower policy complexity.

The theoretical basis of rate distortion theory is notably abstract, making minimal assumptions about the cognitive processes underlying various points on the reward-complexity curve. This approach significantly differs from previous models applied to the similar datasets (Collins, 2018; Steyvers et al., 2019), which investigated detailed mechanistic hypotheses. Both methods have their unique strengths and weaknesses. While the detailed mechanistic theories of cognition, like those developed by Collins, Steyvers, and others, are valuable, exploring general principles at a more abstract level, as done in this work, offers the advantage of making broad claims about cognitive nature that are not confined to specific mechanisms (Gershman, 2020).

Rate distortion theory offers substantial potential as a unifying framework because it merges two theories—information theory and statistical decision theory—each with their extensive explanatory capabilities. Its application has already proven effective in various cognitive domains, including working memory (Sims, 2016; Sims et al., 2012), absolute identification (Sims, 2018), language (Zaslavsky et al., 2018), and motor control (Schach et al., 2018). Ultimately, a comprehensive theory in these areas will likely involve mechanistic models to refine and inform the rate distortion analysis.

References

Amir, N., Suliman-Lavie, R., Tal, M., Shifman, S., Tishby, N., & Nelken, I. (2020). Value-complexity tradeoff explains mouse navigational learning. *PLOS Computational Biology*, *16*(12). doi:10.1371/journal.pcbi.1008497

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Transactions on Information Theory, 18, 14–20.

Bari, B. A., & Gershman, S. J. (2022). Undermatching is a consequence of policy compression. The Journal of Neuroscience, 43(3), 447–457. doi:10.1523/jneurosci.1003-22.2022

Berger, T. (1971). Rate distortion theory: A mathematical basis for data compression. Prentice-Hall.

Bhui R, Lai L, Gershman S. J. (2021). Resource-rational decision making. Current Opinion in Behavioral Sciences.

Bhui, R., & Gershman, S. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. Psychological Review, 125, 985–1001.

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

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

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

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

Culbreth, A. J., Moran, E. K., & Barch, D. M. (2018). Effort-based decision-making in schizophrenia. Current Opinion in Behavioral Sciences, 22, 1–6.

Dassonville, P., Lewis, S. M., Foster, H. E., & Ashe, J. (1999). Choice and stimulus–response compatibility affect duration of response selection. Cognitive Brain Research, 7, 235–240.

Fox, R., Pakman, A., & Tishby, N. (2016). Taming the noise in reinforcement learning via soft updates. In Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence

Hick, W. E. (1952). On the rate of gain of information. Quarterly Journal of Experimental Psychology, 4, 11–26.

Grau-Moya, J., Leibfried, F., & Vrancx, P. (2018). Soft q-learning with mutual-information regularization. In International conference on learning representations.

Hale, D. (1968). The relation of correct and error responses in a serial choice reaction task. Psychonomic Science, 13, 299–300.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. Journal of Experimental Psychology, 45, 188.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. Proceedings of the IRE, 40, 1098–1101

Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Competition and cooperation between multiple reinforcement learning systems. In Goal-directed decision making(pp. 153–178). Elsevier.

Konda, V. R., & Tsitsiklis, J. N. (2000). Actor-critic algorithms. In Advances in Neural Information Processing Systems (pp. 1008–1014).

Matejka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. American Economic Review, 105, 272–298.

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. Journal of Political Economy, 59, 371–404.

McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. Psychonomic Bulletin & Review, 28, 20–39

Lai L., & Gershman S. J. (2021). Policy compression: An information bottleneck in action selection. In: Federmeier KD, editor. The Psychology of Learning and Motivation. vol. 74 of Psychology of Learning and Motivation. Academic Press;. p. 195–232.

Lai, L., & Gershman, S. J. (2023). Human Decision Making Balances Reward Maximization and Policy Compression. doi:10.31234/osf.io/rnz72

Lai, L., Huang, A. Z., & Gershman, S. J. (2022). *Action Chunking as Policy Compression*. doi:10.31234/osf.io/z8yrv

Lerch, R. A., & Sims, C. R. (2018). Policy generalization in capacity-limited reinforcement learning. OpenReview.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. Nature Neuroscience, 17(3), 347–356.

Lieder F, & Griffiths TL.(2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behav Brain Sci.

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. Cognition, 204, 104394

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

Gershman S. J., Lai L. (2021). The Reward-Complexity Trade-off in Schizophrenia. Computational Psychiatry. *2021;5(1):38–53.*

Gigerenzer G, & Selten (2002). Bounded Rationality: The Adaptive Toolbox. MIT Press.

Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. Theory in Biosciences, 131, 139–148.

Parush, N., Tishby, N., & Bergman, H. (2011). Dopaminergic balance between reward maximization and policy complexity. Frontiers in Systems Neuroscience, 5.

Schach, S., Gottwald, S., & Braun, D. A. (2018). Quantifying motor task performance by bounded rational decision theory. Frontiers in Neuroscience, 12.

Steyvers, M., Hawkins, G. E., Karayanidis, F., & Brown, S. D. (2019). A large-scale analysis of task switching practice effects across the lifespan. Proceedings of the National Academy of Sciences, 116, 17735–17740

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

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of information. University of Illinois Press.

Sims, C., Jacobs, R., & Knill, D. (2012). An ideal observer analysis of visual working memory. Psychological Review, 119, 807–830.

Sims, C. R. (2016). Rate-distortion theory and human perception. Cognition, 152, 181–198.

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

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. Institute of Radio Engineers, National Convention Record, 4, 142–163.

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

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. Current Opinion in Neurobiology, 55, 7–14.

Rescorla R. A. , & Wagner A. R. (1972.). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning: current research and theory. vol. 2. Appleton-Century-Crofts.

Rubinstein A (1998). Modeling bounded rationality. MIT press;

Proctor, R. W., & Schneider, D. W. (2018). Hick's law for choice reaction time: A review. Quarterly Journal of Experimental Psychology, 71, 1281–1299.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. The Annals of Mathematical Statistics, (pp. 400–407).

Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. In Perception-action cycle. Springer, pp. 601–636.

Von Neumann, J. (1958). The computer and the brain. Yale University Press.

Wit, E., Heuvel, E. van, & Romeijn, J. (2012). 'all models are wrong...': An introduction to model uncertainty. Statistica Neerlandica, 66(3), 217–236.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. Proceedings of the National Academy of Sciences, 115, 7937–7942.