Paper report: Memory shapes time perception and intertemporal choices

Thomas Michel

March 25, 2024

Contents

1 Introduction

According to the experiences of many, time does not seem to flow at a constant rate. Instead, time appears to slow down when engaging in a new activity and speed up as we become accustomed to it. Both human and non-human subjects experience temporal distortions in their perception and decision-making processes. Experiments indicate that when presented with a sequence of identical stimuli, subjects tend to perceive the duration between stimuli as decreasing. Conversely, when stimuli are different, new stimuli are perceived as lasting longer. This observation led to the hypothesis that the perception of time is linked to the amount of neural energy used to represent stimuli, with this energy being associated with coding efficiency Eagleman and Pariyadath [2009].

Another implication of temporal distortions in perception pertains to intertemporal choices and decision-making. Intertemporal choices involve trade-offs between costs and benefits occurring at different points in time. For instance, when deciding between receiving a reward now or later, individuals must evaluate the value of the reward and the associated time delay. The subjective value of the reward decreases as the delay increases, a phenomenon known as temporal discounting. Temporal discounting is a well-established cognitive bias affecting decision-making across various domains, including economics, psychology, and neuroscience. Over the years, multiple models of temporal discounting have been developed. The most common model is exponential discounting, which posits that the value of a reward decreases at a constant rate over time. Another significant

model is hyperbolic discounting, which suggests that the value of a reward decreases rapidly at first and then more slowly, with the rate of decrease diminishing over time. Empirical evidence indicates that the hyperbolic discounting model better captures the behavior of human subjects in intertemporal choices compared to the exponential discounting model [Laibson, 1997, Bradford et al., 2019].

While existing models can capture human behavior in intertemporal choices, they fail to explain the underlying reasons for these discounting effects or the distortion of time perception. Some research has provided explanations for the hyperbolic discounting model. For instance, Read [2001] suggested that hyperbolic discounting arises from a sub-additive perception of durations. Other studies propose that time perception can be directly modeled based on the effects of classical psychophysical laws Takahashi et al. [2008].

In this report, we will delve into the work of Ortega and Tishby [2016], who introduced a novel model of temporal discounting. This model posits that memory shapes both time perception and intertemporal choices and is grounded in information theory, building upon the hypothesis proposed by Eagleman and Pariyadath [2009]. According to this model, the perception of time is linked to the amount of information stored in memory, particularly the information generated by presented stimuli given the memory of past stimuli. The authors demonstrate that this model aligns with human behavior observed in previous studies on intertemporal choices. Crucially, the model is agnostic to the specific neural mechanisms underlying time perception or the cognitive processes implementing it, such as an internal clock model.

The report will proceed as follows: In the first section, we will present the model of time perception proposed by Ortega and Tishby [2016] and highlight the main findings of their study. Subsequently, in the second section, we will describe the implications of this model for modeling intertemporal choices and how it shapes the perceived values of future gains. Finally, we will discuss some particularity of the model and propose an extension to the experiments conducted in the article to investigate how the agent's policy and objective could influence time perception.

2 Modeling time perception

2.1 General framework

The model proposed by Ortega and Tishby [2016] operates within the general framework of an adaptive agent sequentially interacting with its environment. At each time step t, the agent receives a stimulus s_t from the environment and generates an action a_t in response. Subsequently, the environment provides a feedback signal r_t to the agent. This feedback is stochastically generated according to a probability distribution $\mathbb{P}(r|s_t, a_t)$. The goals of the agent can vary, ranging from maximizing the signal (interpreted as a reward) to maintaining homeostatic equilibrium. Specifically, the authors focus on the scenario where the agent's objective is to maximize the signal (referred to as reward), although their framework can accommodate other objectives.

Additionally, the agent may possess memory of past stimuli and actions, which can inform its current action selection. The memory is updated at each time step t according to a memory update rule $m_t = f(m_{t-1}, s_{t-1}, a_{t-1})$. This memory can be leveraged to generate the current action $a_t = g(m_t, s_t)$. Beyond storing information about the past, such as the history of stimuli and actions, the memory can also represent the agent's belief about the dynamics of the environment and, by extension, future rewards. This last perspective is the one often adopted in the context of Bayesian reinforcement learning.

The authors posit that the memory state serves as the minimal sufficient statistic used by the model to generate the current action. Their approach to modeling memory is reminiscent of the binary code word length associated with a finite sequence of stimuli, building upon ideas introduced by Bialek et al. [2001]. They partition the agent's interactions into two components: x_p, representing past stimuli and actions, and x_f, representing future interactions to predict. The amount of information utilized by the model to enhance prediction of the future is quantified by the difference in code word length between the prior distribution on the future $\mathbb{P}(x_f)$ and the posterior distribution $\mathbb{P}(x_f|x_p)$:

$$-\log P\left(x_{\mathrm{f}}\right)-\left(-\log P\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right)\right)=\log \frac{P\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right)}{P\left(x_{\mathrm{f}}\right)} \tag{1}$$

By averaging over the possible pasts and futures, we obtain the mutual information between the past and the future, which is also called the predictive information:

$$I\left(X_{\mathrm{p}} ; X_{\mathrm{f}}\right)=E_P\left[\log \frac{P\left(X_{\mathrm{f}} \mid X_{\mathrm{p}}\right)}{P\left(X_{\mathrm{p}}\right)}\right]. \tag{2}$$

This predictive information can be seen as the amount of information that the agent needs to store in memory to predict the future.

2.2 Results

The authors conducted a series of experiments to validate their model's predictions. They employed a simple task where the agent must choose between two options, each associated with a different reward. The rewards are probabilistic and depend on the agent's choice. Formally, this agent-environment model is referred to as a multi-armed bandit problem with Bernoulli rewards. Specifically, for all experiments, a bandit with two arms (actions) was considered. The actions are denoted as a and b, and their rewards are sampled from Bernoulli distributions with parameters $\frac{1}{4}$ and $\frac{3}{4}$, respectively. This model is simpler than the general framework presented previously as it is stateless and stationary, meaning rewards depend only on the agent's action and not on the history of past actions, and the environment does not change over time. This simplification enables the authors to focus on the effect of memory on time perception and intertemporal choices. The objective is to maximize the total reward obtained by the agent over a fixed number of steps. Despite its simplicity, this model is powerful as it captures the essence of the trade-off between exploration and exploitation in reinforcement learning. This trade-off refers to the dilemma faced by the agent between exploiting actions that have yielded high rewards in the past and exploring new actions to gather more information about the environment.

In the experiments, the authors consider three types of systems, distinguished by the hypothesis space considered by the agent—the set of environments the agent expects to interact with. In the first model, the agent has perfect knowledge about the environment and can choose the optimal action at each step. In the second model, the agent knows that the environment corresponds to one of two possible environments but does not know which one. The third model considers a more general parametric hypothesis space, where the agent assumes that the environment parameters of the Bernoulli rewards can take any values between 0 and 1. This hypothesis space already encompasses all possible environments the agent can interact with in the context of stationary stochastic Bernoulli bandits. Furthermore, the authors extend their investigation to a non-parametric infinite-dimensional hypothesis space with a more general prior distribution on the environment,

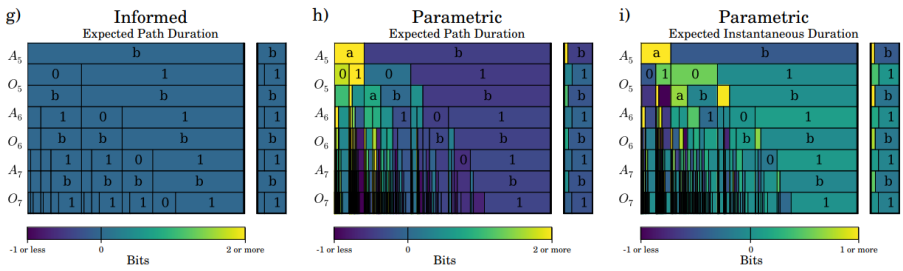

Figure 1: Expected duration conditioned by the past sequence of events $x_p = (a1, a0, b1, b1)$. Figure from Ortega and Tishby [2016].

where models consist of sequences of bandits with possibly different reward distributions. While this last model does not reflect the reality of the environment, it serves as a strict generalization of the previous space, in the sense that for any model from the previous space, there exists a model in the non-parametric space with the same reward distributions.

The decision-making process of the agent is based on the Thompson sampling algorithm, a Bayesian algorithm that iteratively samples the environment parameters from the posterior distribution and selects the action that maximizes the expected reward for the sampled environment. Known to be optimal in achieving the best possible expected total reward in the multi-armed bandit problem, this algorithm is particularly well-suited to the context of the experiments due to its versatility and applicability to various hypothesis spaces.

Within the framework outlined above, the next step is to define the perceived time in relation to memory. The authors propose defining the present as the minimal sufficient statistic of the past. This present represents the information that the agent needs to store in memory to predict the future or adequately recall the past. Formally, the information possessed by the agent about the future is quantified by the log-likelihood ratio of the prior and posterior distribution of future interactions:

$$\text{Present}(x_f) = \log \frac{\mathbb{P}(x_f|x_p)}{\mathbb{P}(x_f)} \tag{3}$$

In this context, the passage of time is simply defined as the change in the number of bits in memory resulting from the current interaction. If x_n represents the signal received by the agent in the current time window, the duration of x_n is the difference between the amount of information possessed by the agent before the interaction and the amount possessed after the interaction.

$$\begin{aligned}
\text{Duration}(x_n) &= \text{Present}(x_f) - \text{Present}(x_f, x_n) \\
&= \log \frac{\mathbb{P}(x_f|x_n, x_p)}{\mathbb{P}(x_f)} - \log \frac{\mathbb{P}(x_f, x_n|x_p)}{\mathbb{P}(x_f, x_n)} \\
&= \log \frac{\mathbb{P}(x_f|x_n, x_p)}{\mathbb{P}(x_f)} - \log \frac{\mathbb{P}(x_f|x_n, x_p)\mathbb{P}(x_n|x_p)}{\mathbb{P}(x_n|x_f)\mathbb{P}(x_f)} \\
&= \log \frac{\mathbb{P}(x_n|x_f)}{\mathbb{P}(x_n|x_p)}
\end{aligned} \tag{4}$$

The experiments yield the expected outcomes. With full knowledge of the environment, the informed agent does not need to store information in its memory, resulting in a constant information

about the future. Consequently, the duration of the interaction is null, and the agent does not experience time.

In contrast, the parametric agent perceives durations that vary depending on the interaction and its understanding of the environment. For example, when the agent repeats the same action and receives the same outcome, the perceived duration of the interaction is close to zero, indicating that the agent is not learning anything new, as illustrated by the trajectory b1b1b1 (Figure 1i). Rare events are perceived as lasting longer because the agent is learning more about the environment and the future, while the repetition of an action for which the outcome is presumed to be well known is associated with small or even negative durations.

3 Decision-making and temporal discounting

The stochastic process associated with the agent-environment couple is influenced by the agent's incentive to gravitate towards events with higher rewards. The authors recognize this transformation of the process as a consequence of maximizing the expected reward while considering the memory capacity of the agent. Formally, this entails maximizing, with respect to the posterior distribution $\tilde{P}$, the free energy functional given by:

$$F\left(x_{\mathrm{p}}\right)[\tilde{P}] := \begin{array}{ll} \sum_{x_{\mathrm{f}}} \tilde{P}\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right)\left[R\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right)+F\left(x_{\mathrm{p}}, x_{\mathrm{f}}\right)\right] & \text{(Expected Rewards)} \\ -\frac{1}{\beta} \sum_{x_{\mathrm{f}}} \tilde{P}\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right) \log \frac{\tilde{P}\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right)}{P\left(x_{\mathrm{f}}\right)} & \text{(KL-Divergence)} \end{array}, \tag{5}$$

where β is the inverse temperature parameter that defines the trade-off between reward maximization and the memory cost of changing the probability of x_f, and $R(x_f|x_p)$ represents the reward associated with the future sequence x_f given the past sequence x_p.

If the probability $P(x_f|x_p)$ is the result of optimizing 5 then the information about the future predicted by the present can be expressed as

$$\log \frac{P\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right)}{P\left(x_{\mathrm{f}}\right)}=\beta\left[R\left(x_{\mathrm{f}} \mid x_{\mathrm{p}}\right)+F\left(x_{\mathrm{p}}, x_{\mathrm{f}}\right)-F\left(x_{\mathrm{p}}\right)\right] \tag{6}$$

which is proportional to the difference between the potential reward for the future x_f and the expected reward for all possible futures given the past x_p. In the context of decision theory, this difference is referred to as the *rejoice*. It's worth noting how these results establish a direct relationship between the present scope and the rejoice, suggesting an alternative formulation of the duration defined in the previous section as the difference in rejoice between two instants.

	$\delta(\tau)$	Growth	Type
Informed	0	0	Infinite
Finite	$\max(0, 0.2274 - 0.1131\tau)$	$\mathcal{O}(1)$	Linear
Parametric	$0.2701e^{-0.6543\tau}$	$\mathcal{O}\left(e^{-c\tau}\right)$	Exponential
Nonparametric	$0.2314\tau^{-0.5805}$	$\mathcal{O}\left(\tau^{-c}\right)$	Hyperbolic

Table 1: Discount functions for different hypothesis spaces. Results from Ortega and Tishby [2016].

After establishing a connection between memory, perceived duration, and predicted reward, the authors propose revisiting the classical concept of reward discounting to align reweighted rewards with subjective values. This is achieved by computing the actual reward optimized by the stochastic

process defined by the interaction between agents and bandits, and then fitting a function to explain the perceived future rewards. Table 1 displays the results obtained by [Ortega and Tishby, 2016] for the specific bandit environment described earlier. Notably, each class of hypothesis allows for the recovery of a different discount function. The informed agent, possessing perfect knowledge of the environment, does not discount future rewards. The finite agent discounts future rewards linearly, while the parametric agent discounts exponentially, and the non-parametric agent discounts hyperbolically. These findings align with the notion that the perception of time is influenced by the memory capacity of the agent and the hypothesis space considered. Moreover, these discount functions are commonly used in the literature to model agents' intertemporal choices and are supported by empirical evidence.

4 Remarks and extensions

4.1 Interpreting duration

In the authors' model, duration is defined as the difference in the information stored in memory between two instants, relative to particular past and future sequences of events (see Eq. 4). This definition assumes that the computational model has a fixed bandwidth, meaning the amount of information that can be modified in memory at each time step is bounded. This assumption appears reasonable within the context of the authors' experiments and biological systems in general.

However, this definition of duration yields some peculiar results. For instance, if we consider any finite sequence of events and compute the instantaneous duration of each event, the sum of these durations is zero, as illustrated in Figure 2. This can be demonstrated from Eq. 4 by decomposing each sequence and identifying terms that cancel out. This result is counter-intuitive, as it implies negative durations for some events. The authors have already identified these negative durations in a broader context and attribute them to the presence of oddball events, which contradict the memory state of the agent. The possibility of experiencing negative durations challenges the conventional understanding of perceived time. This phenomenon warrants further investigation to determine if it truly relates to the perception of time or if it is merely an artifact of the model, which would diminish the utility of the model.

However, it's important to note that the phenomenon mentioned at the beginning of the last paragraph is of a different nature from the one mentioned by the authors. It arises from considering only a finite sequence of events and retrospectively computing the duration of each event, which does not seem to be the approach taken by the authors in their experiments but would be most intuitive based on the definition. One interpretation for this last observation is that as the sequence of future events shrinks, it becomes less relevant to the model than past events. This is consistent with the idea that the present is the minimal sufficient statistic of the past. This is may not the way the authors make use of the proposed definition of duration, but it is well visible in Figure 2, so it seemed important to mention.

4.2 Influence of decision-making on perceived time

The authors demonstrate how the perceived duration of an event is related to the considered hypothesis space. Throughout the experiments, the authors make use of the Thompson sampling algorithm to make decisions. This algorithm is well suited in the context of the experiments, as it follows a Bayesian framework and allows computing explicitly some of the properties of the

stochastic process. This algorithm is designed to minimize regret (difference in total reward between the choices of the agent and the optimal ones). While it models well the behavior of a rational agent, rational model of decision-making are not always the best model to explain observed human behavior. Recently, ideas of alternative objectives for decision-making have been developed in the sequential learning community, from which the multi-armed bandit problem and Thompson sampling originate. A simple alternative to optimization is the idea of satisficing [Simon, 1956]. This concept is related to the idea of bounded rationality, which is the idea that agents have limited computational resources and cannot always make the best decision. In particular, with satisficing, once the agents is able to secure an average reward above a satisfaction level with a certain confidence, then it will judge that the choice is good enough and drastically reduce the exploration.

As an extension of the studied paper, we propose to look at the difference in perception of agents with different behaviors and objectives. In addition to Thompson sampling, we consider the Upper Confidence Bound (UCB) algorithm [Auer et al., 2002], which is a classical algorithm in the multi-armed bandit literature. The UCB algorithm is based on the idea of balancing exploration and exploitation by choosing the action that maximizes the upper confidence bound of the expected reward. The statistics used to compute the upper confidence bound are similar to the ones maintained by the Thompson sampling algorithm, but the decision-making process is different. On important particularity of the UCB algorithm is that the choice of an action is deterministic given the current state of the agent, while it is dependent on the random sampling of the environment parameters in the Thompson sampling algorithm. Then we also consider a modification of the UCB algorithm named Sat-UCB [Michel et al., 2023] which implement the idea of satisficing and drastically change the exploration strategy. These difference may lead to different perception of time and duration. Finally, we consider a random agent which choose actions uniformly at random.

The authors demonstrate how the perceived duration of an event is influenced by the hypothesis space considered. Throughout the experiments, the authors utilize the Thompson sampling algorithm for decision-making. This algorithm, well-suited for the experiments, operates within a Bayesian framework and explicitly computes some properties of the stochastic process. Designed to minimize regret (the difference in total reward between the agent's choices and the optimal ones), the algorithm effectively models the behavior of a rational agent. However, rational models of decision-making may not always accurately explain observed human behavior. Recently, alternative objectives for decision-making, inspired by the literature in economics, have emerged in the sequential learning community, from which the multi-armed bandit problem and Thompson sampling originate. One such alternative is the concept of satisficing [Simon, 1956], which relates to bounded rationality—the idea that agents have limited computational resources and cannot always make optimal decisions. With satisficing, once an agent secures an average reward above a satisfaction level with a certain confidence, it judges the choice to be satisfactory and significantly reduces exploration.

As an extension of the studied paper, we propose to investigate the difference in perception among agents with varying behaviors and objectives. In addition to Thompson sampling, we consider the Upper Confidence Bound (UCB) algorithm [Auer et al., 2002], a classical algorithm in the multi-armed bandit literature. UCB balances exploration and exploitation by selecting the action that maximizes the upper confidence bound of the expected reward. While the statistics used in computing the upper confidence bound are similar to those in Thompson sampling, the decision-making process differs. Notably, UCB's choice of action is deterministic given the current state of the agent, unlike Thompson sampling, which depends on the random sampling of environment parameters. Furthermore, we consider a modification of UCB called Sat-UCB [Michel et al., 2023], which implements the idea of satisficing and drastically alters the exploration strategy. These

differences may result in varying perceptions of time and duration. Lastly, we consider a random agent that chooses actions uniformly at random.

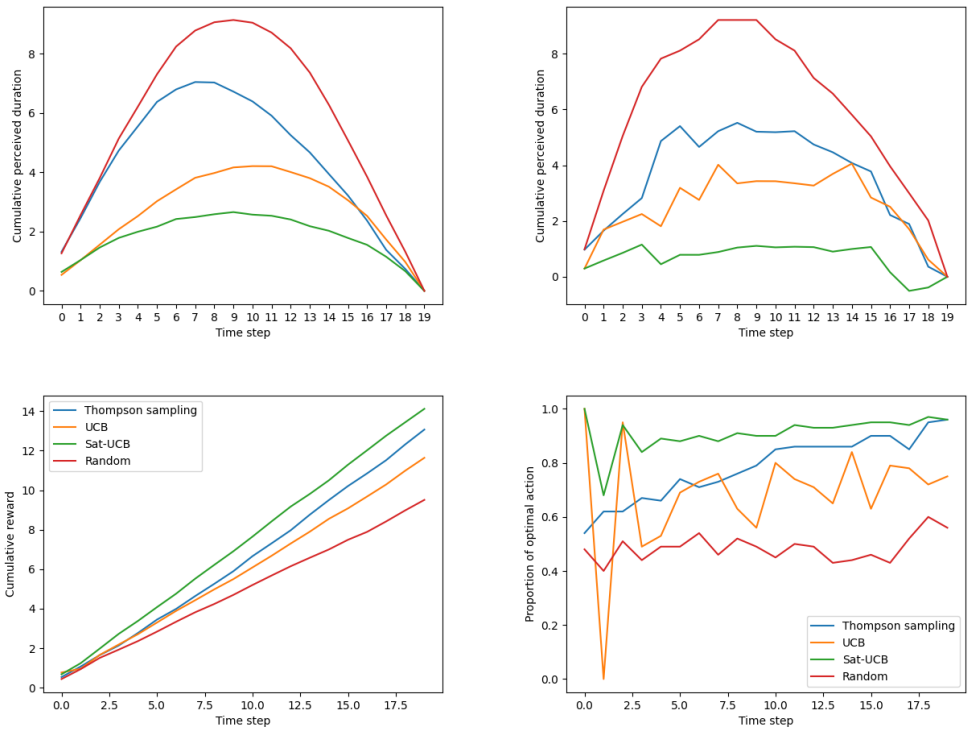

Figure 2: Top left: Average perceived duration of the interaction for different decision-making strategies. Top right: Perceived duration of the interaction for a single run of the algorithm. Bottom left: Average cumulative reward obtained by the agent for different decision-making strategies. Bottom right: Average proportion of optimal actions taken by the agent for different decision-making strategies.

The results of the experiment are depicted in Figure 2. The plots show the average perceived duration of the interaction, the average reward obtained by the agent, and the average proportion of optimal actions taken by the agent. Perceived duration was computed by considering a fixed sequence of 20 actions and rewards experienced by the agents, calculating the duration of each action-reward pair. This corresponds to computing Duration(x_n) defined in Eq. 4 for each element x_n of the sequence. Since deriving the distribution over the stochastic process is challenging when using the UCB algorithm, we approximate probabilities via Monte Carlo simulation. To achieve this, we simulate $10,000$ runs of the algorithm and use them to approximate the probabilities involved in computing the duration. The plots display average durations computed by considering 100 independent runs of the algorithm, which are also independent of those used to compute the probabilities.

The results differ significantly from those proposed by the authors of the article, but we will attempt an interpretation nonetheless. The most notable aspect of the plots is the first half, where we observe that each algorithm perceives time differently, likely due to differences in exploration

strategy. The random strategy exhibits the longest duration, as the agent continually explores the environment and enriches its memory. The UCB algorithm and Thompson sampling are more conservative and explore less, resulting in shorter perceived durations. The Sat-UCB algorithm is highly focused and quickly converges towards playing only the optimal action. This is reflected in the very short perceived duration of the interactions since the future is easily predictable by the agent. Both UCB-based algorithms have shorter durations than the Thompson sampling algorithm and the random sampling algorithm, possibly due to their deterministic nature, which leads to less surprise and shorter perceived durations.

In conclusion, regarding the difference in perception of time among agents with different decision-making strategies, we find that the exploration strategy directly influences the perceived duration of interactions. The more the agent explores, the longer the interactions are perceived, which aligns with the idea that surprising observations leads to longer perceived durations. This result is interesting as it indicates that time perception is not solely linked to the memory capacity of the agent and its hypothesis space but also to the agent's decision-making strategy. However, due to the limitations imposed by the choice of algorithms, it was challenging to delve deeper into the analysis to derive results about intertemporal choices. Additionally, the satisficing framework is not typically studied in terms of reward discounting. Nevertheless, in the specific context of this study, it would have been interesting to explore how this strategy influences the importance attributed to future rewards by the model in the context of satisficing.

5 Conclusion

This study links the perception of time to the memory capacity of an agent interacting with a stochastic environment. The authors propose a model in which the duration of an interaction is defined as the difference in the information stored in memory between two instants. This model builds upon an interpretation of biological models that perform efficient coding of information and perceive time as the rate of change of information stored in memory. The authors demonstrate that the perceived duration of an interaction is associated with the hypothesis space considered by the agent. Our experiments indicate that, within a given hypothesis space, perceived time is also influenced by the decision strategy and the agent's objectives.

The model proposed by the authors presents a simple yet robust framework for studying the perception of time. It can relate time perception to the memory capacity of the agent and general phenomena such as memory plasticity, which correlates with increased perceived durations. The model is theoretically powerful as it connects the approaches that try to model the behaviors of agents to a unified model independent of the biological implementation of memory, and it can recover known results regarding temporal discounting, for instance. However, certain aspects of the model require further clarification, such as the occurrence of negative durations in cases where memory is not coherent with the environment. This phenomenon is counter-intuitive and warrants further investigation to determine if it is merely an artifact of the model or a genuine phenomenon observable in biological systems.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

William Bialek, Ilya Nemenman, and Naftali Tishby. Complexity through nonextensivity. *Physica A: Statistical Mechanics and its Applications*, 302(1-4):89–99, 2001.

W David Bradford, Paul Dolan, and Matteo M Galizzi. Looking ahead: Subjective time perception and individual discounting. *Journal of Risk and Uncertainty*, 58:43–69, 2019.

David M Eagleman and Vani Pariyadath. Is subjective duration a signature of coding efficiency? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1525):1841–1851, 2009.

David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112 (2):443–478, 1997.

Thomas Michel, Hossein Hajiabolhassan, and Ronald Ortner. Regret bounds for satisficing in multi-armed bandit problems. *Transactions on Machine Learning Research*, 2023.

Pedro A Ortega and Naftali Tishby. Memory shapes time perception and intertemporal choices. *arXiv preprint arXiv:1604.05129*, 2016.

Daniel Read. Is time-discounting hyperbolic or subadditive? *Journal of risk and uncertainty*, 23: 5–32, 2001.

Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63 (2), 1956.

Taiki Takahashi, Hidemi Oono, and Mark HB Radford. Psychophysics of time perception and intertemporal choice models. *Physica A: Statistical Mechanics and its Applications*, 387(8-9): 2066–2074, 2008.