Graduate Training Centre of Neuroscience
Neural & Behavioral Sciences
University of Tübingen

Modeling the Role of Serotonin in Patience

Lab Rotation - I

Report submitted by:
Shweta Prasad

Study supervised by:
Dr. Kevin Lloyd
Computational Neuroscience Lab,
Max Planck Institute for Biological Cybernetics

Duration of the lab rotation: Sept. 9, 2023 - Nov. 24, 2023
Report submitted on Dec. 12, 2023

Abstract

Using optogenetic activation of serotonin (5-HT) neurons in the dorsal raphe nucleus (DRN) of mice performing a sequential tone-food task, Miyazaki et. al [1] were able to show that 5-HT promotes waiting longer for delayed rewards, when the probability of reward delivery and the uncertainty in delivery time were both high. In this study, we employ an average reward reinforcement learning model to capture the modulatory effects of 5-HT in the task. Here, it is proposed that 5-HT slows down the agent's subjective perception of time thereby making it more patient, i.e, willing to wait for longer intervals of time before quitting.

Acknowledgements

I would like to extend my deepest gratitude to Dr. Kevin Lloyd for his support and supervision on this lab rotation project; To all the members of the Computational Neuroscience Lab for their enthusiastic camaraderie and discussion; To Pawel Pierzchlewicz, whose kind correspondence and master thesis this exploratory project is built upon; To the Graduate Training Center of Neuroscience for the opportunity to conduct this study; and to my family and friends here for their quiet, constant encouragement.

Contents

1 Introduction **2**
1.1 Serotonin . . . 2
1.1.1 Where is it produced? What does it do? . . . 2
1.1.2 Serotonin in impulsivity/patience . . . 3
1.2 Interval timing . . . 4
1.2.1 Scalar timing theory . . . 5
1.3 Conclusion . . . 5

2 Model background **7**
2.1 Body of experimental evidence . . . 7
2.1.1 Paradigm: Sequential tone-food waiting task . . . 7
2.1.2 Experiments and key findings . . . 8
2.2 Bayesian sequential/repeated decisions model of 5-HT action . . . 11
2.2.1 Model description . . . 11
2.2.2 Predictions . . . 15
2.2.3 Limitations . . . 17

3 Model **19**
3.1 Baseline agent . . . 19
3.1.1 Markov decision process for the baseline agent . . . 20
3.1.2 Results . . . 23
3.1.3 Limitations . . . 26
3.2 Internal clock model of 5HT action . . . 27
3.2.1 Modulating subjective clock speed . . . 27
3.2.2 Results . . . 30
3.2.3 Limitations . . . 31

4 Discussion **32**

Introduction

In this chapter, we present a short sketch of what serotonin is, where it is produced and what some of its identified roles as a neuromodulator are. Then, we hone into ideas about how it may be contributing to the modulation of patience and waiting. Finally, we will briefly touch upon some important concepts related to perception of time intervals, as it is necessary to form the scaffolding for the model at the center of this report.

1.1 Serotonin

Serotonin is a neuromodulator belonging to the monoamine family of neurotransmitters. The IUPAC name of this indoleamine is 3-(2-aminoethyl)-1H-indol-5-ol, more commonly called 5-hydroxytryptamine or 5-HT. It influences and participates in a wide array of biological functions, such as reward signalling, affect, learning, sleep, stress, appetite and other motivation related functions [2, 3, 4]. The consequence of its broad range of effects to someone interested in studying its function as a neuromodulatory substance is that it is immensely difficult to disentangle the different levels at which it may act on the same behavioral read out. This leaves us with often contradictory findings, very difficult to reconcile into one neat theory of its function.

1.1.1 Where is it produced? What does it do?

Serotonin is an L-tryptophan metabolite. Tryptophan is an essential amino-acid, which means it is not naturally produced in the human body in sufficient amounts and thus needs to be supplemented by ingestion. Foods rich in tryptophan include meats, peanuts, seeds, egg whites, fish and milk. Tryptophan depletion has been shown to induce depressive symptoms in humans and rats [5], and consequently, serotonin has been linked to mood disorders, including symptoms of depression. Selective Serotonin Re-uptake Inhibitors (SSRIs) are often prescribed to counter-act these symptoms at the synaptic level.

In humans and mice, the raphe nuclei in the brainstem are the major producers of serotonin. The dorsal raphe nucleus (DRN) is the most prominent serotonergic structure in the brain, accounting for almost a third of all 5-HT neurons in the brain; however, only roughly half of the DRN neurons are serotonergic [6], and the rest have been shown to [co]produce glutamate, GABA, dopamine, nitric oxide and a variety of peptides – some on the same neurons. [4, 6]. This is oft cited as at the root of the contradictory findings on its function/s.

The DRN also sends very heterogeneous projections to different parts of the brain, such as the basal ganglia, different parts of the prefrontal cortex and the motor cortex, further implicating it in value representation and learning [7], motor learning [8] and thus, also behavioral inhibition [9]), cognitive flexibility [10], regulation of social interaction and control of levels of arousal and mood [11, 12]; The overwhelmingly large number of items on this non-exhaustive list stands to support that we may be very far away from pinning down the exact function of serotonin.

1.1.2 Serotonin in impulsivity/patience

While not discussed in any depth here, serotonin's implication in behavioral inhibition and motor learning leaves it well juxtaposed to possibly explain impulsive decision making in people with attention deficit disorders [9, 13]; On the other hand, recent studies such as [14] probing the effects of boosting serotonin have reported that it can reduce subjective cognitive costs and thereby increase information gathering, which can also be interpreted, depending on the task at hand, as a metric of patience.

Studies have shown that over the course of learning an association, serotonin neurons move from firing phasically at the time of reward delivery, to a ramping-phasic firing pattern [15]; Over the course of trials, it begins showing activity at the stimulus onset, ramping up towards, and phasically firing at reward delivery. When reward is omitted, the ramping activity falls off. This suggests that 5-HT may play a supporting role in waiting for rewards, and especially the ramping activity indicates when within a trial its influence may be strongest. [16]

In lieu of these findings, Doya and Miyazaki, et. al embarked on a series of experiments [9, 17, 1, 18], to explore how 5-HT may be modulating patience, and what it may be reporting during its period of activity. These studies are discussed in more depth in section 2.1.

Now, we briefly survey the leading ideas for the function of serotonin with respect to impulsivity and patience. One of the oldest ideas about the role of 5-HT is *behavioral inhibition*, wherein Deakin and Graeff (and later Faulkner) initially propose that serotonin inhibits certain behaviors in the face of punishments [19, 20]. This is usually tested using go/no-go motor tasks. Shortly after,

it was proposed that impulsive decision making can stem from this model of action of serotonin: decreased levels of 5-HT would thus result in poor inhibition of go behaviors, leading to more impulsive actions. Studies have thereafter linked increased aggression and high temporal discounting of rewards to decreased levels of serotonin [21, 19].
Another idea encompassing its interaction with the dopaminergic system is the more recent *dopamine opponency theory*, where serotonin is posited to hold fort in the opposite direction to dopamine; if dopamine encodes reward, serotonin must encode punishment; it suggests that the phasic firing of serotonin reports a so-called punishment prediction error, and the tonic firing of serotonin reports an average reward rate [3, 22]. The average reward reinforcement learning setting [23] is a suitable one to explore this idea. While it is still not conceded that serotonin reports a punishment prediction error, there is consensus that serotonin reports negative and positive rewards.
A more general theory of serotonin's function is *beneficialness*, expressed in terms of the learnt expected reward and a discounting cost that accumulates [2]. The probability of being rewarded is learnt across the course of trial and error, and the costs can be determined by the animals internal states, such as satiety, arousal, patience, motivation and so on, which interact with each other in multiple ways. Here, the key idea is that serotonin reports the learnt overall beneficialness of the current context, and changes (inflates or deflates) the probability of being rewarded in that context; This can thereby make the animal want to wait for longer (or in our vernacular, more *patient*, and less prone to impulsive switching) within a beneficial context.

1.2 Interval timing

The brain must make decisions spanning many orders of magnitudes of time. From instantaneous decisions at the channel level to action potentials in the sub-millisecond and millisecond ranges, to the circadian rhythm in the range of days, to hormone cycling in the range of months; In this section, we look at the order of magnitude of decision making in behavioural tasks, which can be classified as interval timing tasks. Decisions are made before (prospective timing), after (retrospective timing) or at (immediate timing) the estimation of an interval of time. And to do such discrete event timing tasks, the animals/agents need to be able to perform sufficiently accurate interval estimation.

How does they do it? The striatum of the basal ganglia have been shown to play a role in interval timing [24]. Some have suggested that it acts together with the cerebellum, wherein theh cerebellum times discrete events and cognitive control is effected by means of the basal ganglia, providing a channel for attention to modulate perception of time [25, 26]. Considering their projections to the prefrontal and parietal cortices, studies reporting their activation in interval timing tasks [27] give weight to these ideas.

The *pacemaker-accumulator* models of interval timing thus try to weave together the different processes implicated in the previously mentioned studies – namely, attentional, discrete event timing and decision making processes. A typical pacemaker-accumulator model of interval timing (PAM) can be thought to be comprised of three distinct stages. [28].

- At *the clock stage*, a pacemaker sets the time: it generates ticks, which are accumulated in a manner controlled by attention. Already, there arise two different sources of variability in timing. The speed of this clock is thought to be responsible for production of subjective time perception.
- At *the memory stage*, the working memory is involved. A copy of the current state of the accumulator is passed onto the working memory, where it is compared with a previously stored reference state. This stage also becomes a possible source of variability in estimation, as the working memory's representation of the timed duration (both reference and freshly sampled) need not be veridical.
- At *the decision stage*, a decision is made in accordance with the output of the comparator. Do the intervals match up? Is the newly sampled copy faster or slower? and so on.

1.2.1 Scalar timing theory

Proposed by John Gibbon in 1977, the Scalar Expectancy Theory (SET) is a long standing model of interval timing, which can explain the observation that perception of intervals of time follows a sort of Weber-Fechner law [29, 28, 26]; The noise in the estimate of an interval of time scales with the size of the interval, and therefore, the precision of timing is relative to the interval being timed. This is why it is known as *scalar timing*.

To be more specific, the theory says that the variance (σ^2) in time estimates scales with the square of the mean (μ), and

$$\frac{\sigma}{\mu} = \kappa \tag{1.1}$$

where κ is the scalar timing constant.

The model is built on top of PAM, with a focus on the second stage: Gibbon et. al propose that the nature of the comparison operation is a ratio between the current and expected durations, rather than a difference.

1.3 Conclusion

In this introductory chapter, we discussed the roles of serotonin and its proposed mechanisms of action. Considering the task we will be encountering, we also discussed key considerations about interval estimation and a leading theory in

explaining features found in data related to it. We saw that serotonin can affect impulsivity and patience at different levels, and now we will try to relate serotonin and patience at the level of interval timing. The body of literature showing 5-HT's role in modulating time perception is small, however, we would like to speculate with the model presented in the following sections.

Model background

In this chapter, we discuss experimental evidence reported in a series of studies [1, 17, 18] based on which we later propose a putative mechanism, by means of which 5-HT promotes patience.

First, we introduce the experimental paradigm. Then, we discuss what was found in the aforementioned studies, and then discuss Miyazaki et. al's own hypothesis [1] about how 5-HT promotes waiting in the task at hand.

2.1 Body of experimental evidence

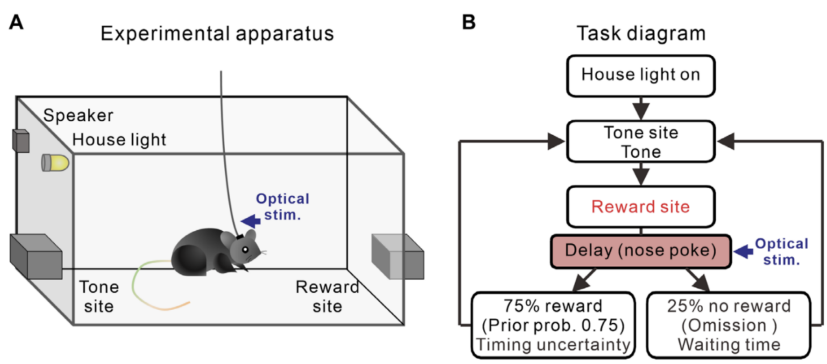

Figure 2.1: Figure taken from [1]. Schematic of the experimental set-up (A) and paradigm (B).

2.1.1 Paradigm: Sequential tone-food waiting task

In this experimental paradigm, the mouse initiates a trial by poking at a tone-site nose-port, upon which a tone is emitted. It then moves to a reward site and pokes at the corresponding nose-port. Here, the mouse is trained to wait for a

specific delay period of T seconds. If it waits for the entire duration successfully, it receives a reward. In the experiments of interest to us, the reward was one of either 1, 2 or 3 pellets of food. The mouse is free to initiate another trial by poking at the tone-site nose-port at any point during the trial; in other words, the mouse could quit/leave at any point during the trial. Refer to Fig. 2.1 for a schematic representation.

During experimental trials, the probability of reward being delivered (henceforth, the prior probability or p_r) was varied as 25%, 50% or 75%. Of interest to us is the performance of the mice in omission trials – trials where no reward was delivered even when they successfully waited for the Ts they were trained to wait for. To note here is that the mouse has no predictive cue to distinguish between reward and omission trials. For instance, in a block of trials with $p_r = 0.75$, 25% of trials are omission trials, wherein no reward is delivered upon successful waiting. The duration until the mouse leaves the reward site is considered the total waiting time, and how this value is affected by 5-HT is of particular interest to us.

2.1.2 Experiments and key findings

In a series of studies [1, 9, 17, 18], Miyazaki et. al made use of the sequential tone-food waiting task in combination with precisely timed optogenetic activation of 5-HT neurons in the dorsal raphe nucleus (DRN) in mice. This allowed them to study timing effects of 5-HT release on the promotion of waiting times during the task.

Previous studies [30] had shown that inhibition of serotonin neurons in the DRN led to mice waiting less for delayed rewards. Using genetically modified mice and an optical stimulation paradigm, Miyazaki et. al first showed that serotonin does indeed have a causal role in promotion of waiting for future rewards [9, 17]. They asked whether the effect of serotonin was a general one across the task interval, and were able to show that its effect was most significant when the stimulation was timed at the initiation of the delay period, signalled by the mouse poking its nose at the reward site nose-port.

They then probed into what factors 5-HT stimulation my be affecting and how. In [1], they showed that 5-HT stimulation promoted waiting most significantly when the prior probability of being rewarded was high, and the variability in reward delivery times was high. They modified the experimental paradigm to study this by introducing variability in the reward delivery times: whereas in the usual case, the reward delivery time was one point in time, say Ts along the trial, they delivered reward randomly at two additional time points $T \pm \Delta$s around T. They split the trials into optogenetic no-stimulation and stimulation trials (or serotonin no-activation and serotonin activation trials), wherein in the former, they shone yellow light transiently (for 0.8 s) at the beginning of the delay period as control and in the latter, they stimulated 5-HT release using

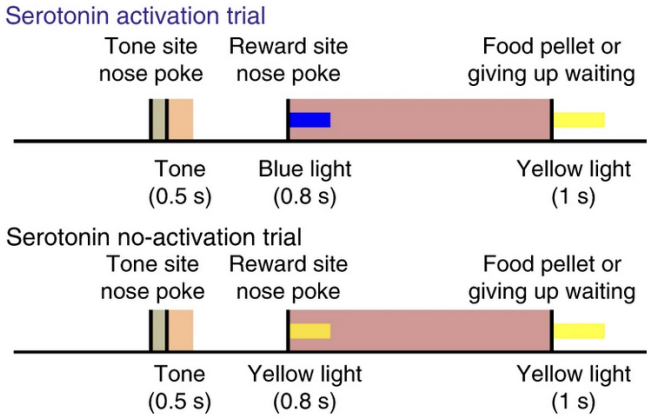

Figure 2.2: Figure taken from [1]. Schematic representation of the optogenetic stimulation paradigm. Blue light represents optogenetic activation of 5-HT neurons in the DRN; Yellow light was shone in control trials.

blue light shone for the same transient duration (Fig. 2.2). In their 2020 study [18], they followed the same modified paradigm, but with light stimulation (in both control and activation conditions) continuing throughout the duration of the delay period in omission trials. We omit further description of the findings in this study since it is not in the scope of the model described later. We summarise the key findings of the discussed studies below.

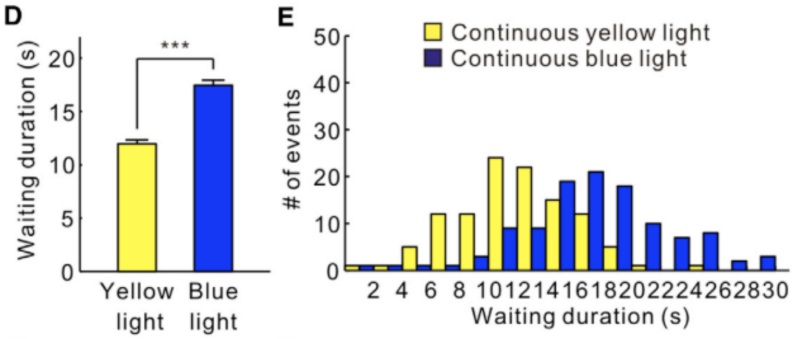

Figure 2.3: Figure taken from [9]. (D) Comparison of the mean waiting times for the yellow and blue light stimulation cases. When the animal is stimulated with blue light, it shows significantly larger average waiting times (E) Distribution of the waiting times.

5-HT neuron activation increases waiting times

The results from [17, 9] show that the stimulation condition does indeed report an increase in waiting times, as can be seen in Fig. 2.3. The authors compare waiting times in the continuous blue and yellow light conditions, and found a significant increase in the blue light (activation) condition.

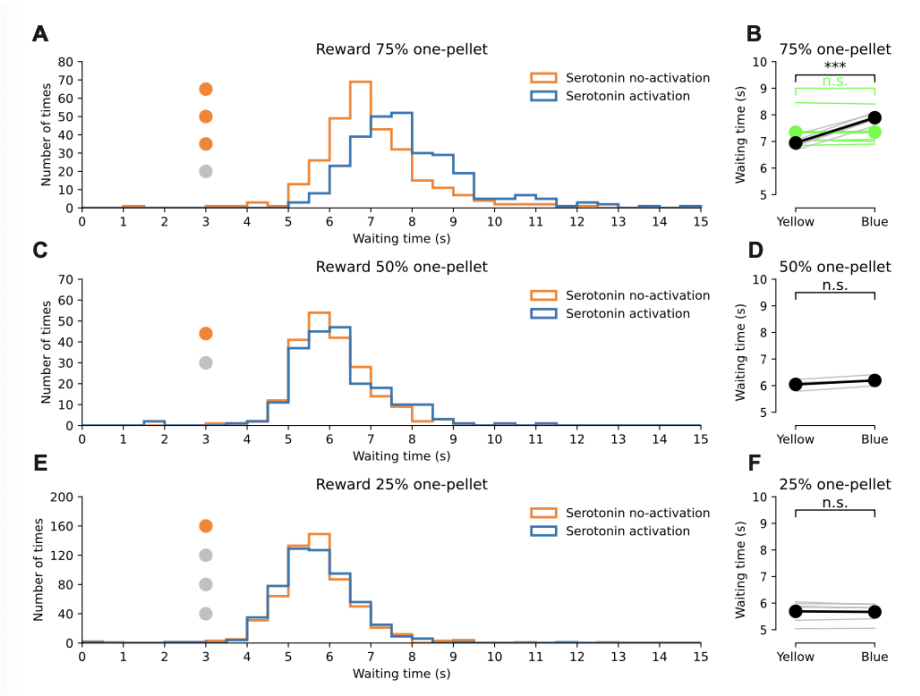

Figure 2.4: Figure courtesy P. Pierzchlewicz, data obtained from authors. (A) Distribution of waiting times for 75% reward trials. (C) Distribution of waiting times for 25% reward trial. (E) Distribution of waiting times for 50% reward trial. (B), (D), (F) show comparison of the mean waiting times. Green lines show results from wild type mice, black from optogenetic mice. A significant increase is observed for optogenetic mice in (B). Orange circles illustrate the timing and number of food pellets presented in rewarded trials. White circles denote omission trials

Serotonin promotes waiting at high reward probability

As can be seen in Fig. 2.4, the larger the prior probability of being rewarded, the later the waiting time distributions. However, the serotonin activation related shift is only significant at higher prior probabilities, as seen in panels (A) and (B) of Fig. 2.4.

Reward size increases waiting times, but not always the 5-HT related shift

The authors found that the baseline waiting distribution (from no-activation omission trials) was subject to a reward size effect, where larger rewards resulted in longer waiting times. They then tested whether this was due to a sensitivity to expected reward magnitude, but found that it was not: they did this by comparing different reward-size – p_r combinations that resulted in the same expected reward value. Their finding was that the reward size affected activation related shift in waiting times most significantly at $p_r = 0.75$

Activation related shift depends on variability in reward delivery time

Two effects are visible from this test. One: the longer the delay in reward delivery times, the later the waiting times and the variance in waiting times. Two: Serotonin activation causes significant rightward shifts in waiting times only in the high reward delivery time variability conditions. This is evident from the data, as can be seen in Fig. 2.5.

To briefly summarize, the effect of serotonin activation seems to be significant promotion of waiting when the prior probability of being rewarded and the uncertainty in reward delivery times are high. Having listed some of the key findings we wish to replicate, we finally discuss the authors' own model of serotonin's action on waiting time distributions.

2.2 Bayesian sequential/repeated decisions model of 5-HT action

In order to explain their findings, Miyazaki et. al propose a Bayesian sequential decision model of the task. They hypothesize that the mechanism by which 5-HT promotes waiting is by inflating the prior and consequently, pushing the waiting time distribution further rightward.

2.2.1 Model description

The model proposed makes the following assumptions:
The task is split into $\boldsymbol{R}$ (rewarded) trials and $\boldsymbol{NR}$ (omission) trials. The animal has no predictive cues regarding the trial type: instead, the animal observes retrospectively what the trial type was. Thus, $\boldsymbol{R}$ and $\boldsymbol{NR}$ can be treated as hidden states. The animal is assumed to make granular decisions every $\Delta\tau$ s, to either wait (W) or leave (L). This forms the set of possible actions the animal can take. If the animal chooses to leave in the current interval, then it gets no reward, and lands in a new trial state (re-initiation by means of nose-poke at

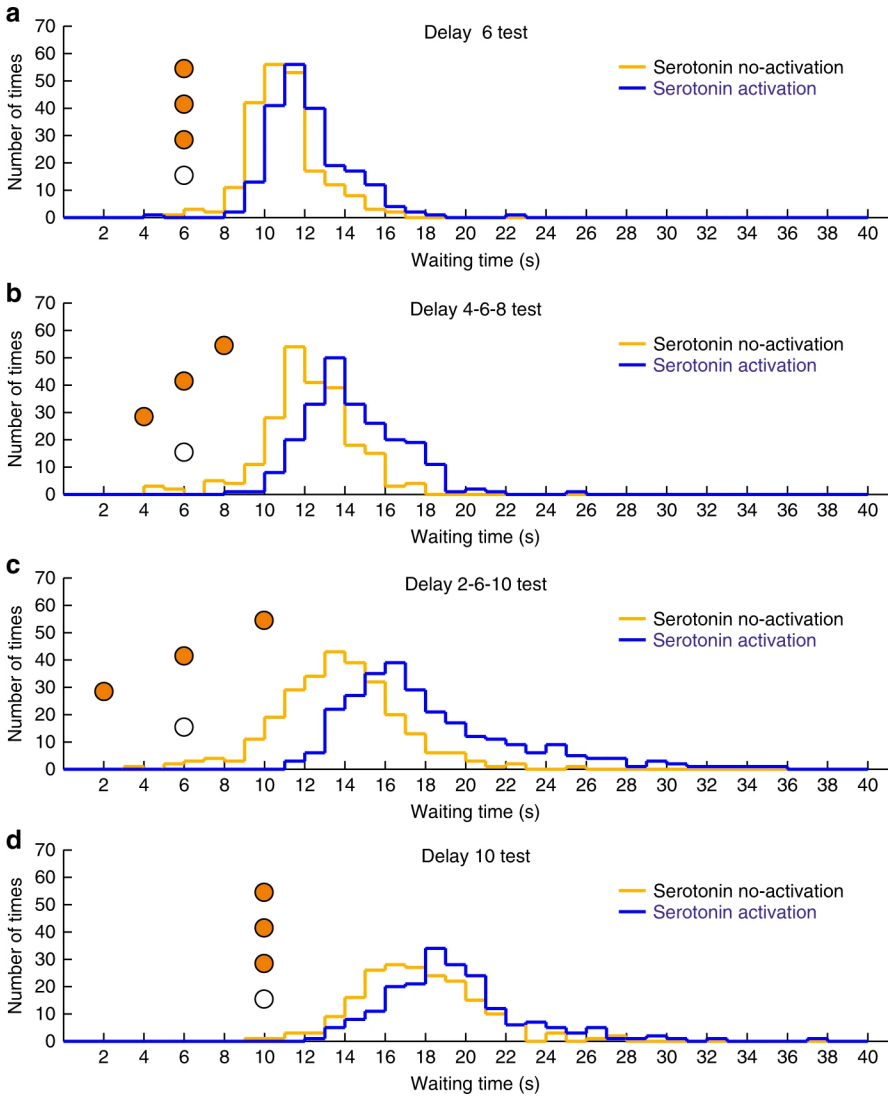

Figure 2.5: Figure taken from [1]. Optogenetic activation of DRN serotonin neurons enhances waiting for temporally uncertain rewards. a Distribution of waiting time during omission trials in the D6 test. b Distribution of waiting time during omission trials in the D4-6-8 test. c Distribution of waiting time during omission trials in the D2-6-10 test. d Distribution of waiting time during omission trials in the D10 test. Orange circles illustrate the timing and number of food pellets presented in rewarded trials. White circles denote omission trials

the tone-site). If it chooses to wait, it can either get a reward in the ensuing interval, or not. If it gets a reward, it is free to re-initiate another trial. If not, it makes another decision in the next $\Delta\tau$ s interval, following the same rules as discussed above. This forms the basis of the sequential decision model.

Formally, let the set of (hidden) states be $\mathcal{S}$; $\mathcal{S} = \{R, NR\}$. $\mathbf{S} \in \mathcal{S}$ is an arbitrary trial state. Let the set of actions be $\mathcal{A} = \{W, L\}$, and $A \in \mathcal{A}$. The values the reward can take are $\mathcal{R} \in \{1, 2, 3\}$. However, the model built here does not take this set into consideration. Let $\mathbf{T_R}$ be the reward delivery time in reward trials. While this value is deterministic, the animal can only perceive time subjectively, and therefore they assume reward timing T is a Gaussian random variable in objective time, defined as

$$T \sim \mathcal{N}(t; \mu = \mathbf{T_R}, \sigma^2)$$

The prior distribution is $P(\mathbf{S} = R)$, which gives the probability of the current trial being a reward trial. We are interested in obtaining the probability of the reward arriving at some arbitrary time t, given by the probability $P(\mathbf{S} = R|t)$ To calculate this (the posterior) using Bayes' theorem, they define the likelihood of the current trial being a reward trial $P(\mathbf{T_R} > t) = 1 - f(t; \mu, \sigma^2)$, where f is the cumulative density function of the random variable $\mathbf{T_R}$. Therefore, we can calculate the posterior as

$$P(\mathbf{S} = R|t) = \frac{P(\mathbf{S} = R)P(\mathbf{T_R} > t)}{P(\mathbf{S} = R)P(\mathbf{T_R} > t) + P(\mathbf{S} = NR) \cdot 1} \tag{2.1}$$

The posterior distribution is necessary in order to calculate the value of each possible action $V(\cdot)$ at a given time step given no reward has yet arrived.

The *value of waiting* given no reward has arrived till current time t is defined as $V(W|t) = P(\mathbf{S} = R|t)$. The *value of leaving* $V(L|t) = 0$ since no reward or punishment followed early termination of the trial. We continue with this assumption, although in their later study, Miyazaki et. al set the value of leaving to be an arbitrary small negative value.

How are decisions made in the model? Miyazaki et. al apply a softmax decision policy at every time step $\Delta\tau$, wherein the probability of choosing to wait i.e, $A = W$, given no reward has arrived until a time t is given by

$$\pi(W|t) = P(W|t) = \frac{1}{1 + \exp(-\beta \cdot V_s)} \tag{2.2}$$

where β is an inverse temperature parameter which defines the stochasticity in the model (higher β corresponds to more deterministic model outputs at every step of the way), and V_s is a function of the action values of waiting and leaving ($V(W|t)$ and $V(L|t)$ respectively), defined as

$$V_s = V(W|t) - V(L|t) \tag{2.3}$$

They then define a trajectory of sequential decisions as the sequence of actions taken at every $\Delta\tau$ seconds, until the mouse leaves the trial. Let the trajectory of decisions be defined as $\mathcal{T} = \{A_1 = W, A_2 = W, \ldots A_{n-1} = W, A_n = L\}$, where n denotes the number of granular decisions made, with the nth decision being the last one, where the animal leaves the trial. We assume that any time interval t can be written as $t = n\Delta\tau \ \forall n \in \mathbb{N}$. We can compute the probability of leaving at a time point t $P_l(t)$ as the probability of the trajectory occurring, obtained by applying the product rule on the individual actions forming the trajectory. Letting $P_w(t)$ denote the probability of waiting at time t, we have:

$$
\begin{aligned}
P_w(0) &= 1 \\
P_w(t) &= P_w(t - \Delta\tau) \times \pi(W|t) \\
P_l(t) &= P_w(t - \Delta\tau) \times (1 - \pi(W|t))
\end{aligned}
$$

The possible role of 5-HT: Within the framework of this decision model, Miyazaki et. al propose that the shifting effect of 5-HT on waiting time distributions is due to its inflation of the prior p_r. To capture this shift in the prior, Δp_r, they devised the following formulation:

$$\Delta p_r = p_r^2 - p_r^3 \tag{2.4}$$

2.2.2 Predictions

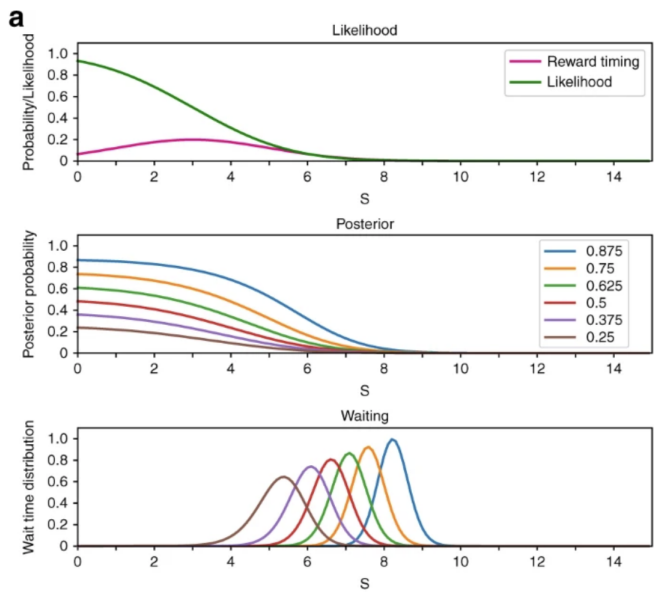

Figure 2.6: Figure taken from [1]. Shown here are the distributions related to agents simulated with reward delivery at $\mu = 3$s with $\sigma = 2$s. The top panel shows the likelihood function in green and the reward timing distribution in pink. The middle panel shows the posteriors obtained using different values of prior probabilities, shown in the corresponding legend. The third panel shows the waiting time distributions corresponding to the different prior values used in the middle panel.

The authors use the following parameter values to make model predictions comparable to the observed data for mice trained on reward delivery around 3 s: The authors assume the reward delivery distribution to be the Gaussian $\mathcal{N}(\mu = 3, \sigma = 2)$, where μ is the mean of the distribution and σ the standard deviation; $\Delta\tau = 0.1$ s and $\beta = 50$. Simulating based on these parameters, the following primary results were obtained:

1. **Higher prior probability corresponds to longer waiting times**. Additionally, the shift in waiting times produced due to 5-HT activation as modelled here increases with higher priors to start with.

2. **Later the reward delivery time, longer the waiting times**. To obtain this in the model, the value of μ is changed from 3s to higher values. This however, did not cause increased variance in waiting time distributions.

3. **Larger the variability introduced in reward delivery times during training, the longer the waiting times**. In order to increase variability, what the authors did is set the $\sigma = 3$ instead of $\sigma = 2$ as defined previously.

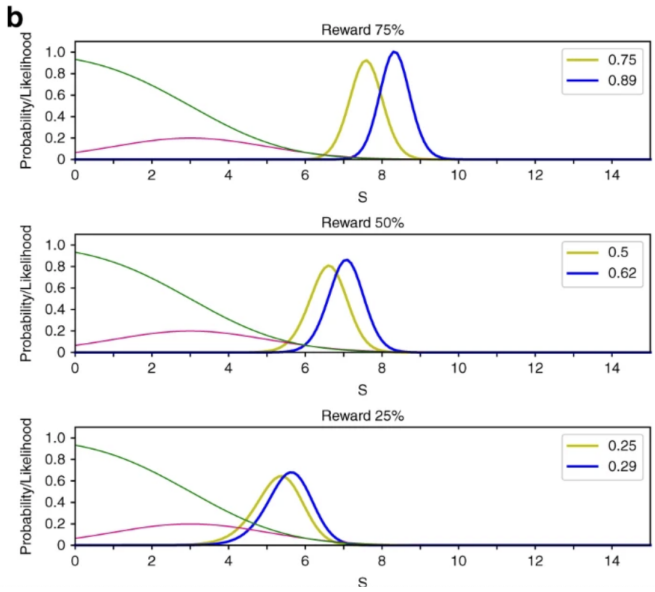

Figure 2.7: Figure taken from [1]. Shown here in yellow are the baseline responses of the model to varying priors, namely: 0.25, 0.5 and 0.75; Shown in blue are the shifted distributions obtained upon stimulation, computed as an inflation to the prior according to eq. 2.4. The corresponding values of the priors are shown in the legend in the respective panels.

4. **No reward size effect obtained.** The model proposed here does not provide a mechanism to account for the effect of reward size on waiting times, in both stimulation and control trials. Consequently, no comment can be made about the shifts caused due to reward size.

Additionally,

5. **The model is sensitive to the size of $\Delta\tau$.** Here, longer intervals result in longer waiting times with larger variances. This can be understood as a consequence of fewer decisions in the trajectory, and therefore, fewer probabilities to multiply. While the mean waiting time increases, the larger resulting variance in waiting times implies the earliest leaving time (qualitatively) shifts further leftwards.

6. **Larger β causes longer waiting times**. In other words, the more deterministic the model's decisions, the longer the resulting waiting times. On the other hand, the more stochastic the model's decisions, the more likely it is to leave at earlier time points. In order to capture what is observed in the data, setting this value suitably is consequential.

7. **Switching out the reward timing distribution from Gaussian to Gamma changes waiting time distributions**. In their 2020 study, Miyazaki et. al propose the use of a Gamma distribution instead of a

Gaussian to capture the reward delivery time distribution, arguing that it is more appropriate to model waiting times. To allow for a direct comparison between Gamma and Gaussian distributions (i.e in terms of μ and σ), the shape and scale parameters of the Gamma distributions were computed as $k = \mu^2/\sigma^2$ and $\theta = \sigma^2/\mu$ respectively. For the same μ and σ parameters, the Gamma distribution based models produced longer waiting times than their Gaussian counterparts. This can be attributed to the thicker tails of the Gamma distributed reward delivery times.

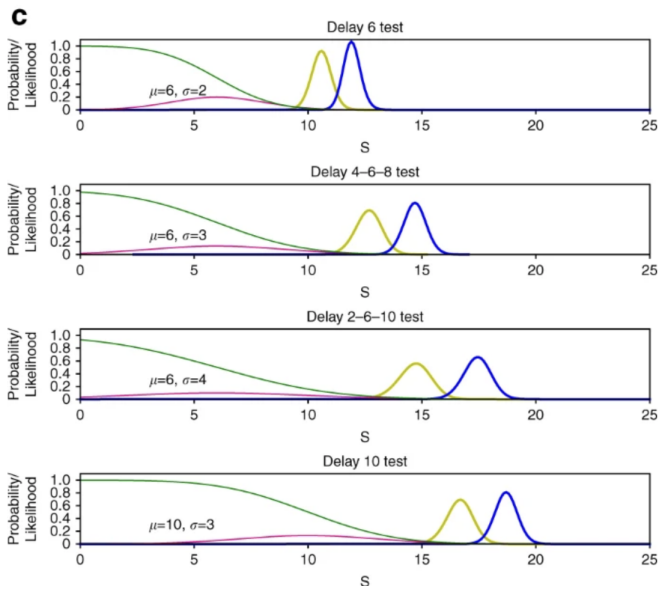

Figure 2.8: Figure taken from [1]. Distributions in yellow indicate the no-stimulation condition and distributions in blue indicate the stimulation condition. All panels show the reward delivery times (μ) they were simulated at with their corresponding delivery time variations (σ). The pink and green curves correspond to the reward delivery time distribution and the likelihood function respectively.

2.2.3 Limitations

First, the repeated decisions model proposed here does not account for reward size effect: the baseline model without invoking the shifting effect of *5-HT ON* should be able to account for the effect reward size has on waiting time distributions, as observed in data. Second, the model predicts longer waiting times when the decision intervals are longer (i.e the frequency of decision making is smaller). There is no data to measure this against. Third, while the reward delivery time is fixed in the experimental data, the model represents it as a Gaussian distribution centered around the delivery time. The authors do not justify this and do not take into consideration the animals' subjective perception

of time. Fourth, the authors modelled the shift in waiting times in stimulation trials as being the result of serotonin increasing the perceived probability of receiving a reward in the trial. They modelled the shift by fitting the data to one parameter, p_r, as shown in eq. 2.4. This, however, can only be used to explain the data at hand, and is not generalizable without further experimentation. Furthermore, it does not allow us to make any predictions about the effect of varying stimulation strength.

Model

The repeated decisions model explored in the previous chapter did not take into account some key aspects found in the data, such as the baseline effects of scalar timing and reward size sensitivity. Moreover, it did not take into account that interval timing may involve subjective values and subjective perception of time, as it only took into consideration what happens in objective time.

Here, we describe an average reward reinforcement learning model based on the previously described repeated decisions model, which takes into account the animal's subjective perception of time, and its performance of the task in objective time. We first describe an agent that operates on subjective time, to capture the baseline behaviour of the mice as is apparent from the experimental evidence previously presented. Then, we hypothesize how serotonin stimulation may be effecting the promotion of waiting times in the task.

3.1 Baseline agent

Let τ represent values in subjective time and t, objective time. The agent makes granular decisions every $\Delta\tau$ seconds in subjective time. Objective time is mapped to subjective time according to

$$p(\tau) \sim \Gamma(\mu = t, \sigma = \kappa t) \tag{3.1}$$

where κ is the scalar property ratio. The Bayesian sequential decisions model presented in [1] applies a Gaussian (Gamma in [18]) temporal blur only to the reward timing in order to introduce variability in the consequent waiting time distribution; Here, the mapping to τ effectively blurs every point in objective time t according to the Gamma distribution defined in eq. 3.1. Therefore, reward timing can be treated as a distribution instead of one point in t. That is, $p(\tau_r) \sim \Gamma(\mu = t_r, \sigma = \kappa t_r)$, where the subscript r denotes reward delivery.

We now define the underlying decision process of the agent that must capture the baseline behaviour of the mouse in the sequential tone-food task.

3.1.1 Markov decision process for the baseline agent

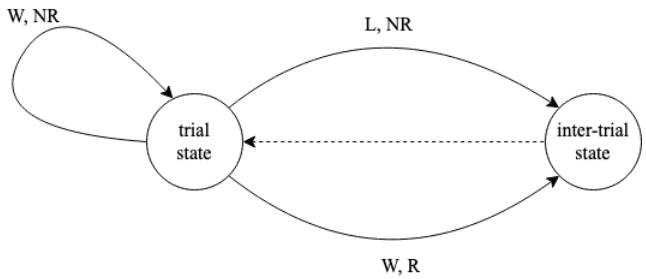

Figure 3.1: This figure illustrates a belief-state Markov decision process with two states, namely a *trial* and an *inter-trial* state. The agent has two possible actions: wait (W) or leave (L), and can either be rewarded (R) or not (NR). Once it enters the inter-trial state, it is immediately transported to the trial state anew.

In figure 3.1, we summarize the decision process described in this section. The agent makes decisions every $\Delta\tau$ s. The sets of actions at the dispense of the agent are the same as described in Section 2.2.1; Actions $A \in \mathcal{A} = \{W, L\}$. The agent can be in the (indirectly observed) reward trial or an omission trial, and either *observe* a reward (R) or not (NR), respectively. Formally, the hidden states $S \in \mathcal{S} = \{H_R, H_{NR}\}$, and the observable states $o \in O = \{R, NR\}$. Reward values can be $r \in \{1, 2, 3\}$. If in one interval it chooses to leave ($A = L$), it receives no reward ($o = NR$), and transitions to the *inter-trial* state. On the other hand, if it chooses to wait ($A = W$), there are two possible transitions: the agent receives a reward ($o = R$) and transitions to the *inter-trial* state, or does not receive a reward ($o = NR$) and stays in the trial state. We note here that there is no terminal state in this model of the task, as once the agent enters the inter-trial state, it transitions to the trial state after an inter-trial interval of τ_{iti} s, where it can get a reward r_{iti}, which is set to 0 here. This notion reflects how the mouse performs trials in the actual task.

The agent makes decisions according to a soft-max policy similar to that described in Section 2.2.1 (eq. 2.2).

$$\pi(\tau) = \frac{1}{1 + \exp(-\beta \cdot V(\tau))} \tag{3.2}$$

where $\pi(\tau)$ represents the probability of choosing $A = W$. However, the value of a state in time, $V(\tau)$, is estimated as a function of an average reward value ρ, which is updated over the course of individual trials. We describe how $V(\tau)$ and ρ are calculated in the following sub-sections. β is the inverse temperature parameter.

A partially observable Markov decision process can be considered to be a belief-state Markov decision process, and can thus be computed. We define the agent's

belief $b(\cdot)$ that a reward will arrive in the next time interval $\Delta\tau$ given the current time is τ, i.e, $\tau_r \in [\tau, \tau + \Delta\tau]$, is as follows. Let $\mathbf{S}$ be a random variable representing the underlying trial type. Let $\mathbf{O}$ be a random variable representing the observation made.

$$b(\tau, \Delta\tau) = \frac{P(\mathbf{S} = H_R) \cdot P(\tau < \tau_r \leq \tau + \Delta\tau | \mathbf{S} = H_R)}{P(\mathbf{O} = NR|\tau)} \tag{3.3}$$

$P(\mathbf{O} = NR|\tau)$ depends on whether the current trial is a reward or omission trial. If the current trial is a reward trial, then τ_r is in the future i.e $\tau_r > \tau$; If the current trial is an omission trial, then no reward arrives. Therefore, we can write

$$P(\mathbf{O} = NR|\tau) = P(\mathbf{S} = H_{NR}) + P(\mathbf{S} = H_R) \cdot P(\tau_r > \tau | \mathbf{S} = H_R) \tag{3.4}$$

$$b(\tau, \Delta\tau) = \frac{P(\mathbf{S} = H_R) \cdot P(\tau < \tau_r \leq \tau + \Delta\tau | \mathbf{S} = H_R)}{P(\mathbf{S} = H_{NR}) + P(\mathbf{S} = H_R) \cdot P(\tau_r > \tau | \mathbf{S} = H_R)} \tag{3.5}$$

$$= \frac{p_r \cdot \Delta\tau \cdot f(\tau)}{(1 - p_r) + p_r \cdot \Phi(\tau)} \tag{3.6}$$

where $f(\tau)$ is the probability density function of the reward delivery distribution, and $\Phi(\tau)$ is the cumulative distribution function. p_r is the probability of the current trial being a reward trial.

Calculating $V(\tau)$

The value of a state (an instant) in time τ during a trial interval depends on the values of the actions of waiting and leaving at that particular instant. Therefore, we first discuss how these action values $Q(\cdot)$ are calculated.

- Let G_τ be the average reward adjusted total return from choosing to wait this instant τ.

$$G_\tau = \sum_{i=0}^{\infty} r_{\tau + i\Delta\tau} - \rho \tag{3.7}$$

- At any instant τ, the value of that instant (i.e state) is

$$V(\tau) = \pi(\tau) \cdot Q(W|\tau) + (1 - \pi(\tau) \cdot Q(L|\tau) \tag{3.8}$$

- The value of leaving at an instant τ given no reward has arrived until then, $Q(L|\tau)$, is the opportunity cost of time; There is no other penalty or value for choosing to leave since the agent only transitions to the inter-trial state in this case. Thus,

$$Q(L|\tau) = -\rho\Delta\tau \tag{3.9}$$

- The value of waiting at an instant τ given no reward has arrived until then, $Q(W|\tau)$ is defined as the expected return following choosing to wait under the current action policy.

$$\begin{aligned} Q(W|\tau) &= \mathbb{E}_\pi[G_\tau|A = W, S = \tau] \\ &= b(\tau, \Delta\tau) \cdot (r + r_{iti}) + (1 - b(\tau, \Delta\tau)) \cdot V(\tau + \Delta\tau) - \rho\Delta\tau \end{aligned}$$

But $r_{iti} = 0$. Thus,

$$Q(W|\tau) = b(\tau, \Delta\tau) \cdot (r) + (1 - b(\tau, \Delta\tau)) \cdot V(\tau + \Delta\tau) - \rho\Delta\tau \tag{3.10}$$

where the first term corresponds to the case where $A = W$ and reward arrives at the next time step; the second term corresponds to the case where $A = W$ and no reward arrives in the next time step, and consequently, depends on the value of the next state; and the third term corresponds to the opportunity cost of time.

Calculating ρ

The average reward ρ is a function of the expected total waiting time, and is defined as the expected return per unit time. We first describe how to calculate the expected total waiting time, and then estimate the expected return per unit time.

Let τ_w represent the total waiting time; $\hat{\tau}_w$, the expected total waiting time; Let $\lambda_{A,\tau_w} = \{A_0 = W, A_{\Delta\tau} = W, A_{2\Delta\tau}, \ldots, A_{\tau_w} = L\}$ describe the trajectory of a trial in terms of decisions A_i made at every decision interval until the agent chooses to leave. Let the optimal expected waiting times in omission trials be $\hat{\tau}_{w,NR}$, and that in the reward trials be $\hat{\tau}_{w,R}$.

For omission trials, the probability of each trajectory can be computed using the product rule. Weighting this with the total waiting time that results and then summing across different trajectories, thus, gives us

$$\hat{\tau}_{w,NR} = \mathbb{E}[\tau_w|\mathbf{S} = H_{NR}] \tag{3.11}$$

$$= \sum_{\tau_w=0} \tau_w \cdot \Pi_{A_i \in \lambda_{A,\tau_w}} \pi(A_i) \tag{3.12}$$

For reward trials, the optimal expected waiting time is approximately the same as the length of the trial, which in this case is when the reward is delivered, t_r.

$$\hat{\tau}_{w,R} = \mathbb{E}[\tau_w|\mathbf{S} = H_R] \approx t_r \tag{3.13}$$

The expected length of each trial $\bar{\tau}_w$ is defined as

$$\bar{\tau}_w = b(0, \hat{\tau}_{w,NR}) \cdot \hat{\tau}_{w,R} + (1 - b(0, \hat{\tau}_{w,NR})) \cdot \hat{\tau}_{w,NR} \tag{3.14}$$

which is obtained by marginalizing over trial type. Here, the agent either receives a reward with the probability $b(0, \hat{\tau}_{w,NR})$ and waits on average for $\hat{\tau}_{w,R}$ s, or doesn't and waits $\hat{\tau}_{w,NR}$ s.

From this, we get the average reward, which is the expected return per unit time.

$$\rho = \frac{b(0, \bar{\tau}_w) \cdot r}{\bar{\tau}_w + \tau_{iti}} \tag{3.15}$$

3.1.2 Results

Several agents based on the above model were simulated. The model was solved using dynamic programming. The agents learnt according to the policy iteration algorithm with the soft-max policy in eq. 3.2 as the target (only the values were learnt, so we did not start from a random policy). $\beta = 100$, so the model was fairly deterministic; This value was chosen such that we could sufficiently replicate the range of values of effects seen in the data. However, $\beta \geq 38$ did not cause significant changes in the outcomes of the models. To make the results comparable to what was found in the data in [1], simulations presented here use the same parameter values for μ for the reward delivery distributions; σ is essentially μ scaled by κ in accordance with the scalar expectancy theory. Additionally, $p_r \in \{0.25, 0.5, 0.75\}$; $r \in \{1, 2, 3\}$; $\kappa = 0.2$;

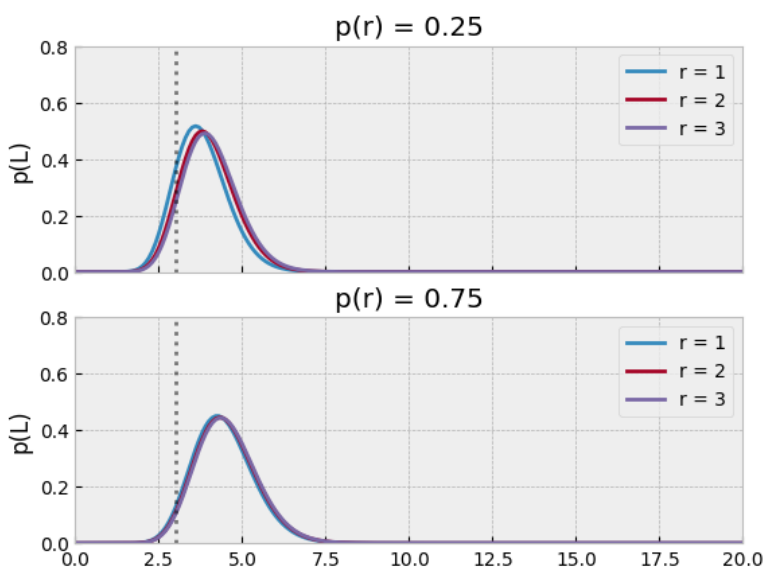

Figure 3.2: The baseline model's sensitivity to different reward sizes at low and high prior probabilities, with reward delivery time at $\mu = 3$s, indicated by the gray dotted line. Upper panel shows waiting time distributions when $p_r = 0.75$; the lower panel shows the waiting time distributions when $p_r = 0.75$

Reward size does not affect waiting times

The baseline model is not sensitive to reward size. While there are slight differences detectable at lower p_r, they are not significant. The differences almost completely disappear at higher reward probabilities. This is illustrated in Fig. 3.2. The waiting time distributions obtained are centered around smaller times than in Miyazaki et. al's model, possibly due to the additional opportunity cost in the value calculation – there were no costs in their model.

Higher prior probability corresponds to longer waiting times

This effect is illustrated in Fig. 3.3. Agents were simulated at $p_r \in \{0.25, 0.5, 0.75\}$. As seen in the data and in the Miyazaki model, higher the prior, longer the waiting times. Additionally, the waiting time distributions get flatter, in accordance with the scalar expectancy theory.

Longer delay times result in later, flatter waiting time distributions

As expected from scalar timing theory, the later reward delivery times produced flatter (higher variance) waiting time distributions. This effect can be seen by comparing the top-most and bottom-most panels in Fig. 3.5 .

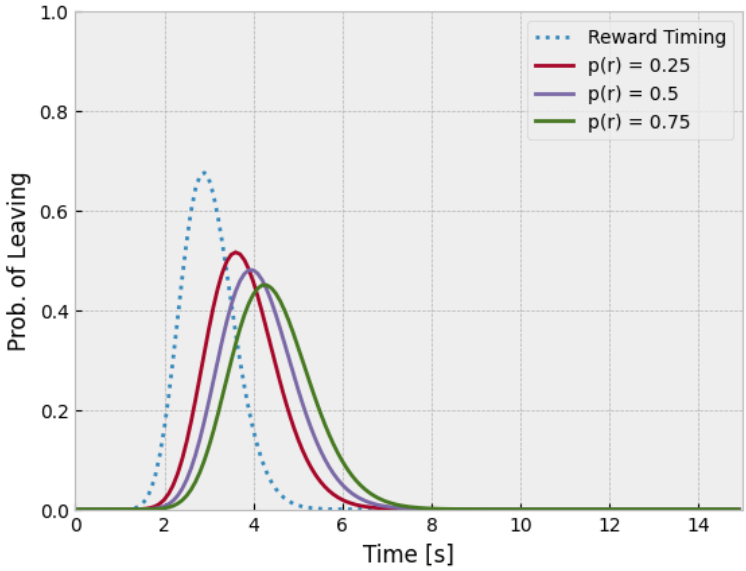

Figure 3.3: The baseline model's response to varying reward probabilities. The dotted-line distribution indicates the reward timing distribution made available to the model, centered at $\mu = 3$s. The thick-lined distributions show the respective waiting time distributions of agents simulated with p_r values of $0.25, 0.5$ and 0.75.

The model is not sensitive to either changes in $\Delta\tau$ or β

Refer to Fig. 3.4 for the relevant illustration. The model showed no significant effects on varying either the inverse temperature parameter or the decision interval (caveat: for very small values, however, the gamma distribution shape changes, and therefore, it stops being valid for our purposes). For larger $\Delta\tau$ values, the curve is less smooth, but no other changes are apparent. For larger βs, there is a slight shift towards higher waiting times, but the effect is negligible. Both of these results are desirable, and an improvement to the model proposed by Miyazaki et. al.

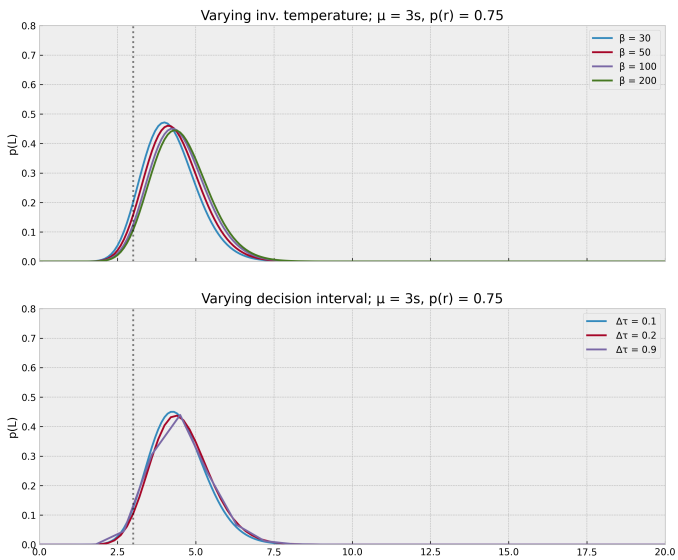

Figure 3.4: The baseline model's response to varying auxillary parameters β and $\Delta\tau$ The colors correspond to the values of the parameters shown in the legend. All agents were simulated with $\mu = 3$s, and with the prior probability $p_r = 0.75$.

Higher variability in reward timing results in longer waiting times

This effect is illustrated in Fig. 3.5. Comparing the three middle panels allows us to look at how waiting time distributions change with increasing variability in reward delivery times (RDT) centered around the same value, namely, $\mu = 6$s. With higher RDT variability ($\Delta = 2$s in the Delay-4-6-8 test, $\Delta = 4$s in the Delay-2-6-10 test), the waiting time distributions are centered around later times, and are qualitatively flatter. This means that the waiting time distributions show higher variances. Moreover, the effect is most evident at $p_r = 0.75$. This is consistent with the data in [1], albeit with smaller values in the model than in the data, where the effect was only significant at higher priors. An observation to make here is that at lower priors, not only are the waiting times centered around shorter times, at $p_r = 0.25$, the Delay-2-6-10

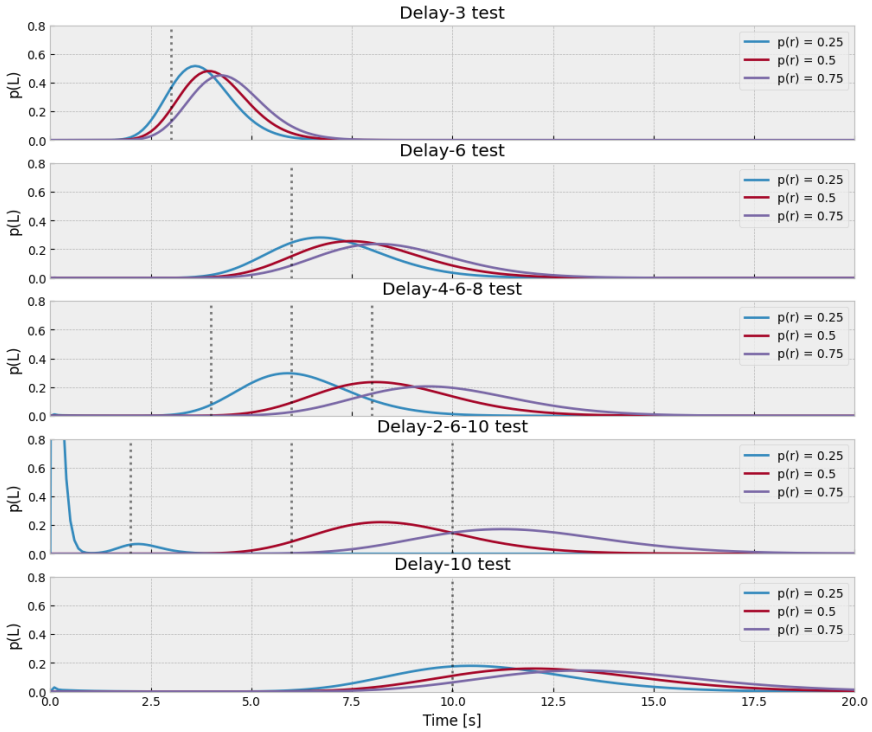

Figure 3.5: The baseline model's response to varying reward delivery times. The label at the top of each panel indicates the training scheme and the corresponding reward delivery times, as in Miyazaki et. al's original experiments [1]. In each panel, the gray dotted line indicates when the reward was delivered. In each panel, the blue, pink and purple curves correspond to the waiting time distributions of agents simulated with priors $0.25, 0.5$ and 0.75 respectively.

($\Delta = 4$s) condition pushes the waiting times to values that indicate the agent does not wait much beyond the RDT, even choosing to leave as soon as the trial is initiated.

3.1.3 Limitations

The model does not show the reward size effect apparent in the data. This is a major limitation, since what we have modelled here is the baseline behavior of the animal performing the task. The waiting times predicted by this model are smaller than those produced by Miyazaki et. al's Bayesian sequential decisions model. The model is slightly sensitive to how the value function is initialized – it does not learn a good value function if initialization is random.

3.2 Internal clock model of 5HT action

In the previous section, we described a baseline RL agent with subjective perception of time. Here, we try to capture the shifts in waiting time distributions observed in 5-HT stimulation trials in [1].

3.2.1 Modulating subjective clock speed

We hypothesize here that serotonin acts by influencing subjective clock speed, in accordance with the previously discussed literature related to scalar timing and dopamine opponency. Consider an agent with a subjective (internal) clock that ticks faster than an objective clock. In this case, the agent constantly overestimates how much objective time has passed, and the frequency of decisions made increases; Consequently, it waits for shorter periods of time. On the other hand, if the agent's internal clock runs slower than an objective clock, it will underestimate how much actual time has passed. In this case, the agent's frequency of decisions decreases, and effectively, its waiting times may increase.

In our model, the interval $\Delta\tau$ is dependent on the speed of the internal clock. We assume that at the start of every trial, the clock speed is reset and stays constant throughout the duration of a trial. We make this assumption for the sake of simplicity.
Let η be a random variable representing clock speed, drawn at the start of each trial. It can vary from trial to trial. Let the clock tick every $\Delta\tau$ seconds, and this interval is mapped according to $\Delta\tau \sim \Gamma(\mu = \Delta t, \sigma = \kappa)$, where all the symbols have their usual meanings. We also note the relation $\eta = 1/\Delta\tau$. In the previous section (3.1.1, **Calculating** ρ), we treated a trial as a trajectory λ, made of n sequential decisions; λ therefore lasts for $\tau = n\Delta\tau$ seconds.

$$\begin{aligned}
\Delta\tau &\sim \Gamma(\mu = \Delta t, \sigma = \kappa\Delta t) \\
n\Delta\tau &\sim \Gamma(\mu = n\Delta t, \sigma = \kappa n\Delta t) \\
\therefore \tau &\sim \Gamma(\mu = t, \sigma = \kappa t)
\end{aligned}$$

We are interested in the change to mean clock speed, over trials. Let the base mean clock speed be $\hat{\eta}_{old}$. Let the new mean clock speed, say, after stimulation, be $\hat{\eta}_{new}$. We define the *shift ratio*

$$\bar{\eta} = \frac{\hat{\eta}_{new}}{\hat{\eta}_{old}} = \frac{\Delta\hat{\tau}_{old}}{\Delta\hat{\tau}_{new}} \tag{3.16}$$

Incorporating the shift ratio, we map subjective time back to objective time as

$$t \sim \Gamma\left(\mu = \frac{\tau}{\bar{\eta}}, \sigma = \frac{\kappa\tau}{\bar{\eta}}\right) \tag{3.17}$$

where the baseline agent can be simulated by setting the shift ratio to 1. Now, how do we relate this to a serotonin activation signal?

Let $\Delta 5\text{-}HT$ represent the relative change in serotonin activation; if optogenetic stimulation of the DRN results in an $x\%$ increase in sertonin, then $\Delta 5\text{-}HT = x\%$. We relate the shift ratio with this relative change in serotonin activation according to

$$\bar{\eta} = 1 - p_r \cdot \Delta 5\text{-}HT \tag{3.18}$$

Here, a larger $\Delta 5\text{-}HT$ results in a smaller $\bar{\eta}$, which implies a decrease in mean clock speed opposed to the baseline, i.e, $\Delta\hat{\tau}_{old} < \Delta\hat{\tau}_{new}$. This formulation of the shift ratio is additionally sensitive to the prior reward probability. An alternative formulation that also introduces a sensitivity to reward magnitude is

$$\bar{\eta} = 1 - p_r \cdot r \cdot \Delta 5\text{-}HT \tag{3.19}$$

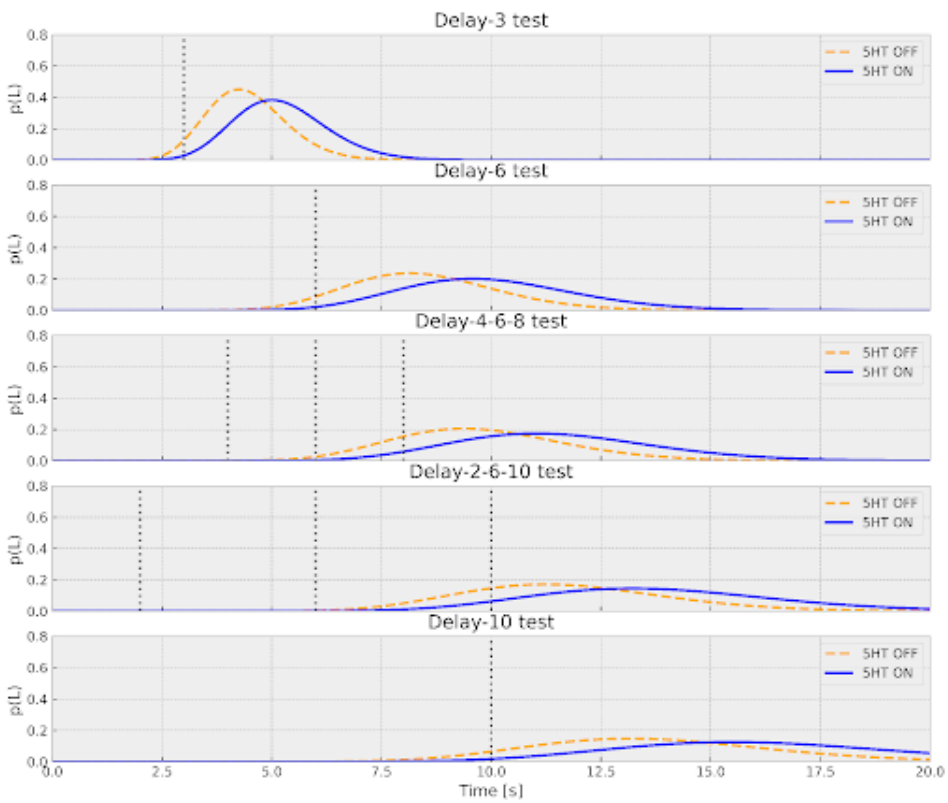

Figure 3.6: Shift in waiting times at different reward delivery delays. All agents were simulated at $p_r = 0.75$. In each panel, the gray dotted like indicates the reward delivery times. The orange dashed-line distribution indicates the baseline waiting times as in no-stimulation trials. The blue distribution indicates waiting times in stimulation trials. The label at the top of each panel indicates the reward delivery scheme the agent was simulated with.

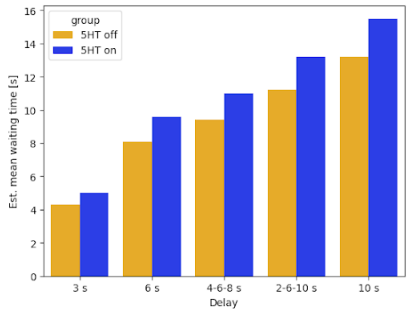

Figure 3.7: Histogram accompanying Fig. 3.6, where the colors have the same meanings.

In the next section, we look at the results obtained by simulating agents with the mapping between subjective and objective times as described in eq. 3.17, with shift ratios formulated as in eq. 3.18 and eq. 3.19. We focus on the stimulation versus non-stimulation cases, as the properties of the baseline agent carries over to the results here. We set $\Delta 5\text{-}HT = 0.2$ based on values reported in [1].

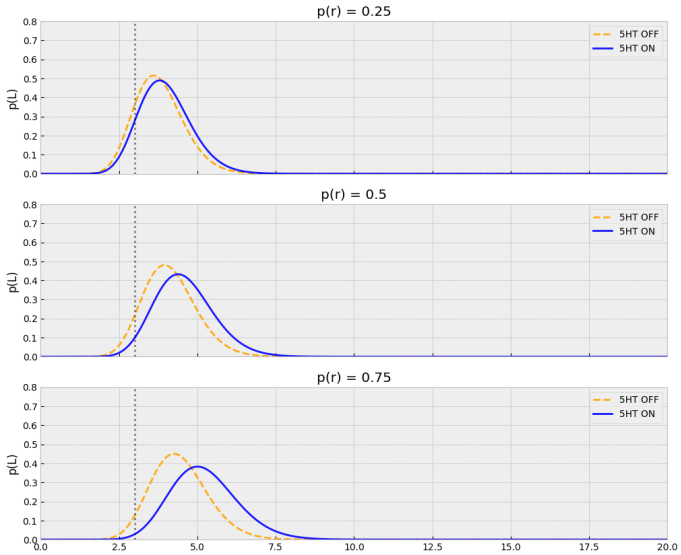

Figure 3.8: Shift in waiting times at different prior probabilities. In each panel: The gray dotted line indicates the reward delivery time, which is 3s here. The orange dashed-line distribution indicates the no-stimulation trials, and reflects baseline waiting times. The blue distribution indicates stimulation trials. The labels at the top of each panel indicates the prior of the simulated agent.

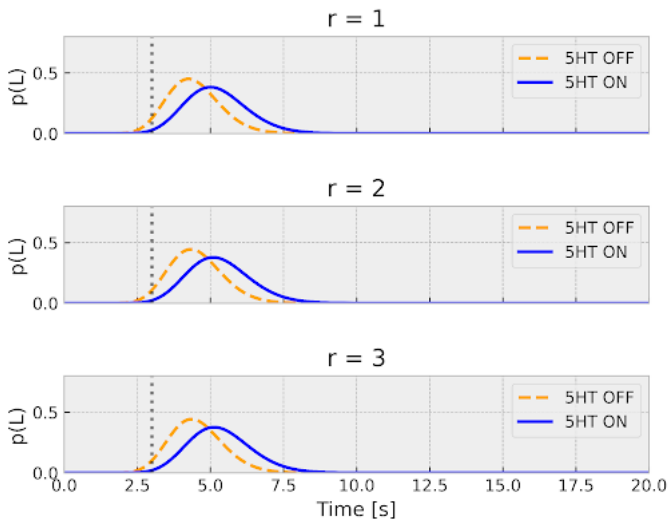

Figure 3.9: Shift in waiting times at different reward sizes for $p_r = 0.75$; $\bar{\eta} = 1 - p_r \cdot \Delta 5\text{-}HT$ in all panels. In each panel, the gray dotted line indicates the reward delivery times. The orange dashed-line distribution indicates the baseline waiting times as in no-stimulation trials. The blue distribution indicates waiting times in stimulation trials. The label at the top of each panel indicates the size of the reward the agent was simulated with.

3.2.2 Results

Here, we discuss the predictions of using the above suggested mappings between subjective and objective times. To note, the baseline agent remains as described in the previous section. What we discuss here is strictly related to the proposed effect of serotonin upon shifts to the waiting time distributions of the agents in different task conditions.

Stimulation causes shifts in waiting times in a p_r dependent manner

This is apparent in Fig. 3.8, wherein you see larger shifts upon stimulation for larger priors, most evidently for $p_r = 0.75$. This is in accordance to what is observed in the data. However, the shifts are not as large as those seen in the data.

Larger variability in RDT makes shifts due to stimulation larger

As can be seen in Fig. 3.6, for a model simulated at $p_r = 0.75$, the larger the Δ, the longer the waiting times. Additionally, the variances in the waiting times also increase. The shifts also show a sensitivity to the delay time itself: longer delays produce larger variances in waiting times, and the shift is larger as well. This is more evident in the histogram corresponding to this result, shown in Fig. 3.7.

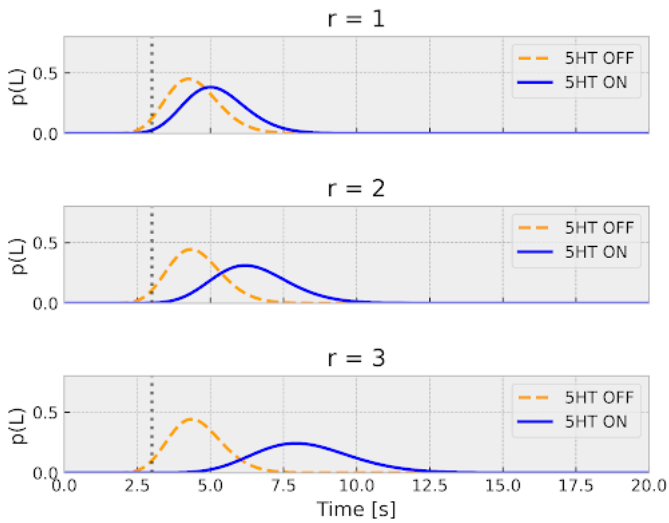

Figure 3.10: Shift in waiting times at different reward sizes for $p_r = 0.75$; $\bar{\eta} = 1 - p_r \cdot r \cdot \Delta 5\text{-}HT$ in all panels. In each panel, the gray dotted line indicates the reward delivery times. The orange dashed-line distribution indicates the baseline waiting times as in no-stimulation trials. The blue distribution indicates waiting times in stimulation trials. The label at the top of each panel indicates the size of the reward the agent was simulated with.

Reward size sensitivity of stimulation related shifts

While the baseline model does not show reward size sensitivity, the second formulation of the shift ratio (eq. 3.19) allows simulation of reward-size sensitivity to serotonin stimulation. This however, does not replicate the full desirable effect, since the baseline agent should be able to show this effect, and stimulation should simply be an additional effect, as per what is apparent from the data. Compare figures 3.9 and 3.10.

3.2.3 Limitations

While the internal clock mechanism suggested here is able to replicate stimulation related shifts in waiting times, the waiting times are not as long as found in the data. The effect of RDT variation on the shifts is not as sharp as reported in [1], but the centers of the reported distributions are relatively closer than those predicted by the Bayesian sequential decisions model discussed in the previous chapter.

Additionally, the reward size effect simulated here applies just to the stimulation related shift in waiting time distributions. It does not allow us to make any comments on the baseline agent's reward size sensitivity. An alternative justification however, is that, since the shift ratio is essentially a knob in the mapping between the agent's subjective perception of time and objective time, it is still a valid approach.

Discussion

In this report, we reviewed a very limited set of effects of the neuromodulator serotonin, namely, its effect on promotion of waiting. However, one must acknowledge the serotonergic system's involvement in myriad functions, which makes experimentally disentangling specific effects of serotonin on behaviour a very difficult task. Several theories have been put forth to explain its mechanisms of action, such as behavioural inhibition [19, 20] and impulsivity [21], dopamine opponency [3, 22, 31], or in an attempt to be more general, the beneficialness theory [2, 16]. Yet, we are not close to comprehensively understanding how serotonin works.

A possible mechanism by which 5-HT could affect the perception of time intervals is proposed here, in alignment with dopamine opponency and SET. The studies in focus in this report are from Miyazaki et. al, who show that serotonin does indeed show regulation of waiting in interval timing tasks. The data in these studies speak to a non-trivial underlying mechanism, as it shows that a simple activation of the serotonergic system is not sufficient to promote waiting. Instead, it promotes waiting in an ethologically relevant manner, reminiscent of what is proposed in the beneficialness view of serotonin action: serotonin promotes waiting only when the animal is fairly certain of reward arriving, only highly uncertain of when it will arrive around the time it has learnt to estimate as the delay period. Here, one needs to take into account not just subjective variability in interval estimation, but also the external variability in timing introduced by the experimenters, which makes interval estimation not-so-straightforward.

Miyazaki et. al proposed a Bayesian sequential decisions model which captured some features of the data but failed to replicate some key others. Even so, in order to replicate what they were able to, they fit larger variances to the model's internal representation of reward delivery time than predicted by SET. Moreover, what is most evident is that the mechanism they provided to explore increase in waiting times is wanting for biological plausibility. It also provides no way to capture the reward size effect.

Here, increase in waiting times is posited to be a result of slowing internal clock speed. A slower clock means that the agent perceives lesser objective time to have passed, and therefore, makes more patient decisions – the decisions in question being to wait or to leave the trial in the current decision interval. Effectively, a tempting interpretation of this in terms of beneficialness of the current context (of waiting), is that serotonin increases motivation to stay in the state by reporting the benefits of staying in that state, and the report is translated in terms of internal clock speed. However, since we posit that serotonin acts to slow clock speed in order to increase waiting times/patience, it is positioned as acting in opponency to dopamine.

The discussed internal clock based mechanism also does not show a key feature of the behavioral data: reward size sensitivity. While the mapping from subjective to objective in eq. 3.19 allows for a reward size effect, it does not act completely in accordance with what is observed in the data as it also causes shifts in waiting times at lower probability conditions, with only the reward size being able to modulate shift enough to produce a significant effect. This model can explain why the shift effect of serotonin is not always observed in experiments, but it employs the relationship between the neuromodulator and clock speed, which is still a link up for debate due to lack of evidence.

Bibliography

[1] K. Miyazaki, K. W. Miyazaki, A. Yamanaka, T. Tokuda, K. F. Tanaka, and K. Doya, "Reward probability and timing uncertainty alter the effect of dorsal raphe serotonin neurons on patience," *Nat. Commun.*, vol. 9, June 2018.

[2] L. Liu, M. Fuller, T. P. Behymer, Y. Ng, T. Christianson, S. Shah, N. K. K. King, D. Woo, and M. L. James, "Selective serotonin reuptake inhibitors and intracerebral hemorrhage risk and outcome," *Stroke*, vol. 51, pp. 1135–1141, Apr. 2020.

[3] P. Dayan and Q. J. M. Huys, "Serotonin in affective control," *Annu. Rev. Neurosci.*, vol. 32, no. 1, pp. 95–126, 2009.

[4] K. Z. Peters, J. F. Cheer, and R. Tonini, "Modulating the neuromodulators: Dopamine, serotonin, and the endocannabinoid system," *Trends Neurosci.*, vol. 44, pp. 464–477, June 2021.

[5] L. A. W. Jans, G. A. H. Korte-Bouws, S. M. Korte, and A. Blokland, "The effects of acute tryptophan depletion on affective behaviour and cognition in brown norway and sprague dawley rats," *J. Psychopharmacol.*, vol. 24, pp. 605–614, Apr. 2010.

[6] K. W. Huang, N. E. Ochandarena, A. C. Philson, M. Hyun, J. E. Birnbaum, M. Cicconet, and B. L. Sabatini, "Molecular and anatomical organization of the dorsal raphe nucleus," *Elife*, vol. 8, Aug. 2019.

[7] E. S. Bromberg-Martin, O. Hikosaka, and K. Nakamura, "Coding of task reward value in the dorsal raphe nucleus," *J. Neurosci.*, vol. 30, pp. 6262–6272, May 2010.

[8] T. Kawashima, M. F. Zwart, C.-T. Yang, B. D. Mensh, and M. B. Ahrens, "The serotonergic system tracks the outcomes of actions to mediate short-term motor learning," *Cell*, vol. 167, pp. 933–946.e20, Nov. 2016.

[9] K. Miyazaki, K. W. Miyazaki, and K. Doya, "The role of serotonin in the regulation of patience and impulsivity," *Mol. Neurobiol.*, vol. 45, pp. 213–224, Apr. 2012.

[10] S. Matias, E. Lottem, G. P. Dugué, and Z. F. Mainen, "Activity patterns of serotonin neurons underlying cognitive flexibility," *Elife*, vol. 6, Mar. 2017.

[11] G. Dölen, A. Darvishzadeh, K. W. Huang, and R. C. Malenka, "Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin," *Nature*, vol. 501, pp. 179–184, Sept. 2013.

[12] R. Cools, A. C. Roberts, and T. W. Robbins, "Serotoninergic regulation of emotional and behavioural control processes," *Trends Cogn. Sci.*, vol. 12, pp. 31–40, Jan. 2008.

[13] J. F. Quist, C. L. Barr, R. Schachar, W. Roberts, M. Malone, R. Tannock, V. S. Basile, J. Beitchman, and J. L. Kennedy, "The serotonin 5-HT1B receptor gene and attention deficit hyperactivity disorder," *Mol. Psychiatry*, vol. 8, pp. 98–102, Jan. 2003.

[14] J. Michely, I. M. Martin, R. J. Dolan, and T. U. Hauser, "Boosting serotonin increases information gathering by reducing subjective cognitive costs," *J. Neurosci.*, vol. 43, pp. 5848–5855, Aug. 2023.

[15] W. Zhong, Y. Li, Q. Feng, and M. Luo, "Learning and stress shape the reward response patterns of serotonin neurons," *J. Neurosci.*, vol. 37, pp. 8863–8875, Sept. 2017.

[16] Y. Li, W. Zhong, D. Wang, Q. Feng, Z. Liu, J. Zhou, C. Jia, F. Hu, J. Zeng, Q. Guo, L. Fu, and M. Luo, "Serotonin neurons in the dorsal raphe nucleus encode reward signals," *Nat. Commun.*, vol. 7, p. 10503, Jan. 2016.

[17] K. W. Miyazaki, K. Miyazaki, K. F. Tanaka, A. Yamanaka, A. Takahashi, S. Tabuchi, and K. Doya, "Optogenetic activation of dorsal raphe serotonin neurons enhances patience for future rewards," *Curr. Biol.*, vol. 24, pp. 2033–2040, Sept. 2014.

[18] K. Miyazaki, K. W. Miyazaki, G. Sivori, A. Yamanaka, K. F. Tanaka, and K. Doya, "Serotonergic projections to the orbitofrontal and medial prefrontal cortices differentially modulate waiting for future rewards," *Sci. Adv.*, vol. 6, p. eabc7246, Nov. 2020.

[19] J. F. Deakin and F. G. Graeff, "5-HT and mechanisms of defence," *J. Psychopharmacol.*, vol. 5, pp. 305–315, Jan. 1991.

[20] P. Faulkner and J. F. W. Deakin, "The role of serotonin in reward, punishment and behavioural inhibition in humans: insights from studies with acute tryptophan depletion," *Neurosci. Biobehav. Rev.*, vol. 46 Pt 3, pp. 365–378, Oct. 2014.

[21] P. Soubrié, "Reconciling the role of central serotonin neurons in human and animal behavior," *Behav. Brain Sci.*, vol. 9, pp. 319–335, June 1986.

[22] J. A. Kauer and A. M. Polter, "Two-pronged control of the dorsal raphe by the VTA," *Neuron*, vol. 101, pp. 553–555, Feb. 2019.

[23] S. Mahadevan, "Average reward reinforcement learning: Foundations, algorithms, and empirical results," *Mach. Learn.*, vol. 22, no. 1-3, pp. 159–195, 1996.

[24] G. B. M. Mello, S. Soares, and J. J. Paton, "A scalable population code for time in the striatum," *Curr. Biol.*, vol. 25, pp. 1113–1122, May 2015.

[25] C. Malapani, B. Rakitin, R. Levy, W. H. Meck, B. Deweer, B. Dubois, and J. Gibbon, "Coupled temporal memories in parkinson's disease: a dopamine-related dysfunction," *J. Cogn. Neurosci.*, vol. 10, pp. 316–331, May 1998.

[26] C. Malapani and S. Fairhurst, "Scalar timing in animals and humans," *Learn. Motiv.*, vol. 33, pp. 156–176, Feb. 2002.

[27] C. V. Buhusi and W. H. Meck, "What makes us tick? functional and neural mechanisms of interval timing," *Nat. Rev. Neurosci.*, vol. 6, pp. 755–765, Oct. 2005.

[28] M. J. Allman, S. Teki, T. D. Griffiths, and W. H. Meck, "Properties of the internal clock: First- and second-order principles of subjective time," *Annu. Rev. Psychol.*, vol. 65, pp. 743–771, Jan. 2014.

[29] J. Gibbon, "Scalar expectancy theory and weber's law in animal timing," *Psychol. Rev.*, vol. 84, pp. 279–325, May 1977.

[30] B. L. Jacobs and C. A. Fornal, "Activity of serotonergic neurons in behaving animals," *Neuropsychopharmacology*, vol. 21, pp. 9S–15S, Aug. 1999.

[31] Y.-L. Boureau and P. Dayan, "Opponency revisited: competition and cooperation between dopamine and serotonin," *Neuropsychopharmacology*, vol. 36, pp. 74–97, Jan. 2011.