Fitting a reinforcement learning model to data from a perceptual decision task in mice with asymmetric midbrain dopamine stimulation

Catherine Hastings

Abstract

We describe what is known about the role dopamine plays in making decisions. We give an introduction to reinforcement learning (RL) computational models and explain why it is appropriate to apply such models to perceptual decision tasks.

We give a description of a perceptual, two-alternative, forced decision task given to mice. The mice receive stimulation of midbrain dopamine neurons in an asymmetric manner according to the actions they take. We fit a RL model to previously collected data. Analysing the results of such a fitting gives an estimate of the learning rates of animals, how much they value dopamine stimulation in comparison with a water reward and their uncertainty in the position of a visual stimulation.

1 Introduction

1.1 The role of dopamine in decision-making

Dopamine has long been shown to play a significant role in reward-motivated behaviour. Olds and Milner (1954) electrically stimulated various parts of the brains of rats when a lever was pushed. When the part of the brain being stimulated was associated with dopamine-producing neurons, the rats continued to press the lever to the exclusion of eating and drinking. The parts of the brain eliciting this behaviour in rate are those where the majority of dopamine neurons are located, namely the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc) (see Figure 1.1). The axons of the dopamine neurons project to areas such as the striatum, nucleus accumbens, and frontal cortex, which have been associated with motivation and goal-oriented behaviour (Schultz et al., 1997).

Further evidence of dopamine's role in the reward process is given by Koob (1992), when discussing how dopamine affects addiction to substances like amphetamine and cocaine. Both drugs have been shown to reduce the reuptake of dopamine, prolonging its effects on target neurons. It is thought that this is the primary reason both for the strong reinforcing effects of taking these drugs and the cravings experienced during withdrawal.

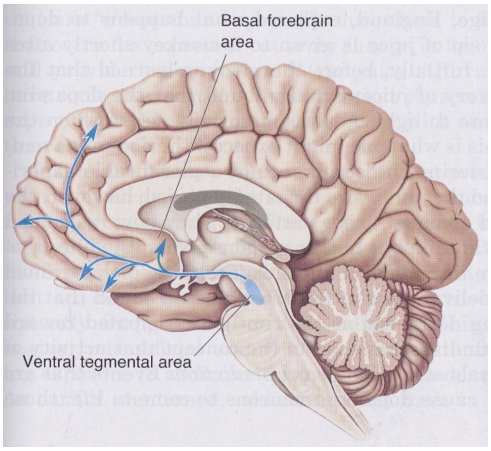


Figure 1.1: Showing the ventral tegmental area (VTA) and the regions to which the dopamine axons project. Figure from Bear et al. (2016).

A particularly significant experiment in developing the theory of the mechanism of dopamine was conducted by Hollerman and Schultz (1998). The experiment required monkeys to touch a lever after a light was turned on and, a short period of time later, the monkey would receive a reward of a drop of juice. The activity of single dopamine neurons was recorded with the monkey being alert, over the course of multiple trials. For the first few trials, the neurons were observed to produce dopamine shortly after the monkey received the juice reward. But after multiple repeats, the neurons produced dopamine just after the light was shown and no longer just after the reward. This shift in dopamine activity mirrors the shift in behaviour of the monkey. Initially, the monkey only shows approach behaviour on the delivery of the juice. But after many repeats, the monkey presses the lever immediately after the light appears, thus now displaying approach behaviour towards the appearance of the light.

In addition, the dopamine neurons displayed reduced firing from the basal rate when the light was shown but no reward was received. The decrease in neuronal activity was observed at the exact time that the reward was expected. In this sense, it can be interpreted that the dopamine activity is not purely linked to reward, but rather to the *prediction error* in the expected reward. Equally, the observation that the magnitude of a dopamine response decreases over time to an expected reward reinforces this interpretation. Also, it is supported by the the fact that there is no negative dopamine response to aversive stimuli.

It was observed by Schultz et al. (1997) that this interpretation of dopamine as a mechanism for recording the error in the prediction of the value of an environmental cue has strong analogies with an area of computer science, namely reinforcement learning.

1.2 Reinforcement learning models

Reinforcement learning (RL) is a type of machine learning with a focus on goal-directed interaction learning.

The reinforcement learning problem is defined by an agent interacting with its environment. The agent uses information from the environment to determine the current state of the system. The agent has a *policy* for taking an action dependent upon its state. The action results in a *reward* for the agent, and also an altered state of the environment. In addition, the agent stores a *value function* which encodes information about the agent's expected reward in any state for a given policy. For every iteration, or trial, the agent can update its policy and value function. The agent's aim is to maximise the total reward it receives over many trials (Sutton and Barto, 1998).

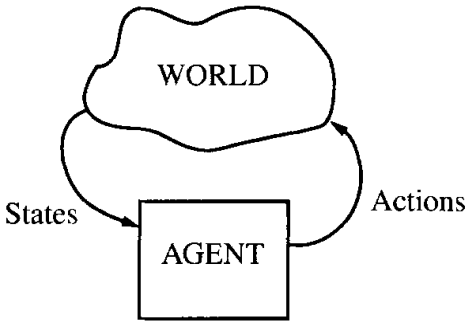


Figure 1.2: A representation of an agent interacting with its environment. Figure from Kaelbling et al. (1998).

A policy, reward and value function are the three basic features of any RL problem. In addition, the agent can define a *model* of its environment, which can define the state of the environment if the state is not fully known. This framework is very abstract and flexible, and can be applied to very varied problems. The field of reinforcement learning studies and develops different ways of solving these problems.

Considering the experiment with monkeys described in section 1.1, it can be seen that one could model the monkey as an agent in a RL problem. The monkey receives information about its environment from the turning on of the light, it is required to take the action of pressing the lever and this results in its being given a reward of juice.

Schultz et al. (1997) recognised that some of the methods used in solving RL problems involved updating the stored value function using a reward prediction error. These methods corresponded closely to their proposed dopamine mechanism being associated with reward-motivated behaviour, and consequently that it might be possible to use existing RL knowledge to create new models of reward-related behaviours. They proved that RL methods do indeed provide good models of behaviour, and the fact the models can so directly be interpreted in behavioural terms is the reason a RL model is implemented here.

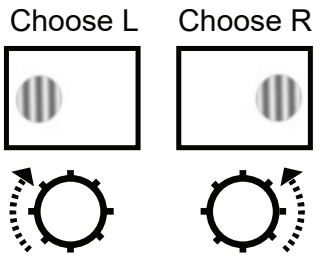
1.3 Experimental method

Below is outlined the task to be modelled using a reinforcement learning approach. The task involves presenting a mouse with a basic two-alternative forced choice task as described in Burgess et al. (2016).

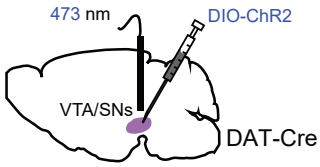
The mouse was head-fixed with its front paws placed on a steering wheel, with three screens positioned around it. First, the mouse was required to hold the steering wheel still in order to initiate the trial. At the onset of each trial a stimulus appeared on either the left or right screen, or neither if the contrast value was 0. An audible click signalled to the mouse that the trial had started. The animal was then required to move the steering wheel in order to transfer the stimulus on to the central screen. The movement of the wheel was coupled to the the movement of the stimulus so that if the stimulus appeared on the left, the mouse was required to move the wheel to the right and the opposite was true for a stimulus on the right. This was to simulate the mouse's moving so as to orient its body towards the stimulus (see Figure 1.3a).

If the mouse successfully identified the side of the stimulus, and moved the steering wheel accordingly, the animal received a water reward from a water spout fixed in front of the mouse. For this experiment the water reward was fixed at 2 µL. In cases when no stimulus occurred, the mouse received a reward 50% of the time.

For some trials the mouse received a reward via phasic optogenetic stimulation of the midbrain dopamine neurons in addition to the water reward. The optogenetic stimulation was achieved by injecting a virus into the area of the brain which required stimulation, here the VTA and the SNc, where the majority of the dopaminergic cell bodies are located (see section 1.1). The virus genetically altered the dopamine neurons so that they became light sensitive. For these cells that meant that they now produced dopamine in the presence of light (Cardin et al., 2010). The virus used here contained Cre-dependent Channelrhodopsin-2 (ChR2), and the mice were of the $DAT^{IREScre}$ variety. After the experiment, it was confirmed using immunochemistry that 77% of the dopamine neurons became light-sensitive and also that very few cells of other types were affected. For more precise details, see Burgess et al. (2016). During the experiment, an optical fibre was inserted into the mouse's brain just above the VTA (see Figure 1.3b). A laser sent multiple (between 6 and 10) very short bursts of light for, in total, not more than 0.5s. This resulted in the stimulation of dopamine production. It has been shown that optogenetic stimulation of dopamine neurons in this manner is sufficient to alter behaviour in mammals (Tsai et al., 2009).



(a) Demonstration of the required movement of the wheel by the mouse at the appearance of the stimulus.



(b) The mouse's brain is injected with a virus containing ChR2, to genetically modify dopamine neurons. An optical figure is placed above the VTA/SNc so that light can stimulate a dopamine reward.

Figure 1.3: Figures demonstrating elements of the experimental set-up. Figures by A. Lak.

The dopamine stimulation was given as a reward in a block structure and in an asymmetric manner. That is, during one block of trials, dopamine stimulation was given as a reward *only* after, say, rightward successful choices. Then during the next block, dopamine was given after only leftward successful trials. Water reward was given after all successful trials. If the mouse moved incorrectly, or failed to move, the animal received no reward and was given a 2s timeout accompanied by a noise burst. During a session a mouse would perform between 300 and 1200 trials. The blocks were typically between 50 and 400 trials long, meaning between 2 and 4 blocks

were conducted during a session.

6 mice were trained to perform this task. During training only water was given as a reward. The training typically took 5 weeks for the animals to reach their highest rate of performance. This rate of performance varied across mice, but with most choosing correctly on at least 80% of trials.

2 Method

2.1 Model implementation

The model implemented is a Partially Observable Markov Decision Process (POMDP) with a belief state, following a process by Lak et al. (2017). The aim of using the model is to simulate the behaviour of a mouse over many trials of the experimental procedure described in Section 1.3. Formalism and further information about POMDPs can be found in Kaelbling et al. (1998).

This model requires a sequence of stimulus contrast values as input. Each contrast value represents a single trial. The model outputs a sequence of the agent's actions (its choosing left or right) in response to each stimulus. The agent aims to learn to choose the optimal action to maximise the total reward it receives, by using reinforcement learning methods described in Section 1.2. It is hoped that this kind of learning strongly mirrors the learning of mammals like mice, and due to the evidence provided in Section 1.1 it is believed that this hope is reasonable.

For each trial, the agent receives some stimulus s, where s is chosen at random from a discrete set of values $\{s_i\}$ with $s_i \in [-0.5, 0.5]$. The value of s represents the contrast of the stimulus: the greater the value of $|s|$, the higher the contrast, and consequently the more easily the stimulus is perceived. When $s < 0$ the stimulus appears on the left, when $s > 0$ the stimulus appears on the right and when $s = 0$ no stimulus appears. Therefore a correct action C is now defined for each contrast value s, where

$$C(s) = \begin{cases} \mathrm{L} & s < 0\,, \\ \mathrm{R} & s > 0\,. \end{cases}$$

When $s = 0$, we define C for that trial to be either L or R with a probability of 0.5.

Due to the lack of clarity in the stimulus, the model assumes that the agent does not perceive the exact value of the stimulus, but a noisy estimate $\hat{s}$. $\hat{s}$ is sampled from a normal distribution with mean of the actual contrast value s and variance σ^2, with σ being one of the parameters of the model. Formally, $\hat{S} \sim \mathrm{N}(s, \sigma^2)$.

The agent forms a belief about the value of the stimulus. This belief comprises a distribution over the range of the stimulus values. Here this is denoted as $B(s, \hat{s}; \sigma) = \mathrm{P}\,(s \mid \hat{s})$, and $B \sim \mathrm{N}(\hat{s}, \sigma^2)$. The belief is defined thus because this model is a simplification of a Bayesian approach, where the belief distribution is updated after each trial as described by Whiteley and Sahani (2008).

Thus, the agent's belief of the stimulus being on the left is $b_{\mathrm{L}}(\hat{s}) = \mathrm{P}\,(B < 0 \mid \hat{s})$ and the belief of the stimulus being on the right is $b_{\mathrm{R}}(\hat{s}) = \mathrm{P}\,(B > 0 \mid \hat{s})$. So the agent's belief of the correct choice based upon its perceived stimulus value is

$$\hat{C} = C(\hat{s}) = \begin{cases} \mathrm{L} & b_{\mathrm{L}} > b_{\mathrm{R}}\,, \\ \mathrm{R} & b_{\mathrm{L}} < b_{\mathrm{R}}\,. \end{cases}$$

In the rare case that $b_{\mathrm{L}} = b_{\mathrm{R}}$, it could be assumed that $C(\hat{s})$ would be defined to be either L or R with a probability of 0.5, as above for $C(s)$.

Now the agent must make a choice, and decide upon an action. Here, the agent could simply choose according to its perceived value of the stimulus and choose left whenever $b_{\mathrm{L}} > b_{\mathrm{R}}$ and right whenever $b_{\mathrm{R}} > b_{\mathrm{L}}$. However, for this task the animal is being rewarded unevenly for left and right choices. As a result of this it is possible that the agent develops a preference for either left or right choices due to its having received larger rewards on that side. In order to incorporate the value of each choice based upon the agent's expected reward, Q-values are introduced into the model.

Q-values store an estimate of the agent's expected reward for a given action, i.e. the *value* of a choice. $Q_{\hat{C}A}$ is defined to be the value of taking action A given the stimulus has defined the choice $\hat{C} = C(\hat{s})$, with $A, \hat{C} \in \{\mathrm{L}, \mathrm{R}\}$. To give an example, this means that Q_{RL} would represent the value of taking action left given that the agent believes the stimulus is on the right.

For the first trial, Q_{LL} and Q_{RR} are initialised at 1 and Q_{RL} and Q_{LR} at 0. In addition, the Q-values are reset to these same values for the first trial of each day, as the animal appears to have no memory about rewards it has received on previous days of the experiment.

Now, the overall values of taking actions left and right are

$$Q_A = b_{\mathrm{L}} Q_{\mathrm{L}A} + b_{\mathrm{R}} Q_{\mathrm{R}A} \ ,$$

where $A \in \{\mathrm{L}, \mathrm{R}\}$. The agent makes a choice about which action to take according to

$$A = \mathrm{argmax}\left\{Q_{\mathrm{L}}, Q_{\mathrm{R}}\right\} \ .$$

The agent receives a reward R based upon its decision. For this task, the agent's reward varies according to the current 'reward block'. If the current trial is in a reward block where

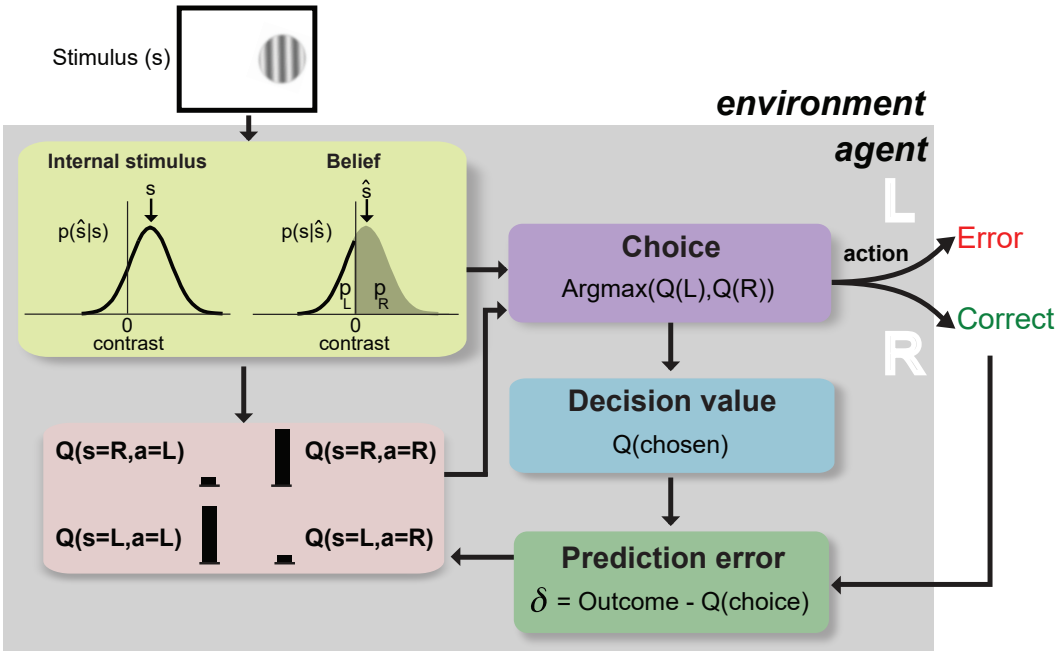


Figure 2.1: Schematic illustrating the model implemented. (Figure by A. Lak, adjusted from Lak et al. (2017).)

only successful rightward choices are being rewarded with dopamine stimulation, then when the agent chooses *left* correctly, the reward R is defined to be 1, to represent the water reward. In the same block, when the agent chooses *right* correctly, the reward is defined to be $(1 + x)$, where x represents the agent's value of the DA stimulation. x can then be interpreted as how much the agent values the DA stimulation in comparison with the water reward. This x value is another parameter of the model. When an incorrect choice is made, $R = 0$.

The difference between the agent's prediction about its reward and the actual outcome (i.e. prediction error) is defined as

$$\delta = R - Q_A \ .$$

If the agent receives more reward than it is expecting, i.e. if $R > Q_A$, then δ will be positive. Equally, δ will be negative if the agent receives less reward than it expects. With this prediction error, the Q-values are updated according to

$$Q'_{\hat{C}A} = Q_{\hat{C}A} + \alpha \, \delta \, b_{\hat{C}} \ ,$$

where $Q'_{\hat{C}A}$ is the updated value of $Q_{\hat{C}A}$ and α is the *learning rate* of the agent, the third parameter of the model. Note that for each trial only two of the four Q-values are updated, $Q_{\mathrm{L}A}$ and $Q_{\mathrm{R}A}$. Also note that, as both α and $b_{\hat{C}}$ are greater than zero, the sign of δ determines whether $Q_{\hat{C}A}$ increases or decreases.

The newly calculated Q-values are retained and are then used to make a choice during the next trial. So this process is iterative and the updating of the Q-values corresponds to the agent 'learning'.

2.2 Fitting the model to data

For any sequence of stimulus values $\{s_t\}$, where $t = 1, ..., N$ represents the trial number, a mouse 'outputs' a sequence of actions $\{a_t\}$. Equally, for the same sequence of stimulus values $\{s_t\}$ and a fixed set of parameters $\{\alpha, \sigma, \text{DA value}\}$, the model outputs its own sequence of actions $\{A_t\}$. The aim now is to choose the set of parameter values $\{\alpha, \sigma, x = \text{DA value}\}$ which results in the model's best approximating the mouse behaviour, so that $\{A_t\}$ most resembles $\{a_t\}$.

Firstly, it should be noted that the sequences $\{s_t\}$ and $\{a_t\}$ are fixed data collected from experiment. Therefore the stimuli sequences are those to be fed to the model as input.

Secondly, the sequences of stimuli are specific to each mouse. This is due to the variation in ability between mice. Mice that were better at the task were only shown stimuli with contrast values between -0.2 and 0.2, whereas other mice were shown stimuli with contrast values between -0.5 and 0.5. In addition, some mice were capable of performing many more trials in one session than others. As a consequence, it is expected that the model parameter values $\{\alpha, \sigma, \text{DA value}\}$ will differ between mice.

Thirdly, it is noted that a sequence of model actions $\{A_t\}$ is not deterministic. That is, if run multiple times on the same sequence of stimuli and with the same parameter values, the sequence of actions the model produces will be different each time. This is expected due to the probabilistic nature of the model's perceived stimulus value and belief state during each trial. However, due to the large number of trials being run, when viewed on a macroscopic level, the model's behaviour will vary very little between runs.

For a single set of mouse data, (i.e. for one sequence $\{s_t\}$ and the corresponding $\{a_t\}$), in order to find the set of model parameters to best match the mouse data, the model was run with many different parameter values. These parameter values were fixed at the beginning of each run, with a 'run' being defined as the passing to the model of a sequence $\{s_t\}$ and its output the corresponding $\{A_t\}$. At the end of each run the following were calculated:

- $\{s_i\}$, the vector of possible contrast values used as stimuli (typically around 6),
- $\{n_i\} = \{n(s_i)\}$, the number of times each contrast value was used as a stimulus,
- $\{p_{\mathrm{m},i}\} = \{p_{\mathrm{m}}(s_i)\}$, the fraction of rightward choices made by the *mouse* for each contrast value,
- $\{p_{\mathrm{r},i}\} = \{p_{\mathrm{r}}(s_i)\}$, the fraction of rightward choices made by the *model* for each contrast value.

Then the negative log-likelihood (NLL) of the parameter values used being the correct ones given both the mouse and the model choices was calculated as

$$-\sum_i n_i \big[\, p_{\mathrm{m},i} \ln(p_{\mathrm{r},i}) + (1 - p_{\mathrm{m},i}) \ln(1 - p_{\mathrm{r},i}) \,\big] \ .$$

The NLL was calculated for each set of parameter values and the parameters values giving the minimum value for the NLL were then deemed to be the optimal values for matching the mouse and model behaviour. This method of minimising the negative log-likelihood is equivalent to maximum likelihood estimation (MLE).

3 Results

The model was fitted to data successfully and results are plotted below for two mice, with further results in Appendix A.

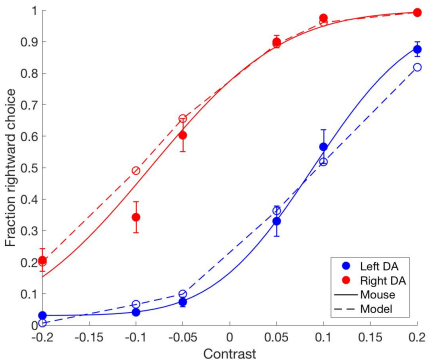
Figure 3.1a shows that mouse behaviour was significantly altered by the asymmetric dopamine (DA) stimulation. The mouse shows a significant preference for choosing the side which offers a DA reward. The mouse shows such a preference that it makes incorrect decisions about the side of the stimulus approximately 20% of the time even for the clearest contrast levels.

Figure 3.1a also shows that the model mirrors the mouse behaviour well. However, firstly, this does not necessarily mean that the model is a good predictor of mouse behaviour. The method by which the model was fitted to the mouse data effectively ensures that the two psychometric functions will be relatively similar. It is possible that the psychometric functions of the mouse and the model are very similar, but that the learning rates are very different. Secondly, the error function is generally a better fit to the mouse data than the POMDP model. This could suggest that the method by which the POMDP is fitted could be improved.

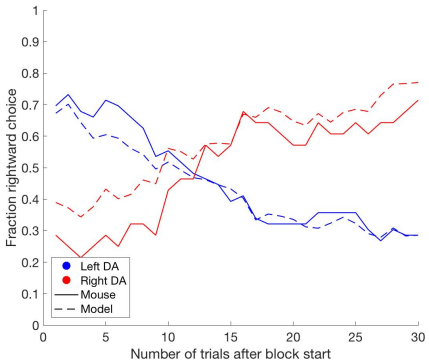
Considering figures 3.1b and 3.1c, it can be seen that the model also mirrors mouse behaviour well on a trial by trial basis. From figure 3.1c, it can be seen that during each block, both the mouse and the model show a distinct preference for the side where a DA reward is received. It can also be noted that the magnitude of the oscillations in the mouse behaviour are larger than those in the model's behaviour, which suggests that the model is less influenced by the effect of the DA stimulation. This is consistent with what is shown by figure 3.1a, that the mouse makes

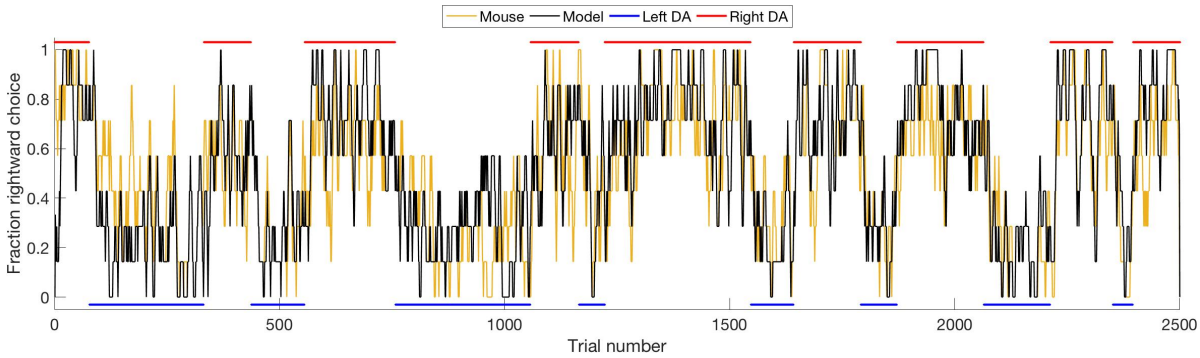
Mouse SS040.

(a) Psychometric data.



(b) Average behaviour at the start of a block.





(c) Trial by trial actions of mouse and model.

Parameters	
α	0.2
σ	0.1
DA val.	4.5

Total trials	2500
No. of contrast levels	6
Median trials per contrast	451

Figure 3.1: Mouse SS040. Plots where the model was run with the parameter values giving the minimum negative log-likelihood of all parameter values tested. Model averages were calculated from 21 runs. **(a)** Plotted are the mouse psychometric data for each contrast value and an error function fitted using MLE to the mice data (the solid line). In addition the POMDP model data are plotted (dashed line). The model data were collected as the mean choice for a specific trial over many runs. Error bars show the standard error for each mouse data point across trials. **(b)** Smoothed trial by trial actions of the mouse (solid) and model (dashed). Plotted are the average choices of one mouse over many blocks, specifically at the beginning of the block. Only blocks not at the start of the a session are included. In blue are choices made in a block where dopamine was given after correct left choices, and in red are choices made in a block where dopamine was given after correct right choices. The model data were collected as the mode choice for a specific trial over many runs. **(c)** Smoothed trial by trial actions of the mouse (yellow) and the model (black). The model data were collected as the mode choice for a specific trial over many runs. The blue and red lines, represent the current block, where blue represents DA reward after left choices and red represents DA reward after right choices.

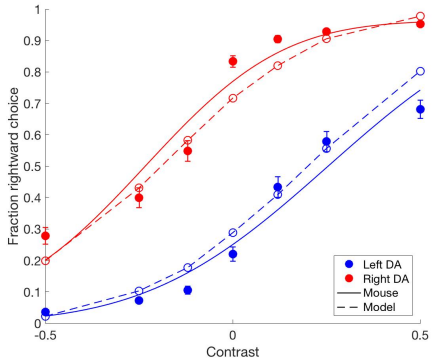
more correct decisions than the model, with the exception being for DA reward after leftward choices for contrast values of 0.05.

Figure 3.1c also shows that both the mouse and the model respond within a very few trials to a change in the DA reward structure. This is particularly evident in figure 3.1b. This figure plots the average choices of both the mice and the model early in a block, i.e. shortly after there has been a change in the reward structure. And because choices are only included for blocks that are not the first in the day, the DA reward will have switched from being after left/right choices to right/left choices (for the red/blue lines respectively). For the model, it takes approximately 10 trials from the start of a block for the behaviour to significantly qualitatively change, and approximately 12 trials for the mouse. The fact that these numbers are quite similar means that the learning rates for the mouse and the model are also similar.

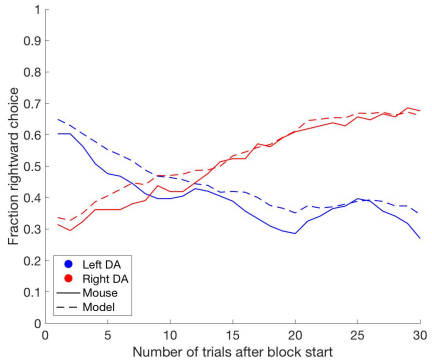
Comparing figures 3.1 and 3.2, it can be seen that there are some significant differences in behaviour between mice. Firstly, it should be noted that mouse SS040 in 3.1 was only shown stimuli with contrast values between -0.2 and 0.2, whereas mouse ALK017 in 3.2 was shown contrast values between -0.5 and 0.5. This suggests that mouse SS040 was much better at

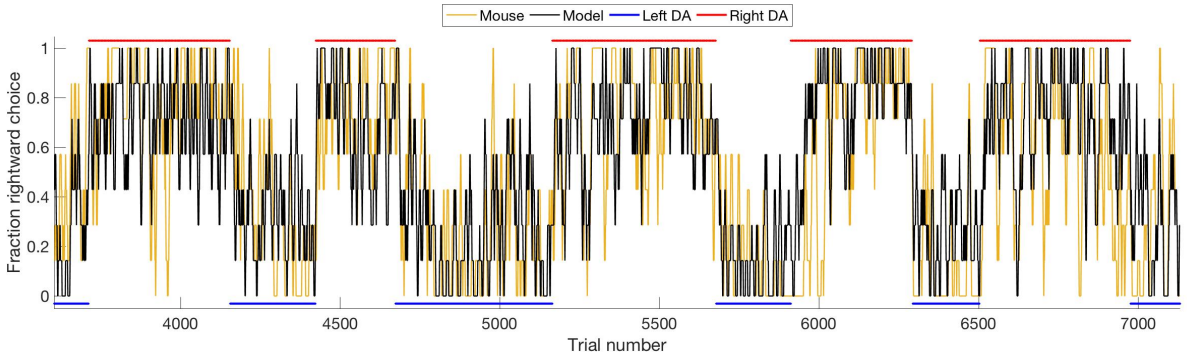
Mouse ALK017.

(a) Psychometric data.



(b) Average behaviour at the start of a block.





(c) Trial by trial actions of mouse and model.

Parameters	
α	0.25
σ	0.325
DA val.	1.5

Total trials	10244
No. of contrast levels	7
Median trials per contrast	1503

Figure 3.2: Mouse ALK017. Plots where the model was run with the parameter values giving the minimum negative log-likelihood of all parameter values tested. For more details, see Figure 3.1.

performing the task than mouse ALK017. Also, ALK017 performed more than four times more trials than mouse SS040. However, the number of trials per block remained similar. The psychometric curves and the learning rates are relatively similar for both mice, which is consistent with α values of 0.2 and 0.25 being similar for mice SS040 and ALK017 respectively.

4 Discussion

Figures 3.1 and 3.2 give the parameter values which correspond to the minimum negative log-likelihood values of being the correct parameter values of all parameter values checked. The α parameter values are 0.2 and 0.25 for mice SS040 and ALK017 respectively. Their being similar suggests that the mice have similar learning rates. The absolute value of alpha can be interpreted as a learning rate due to its being a multiplicative factor in the updating of the Q-values. When α is high, the Q-values change by a greater amount from trial to trial. This means that when a qualitative changes in the reward structure occurs, the Q-values will update to reflect this change within a very few trials. The fact that, for both mice, the α value is quite low means that the Q-values will update relatively slowly and the mouse behaviour will remain more steady over time.

The σ parameter values are 0.1 and 0.325 for mice SS040 and ALK017 respectively. This shows some significant difference between the two mice. However, it is quite difficult to directly compare these values, because these mouse were shown stimuli with different ranges of contrast values ($[-0.2, 0.2]$ and $[-0.5, 0.5]$ respectively). The differences in σ values reflect this difference. The absolute value of σ represents the animal's uncertainty both in its perception of the stimulus and in its belief distribution of the correct stimulus value. The higher the value of σ, the greater the animal's uncertainty. We can make the observation that the ALK017 value of σ being higher than that of mouse SS040 indicates that the former was much more uncertain in its knowledge of the stimulus, and consequently performed were on the task (i.e. made fewer correct choices). The accords with collected data as the two mice show similar levels of performance, despite mouse SS040 effectively performing a much harder task.

The parameter values representing the mice's value of the dopamine in comparison with water drop were 4.5 and 1.5 for mice SS040 and ALK017 respectively. This is slightly surprising. One might expect that mouse SS040 having a higher value for the dopamine stimulation would result in that mouse making more decisions seeking the dopamine stimulation, which in turn would result in worse performance in the task. Considering that this behaviour is not witnessed, it is possible that there is some trade-off during the fitting for the parameter values of α and the DA value. Lowering the value of α allows for an increase in the DA value, with the model still outputting similar behaviour.

There are multiple pieces of evidence that suggest that the method used for fitting the model to the data could be improved. Firstly, one would expect that the POMDP model should approximate mouse behaviour better than the error function, but figures 3.1a and 3.2a show that this is not the case. Also, the model makes more incorrect choices than the mice points to an error in the fitting.

It could be argued that the model fitting is not the cause for the error in the results, but rather that the choice of model is the larger issue. However, previous studies (such as Schultz et al. (1997)) have shown significant benefit from using reinforcement learning models. Also, the

theoretical arguments for using such models are very strong. The fact that RL model parameters give such strong behavioural interpretations is another good argument for their continued use. Still, there ways in which the model could be improved. For example, the belief distribution could be updated in a Bayesian manner after each trial as suggested by Whiteley and Sahani (2008).

Acknowledgements

I would like to thank Armin Lak for his supervision during this project. He is also to be credited with collecting the data analysed here, for writing the initial code used to run the model and for producing several of the figures. I would also like to thank Kenneth Harris and Matteo Carandini for their supervision and support.

References

Mark F Bear, Barry W Connors, and Michael A Paradiso. *Neuroscience: Exploring the Brain*. Lippincott Williams & Wilkins, 4th edition, 2016.

Christopher P Burgess, Nicholas Steinmetz, Armin Lak, Zatka-Haas Peter, Adam Ranson, Miles Wells, Sylvia Schroeder, Elina A K Jacobs, Charu Bai Reddy, Sofia Soares, Jennnifer F Linden, Joseph J Paton, Kenneth D Harris, and Matteo Carandini. High-yield methods for accurate two-alternative visual psychophysics in head-fixed mice. *bioRxiv*, 2016.

Jessica A Cardin, Marie Carlén, Konstantinos Meletis, Ulf Knoblich, Feng Zhang, Karl Deisseroth, Li-Huei Tsai, and Christopher I Moore. Targeted optogenetic stimulation and recording of neurons in vivo using cell-type-specific expression of channelrhodopsin-2. *Nature protocols*, 5(2):247–254, 2010.

Jeffrey R Hollerman and Wolfram Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature neuroscience*, 1(4):304–309, 1998.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

George F. Koob. Dopamine, addiction and reward. *Seminars in Neuroscience*, 4(2):139–148, 1992. Milestones in Dopamine Research.

A. Lak, K. Nomoto, M. Keramati, M. Sakagami, and A. Kepecs. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, 27:1–12, 2017.

James Olds and Peter Milner. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology*, 47 (6):419, 1954.

Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.

Hsing-Chen Tsai, Feng Zhang, Antoine Adamantidis, Garret D Stuber, Antonello Bonci, Luis De Lecea, and Karl Deisseroth. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, 324(5930):1080–1084, 2009.

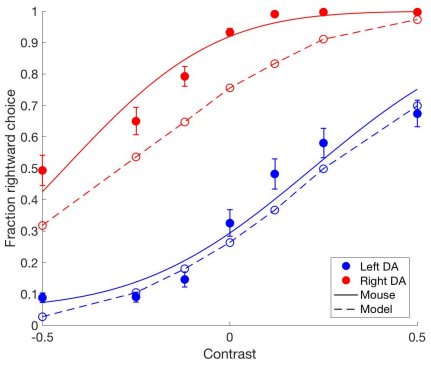
Louise Whiteley and Maneesh Sahani. Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision*, 8(3):2, 2008.

A Additional figures

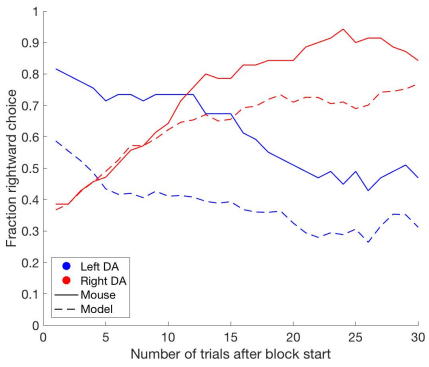
We include here figures representing data for mice other than SS040 and ALK017 as shown in section 3.

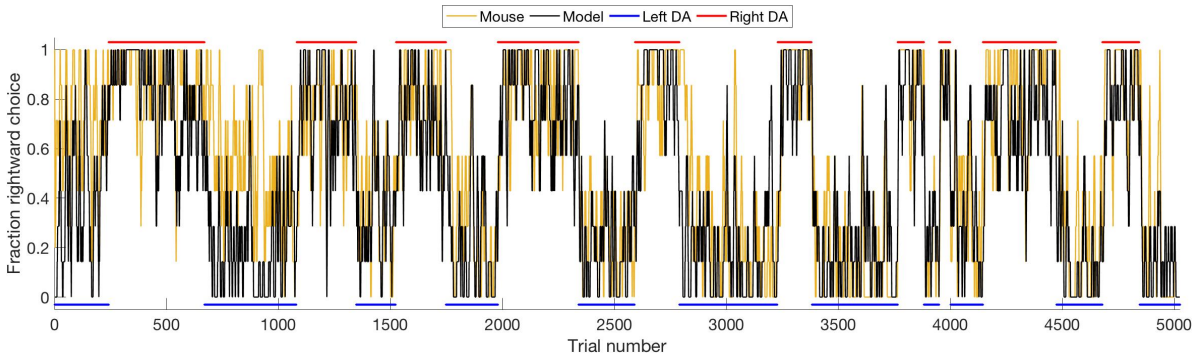
Mouse ALK028.

(a) Psychometric data.



(b) Average behaviour at the start of a block.





(c) Trial by trial actions of mouse and model.

Parameters	
α	0.55
σ	0.35
DA val.	2.5

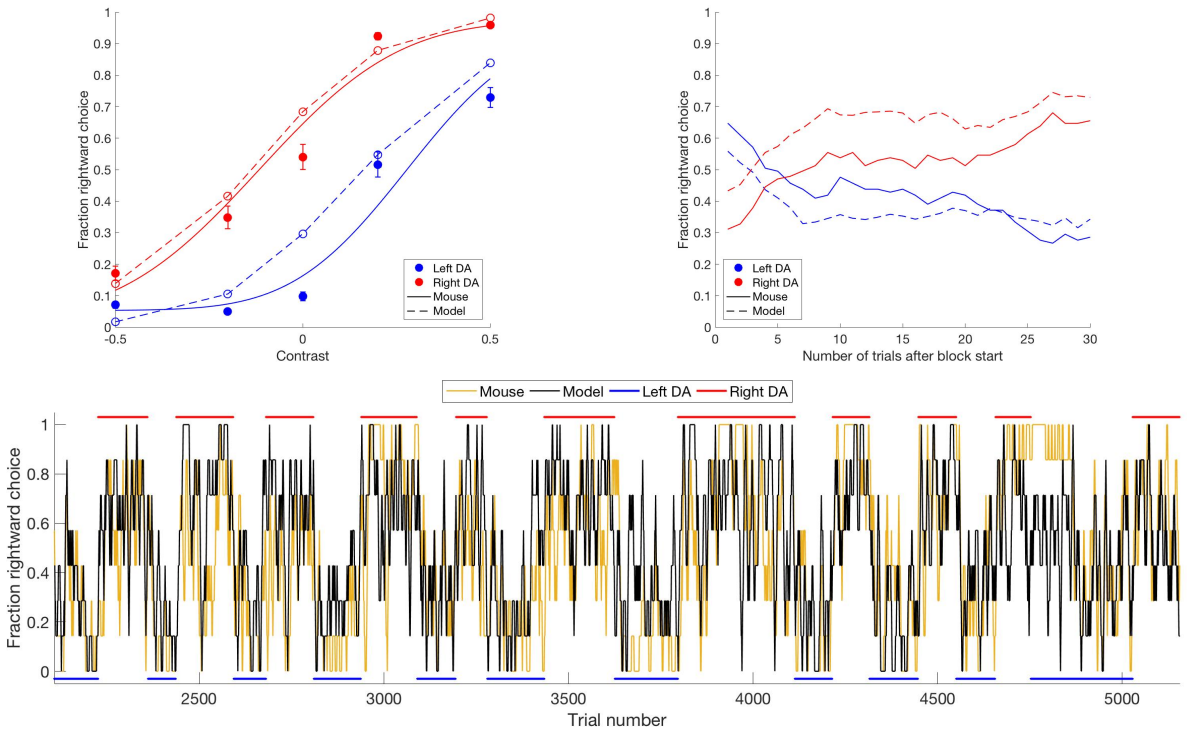
Total trials	5024
No. of contrast levels	7
Median trials per contrast	717

Figure A.1: Mouse ALK028. Plots where the model was run with the parameter values giving the minimum negative log-likelihood of all parameter values tested. Model averages were calculated from 21 runs. **(a)** Plotted are the mouse psychometric data for each contrast value and an error function fitted using MLE to the mice data (the solid line). In addition the POMDP model data are plotted (dashed line). The model data were collected as the mean choice for a specific trial over many runs. Error bars show the standard error for each mouse data point across trials. **(b)** Smoothed trial by trial actions of the mouse (solid) and model (dashed). Plotted are the average choices of one mouse over many blocks, specifically at the beginning of the block. Only blocks not at the start of the a session are included. In blue are choices made in a block where dopamine was given after correct left choices, and in red are choices made in a block where dopamine was given after correct right choices. The model data were collected as the mode choice for a specific trial over many runs. **(c)** Smoothed trial by trial actions of the mouse (yellow) and the model (black). The model data were collected as the mode choice for a specific trial over many runs. The blue and red lines, represent the current block, where blue represents DA reward after left choices and red represents DA reward after right choices.

Mouse SS031.

(a) Psychometric data.

(b) Average behaviour at the start of a block.



(c) Trial by trial actions of mouse and model.

Parameters	
α	1.00
σ	0.25
DA val.	3.0

Total trials	5643
No. of contrast levels	5
Median trials per contrast	1121

Figure A.2: Mouse SS031. Plots where the model was run with the parameter values giving the minimum negative log-likelihood of all parameter values tested. For more details, see Figure A.1.