Homework 5
Systems & Theoretical Neuroscience [SWC]

Due: Monday January 15th

1 Temporal difference learning

1.1

In classical conditioning, reinforcement learning agents learn to predict rewards by keeping a record of the expected reward or value v, and learning the relationship between this value and the stimuli presented u by updating a set of weights w.

$$v = wu \tag{1}$$

To test this learning effect experimentally, you set out to replicate a classic experiment with your pet dog, Ivan. For the first 100 trials, you ring a bell and give Ivan some dog food. Then, you repeat another 100 trials in which you ring the bell, but do not deliver a reward. You then repeat the entire experiment another time.

You hypothesise that Ivan uses the Rescorla-Wagner rule to update his expectation of the reward. The Rescorla-Wagner rule says that an agent updates his weights every trial as follows:

$$w \to w + \epsilon\delta u \tag{2}$$

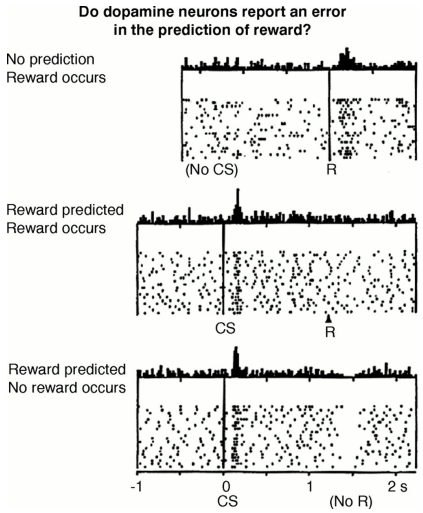
$$\delta = r - v \tag{3}$$

where δ is the difference between the reward and the expected reward, and ϵ is a learning rate taking values between 0 and 1.

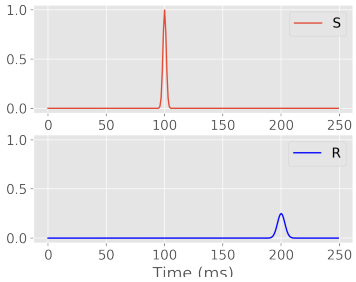
a) Simulate the experiment described above, using $r = 1$ when there is a reward, and $r = 0$ otherwise, $u = 1$ if a stimulus is present, and 0 otherwise, with a learning rate $\epsilon = 0.05$. Plot the weight parameter w over trials. What do you see? How would you expect this curve to differ in your real experiment with Ivan?

After your observation that Ivan can learn the stimulus-reward association, you wonder how different stimuli interact when they are both predictive of reward. To test this, you first associate the bell sound with a food reward. Then, you start adding a flashing light as a second stimulus on every trial, while continuing to administer the reward. When you remove the bell sound to only present the flashing light, you observe that Ivan does not expect the food reward anymore.

b) Can you explain this in terms of the Rescorla-Wagner rule? Simulate how the weights evolve during this experiment (hint: when there are multiple stimuli, you work with vectors of stimuli and weights, and the value equation becomes the dot product $v = \vec{w} \cdot \vec{u}$).



(a) Activity of dopaminergic neurons in the ventral tegmental area (VTA) from a rat during a classical conditioning task. CS: conditioned stimulus, R: reward (Schultz, Dayan & Montague, 1997).



(b) Stimulus and reward example for your simulations.

Figure 1

In addition to learning associations between stimuli and rewards and punishments, animals can learn to predict the time at which rewards will arrive within a trial. You are curious how this works in the brain, but rather than subjecting Ivan to an invasive experiment you pick up an old copy of *Science* and you find the interesting result from Schultz et al. shown in figure 1a. Before learning, dopaminergic neurons in the midbrain fire when the reward is administered. After learning, they fire at the onset of the conditioned stimulus.

c) Qualitatively, what are these neurons coding? Why would this be useful for learning?

We model this phenomenon by assuming that, at each time point t, the rat is estimating the *expected total future reward* from t until the end of the trial, at time T. As before, we model this estimate as a linear function of the observed stimuli $u(t)$, now expressed over time, where we take the full stimulus history of the current trial into account:

$$v(t) = \sum_{\tau=0}^{t} w(\tau)u(t-\tau) \tag{4}$$

To get a good estimate of the expected future reward $v(t)$, the rat must therefore learn an appropriate set of weights $w(\tau)$.

Note that since $v(t)$ denotes the expected future reward, it can be expressed recursively as:

$$v(t) = r(t) + v(t+1) \tag{5}$$

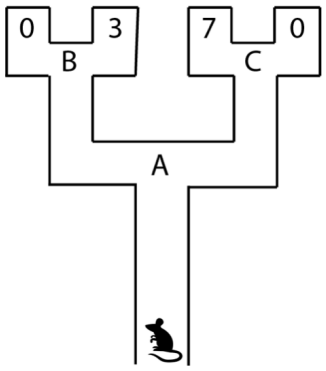


Figure 2: The maze task. The possible states are A, B and C, and the possible actions at each state are right or left. The rat starts each trial at state A, and its goal is to maximise its total future reward.

So, if the reward $r(t)$ received at timestep t doesn't satisfy this equation, the weights defining $v(t)$ (and $v(t+1)$) must be adequately modified. This results in the *temporal difference learning* algorithm (Sutton, 1988):

$$\begin{aligned} \delta(t) &= r(t) + v(t+1) - v(t) \\ w(\tau) &\rightarrow w(\tau) + \epsilon\delta(t)u(t-\tau) \end{aligned} \tag{6}$$

where $\delta(t)$ is called the *temporal difference prediction error*.

d) Using equations (4) and (6), implement a temporal difference learner that qualitatively reproduces figure 1a. You can take $u(t)$ and $r(t)$ to be Gaussian bumps centred at 100 ms and 200 ms, respectively, as shown in figure 1b. Plot the value $v(t)$ and the prediction error $\delta(t)$ for all t before and after learning with a conditional stimulus. *Hint: equation 6 contains the value of the next time step $v(t+1)$, which you cannot compute at time t. You can only update $v(t)$ at time t+1.*

e) What happens if, after learning, you show the stimulus but not the reward?

1.2

The temporal difference learning algorithm that you implemented in question 1.1 can be used not only for learning stimulus-reward associations, but also to learn to take the optimal actions when performing a task. In this question, we will explore a specific temporal difference strategy called actor-critic learning.

Consider the maze task depicted in figure 2. The goal of the actor-critic learner will be to learn an optimal *policy* $\pi(s, a)$, which describes which actions to take in which states. Recall that the

value $v(s)$ of state s is the expected total future reward in a trial when starting from that state. Because this future reward depends on the actions taken, we take the expectation with respect to our policy: $v(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} r_t\right]$.

a) What are the true values of states A, B and C under a random policy (i.e. when left and right are chosen randomly, regardless of state)?

As the agent walks through the maze, any action a might take it from state s to state s'. As in question 1.1, the temporal difference learner can update its value of the current state using the temporal difference update rule:

$$\begin{aligned} v(s) &\rightarrow v(s) + \epsilon\delta \\ \delta &= r_a(s) + v(s') - v(s) \end{aligned} \tag{7}$$

b) Implement an agent that takes random left or right actions at each time step, while learning the value of each state using the learning rule in equation 7. Does your algorithm converge to the correct true values?

c) The part of the algorithm you just implemented is called the critic. Why do you think it is called that way?

We now need a separate algorithm to use these evaluations to actually learn a policy. We define *action values* $m_a(s)$ as the value of taking action a while being in state s. This is given by the actual reward received upon taking that action plus the rewards that are expected to follow: $m_a(s) = r_a(s) + v(u')$.

d) What is the action value $m_{Right}(B)$ in the maze task?

To go from action values to the probability of taking an action, we use the *softmax* equation:

$$P_\pi(a|s) = \frac{exp(\beta m_a(s))}{\sum_{a'=1}^{N_a} exp(\beta m_a(s))} \tag{8}$$

where $P_\pi(a|s)$ denotes the probability of taking action a while in state s, according to our current policy π (implicitly defined by the action values).

e) What happens to the distribution of action probabilities if you vary β? What does this mean in terms of the behaviour of the learning agent?

It turns out that we can use the same temporal difference prediction error δ to update the action values. The policy improvement or actor learning rule is then:

$$m_{a'}(s) \rightarrow m_{a'}(s) + \epsilon(\delta_{aa'} - P_\pi(a'|s))\delta \tag{9}$$

where $\delta_{aa'}$ is the Kronecker delta function, which is 1 if $a = a'$ and 0 otherwise.

f) Write down the update for $m_{a'}$ when the current action $a = a'$ and when $a \neq a'$. Why are these reasonable updates for each of these cases?

g) Implement the full actor-critic algorithm to learn the maze task. Plot the probabilities of going left or right at each state, before and after learning. You can use a learning rate $\epsilon = .5$, and $\beta = 1$. Set the initial values and action values all to zero, and select actions using the softmax distribution. Does your artificial rat learn the optimal policy?

h) Actor-critic learning is considered to be a biologically plausible implementation of reinforcement learning, partly because the prediction error δ is encoded by dopaminergic neurons in the midbrain (figure 1a). What neural structures do you think could perform the actor and critic role, respectively?

2 The role of dopamine in the brain

As seen in the previous question, dopaminergic neurons in the ventral tegmental area and substantia nigra show response properties consistent with a role in reinforcement learning, specifically in encoding prediction errors.

However, other theories - namely, incentive salience theory - suggest that these firing properties can be equally well explained by dopamine attributing Pavlovian incentive value to cues that signal reward. Under incentive salience theory, dopamine makes stimuli desirable by acting as a motivational signal.

While not necessarily contradictory to the prediction error hypothesis, it is challenging to dissociate these two hypotheses because predictive and motivational properties of reward-associated cues are often acquired together. In this question, we will assess two papers addressing these two hypotheses.

Our first paper studies the difference between two breeds of rats that show different behaviours during classical conditioning paradigms. If a conditioned stimulus (CS) is presented immediately before unconditioned stimulus (US) delivery at a separate location, some animals approach and engage the CS itself and go to the location of food delivery only upon CS termination. This conditioned response (CR), which is maintained by Pavlovian contingency, is called 'sign-tracking' because animals are attracted to the cue or sign that indicates impending reward delivery.

Other individuals do not approach the CS, but during its presentation engage the location of US delivery, even though the US is not present until CS termination. This CR is called 'goal-tracking'.

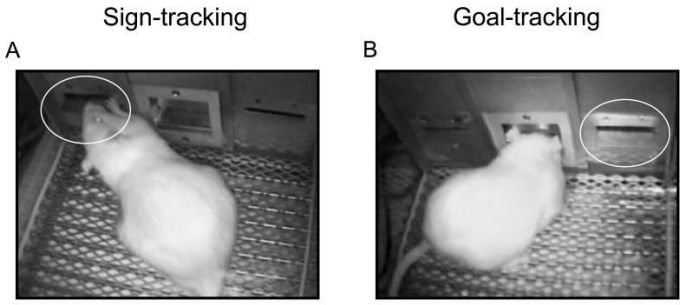


Figure 3: Rats of different breeds, having undergone identical training, show distinct behavioural phenotypes (sign- and goal-tracking)

Surprisingly, it turns out you can actually breed rats to be 'goal-trackers' or 'sign-trackers', sug-

gesting that the circuitry involved in learning might differ between them in some fundamental way. Figure 3 shows the difference in behaviour between rats selectively bred for differences in locomotor responses to a novel environment, with high responders to novelty (bHR rats) consistently learn a sign-tracking CR but low responders to novelty (bLR rats) consistently learn a goal-tracking CR.

a) If dopamine encodes a prediction error signal as described above, how would you expect VTA activity during learning to differ between the two breeds of rat?

Figure 4 shows fast-scan cyclic voltammetry recordings of dopamine levels in the nucleus accumbens, a site previously shown to be important for the acquisition and/or performance of Pavlovian conditioned approach behaviour.

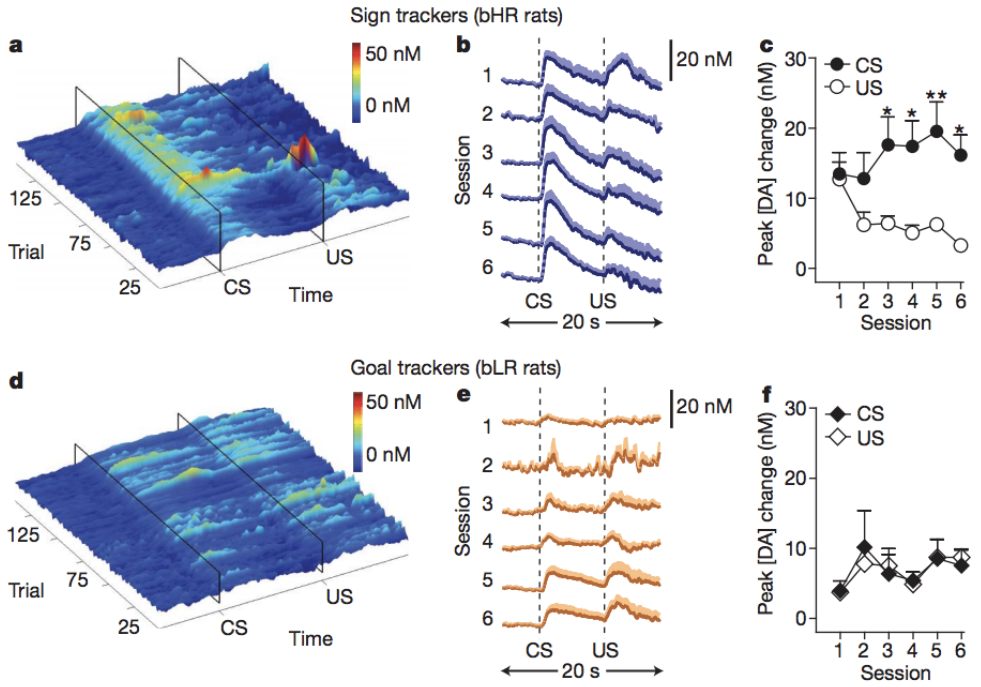


Figure 4: Phasic dopamine signaling in response to CS and US presentation during the acquisition of Pavlovian conditional approach behaviour in bHR and bLR rats. Dopamine concentration [DA] in the nucleus accumbens was measured over six days of training. Single trial [DA] traces for two individual animals are shown in a,d, with session (day) averages (normalized) shown in b,e. c, f, show the change in peak amplitude of [DA] observed in response to CS and US for each session of conditioning.

b) How does the dopamine signal in the nucleus accumbens differ between sign and goal trackers over learning?

c) Does this match expectations from the Rescorla-Wagner rule? Speculate on reasons for any

divergence from the hypothesized dopamine prediction error signals.

d) What are the limitations of this study in addressing whether dopamine acts as a prediction error or incentive salience signal?

In our next set of experiments, the authors utilize optogenetic stimulation of dopamine neurons in the ventral tegmental area to address whether or not dopamine encodes prediction errors.

They start with a paradigm, shown in figure 5, in which an association is first made between a cue and a reward, and then a second cue is introduced which also precedes a reward. The test trials then test whether the animal has learned to associate the second cue with reward.

e) What does figure 5 show about the response to the second stimulus (shown as 'X')? What type of paradigm is this?

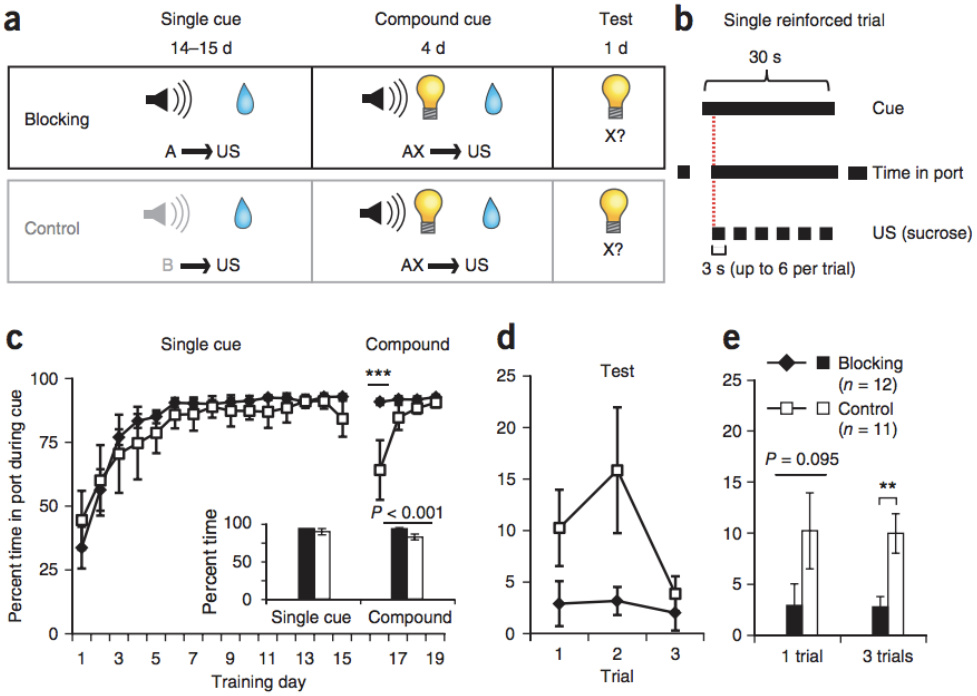


Figure 5: Behavioral data. (a) Experimental design of the task. A, cue A; X, cue X; AX, compound presentation of cues A and X; US, unconditioned stimulus. (b) During reinforced trials (single cue and compound cue sessions), sucrose delivery was contingent on reward port entry during the 30s cue. As long as the rat stayed in the reward port, sucrose was delivered repeatedly for 3s at a time with 2s timeouts in between each delivery. Up to six sucrose rewards could be earned per trial if the rat stayed in the reward port constantly. (c) Performance across all single cue and compound training sessions. (d) Performance during test trials. (e) Performance in the first test trial and the average over all three test trials.

To address whether dopamine acts as a prediction error signal, the authors now stimulate dopaminergic neurons in the VTA during this paradigm, as shown in figure 6.

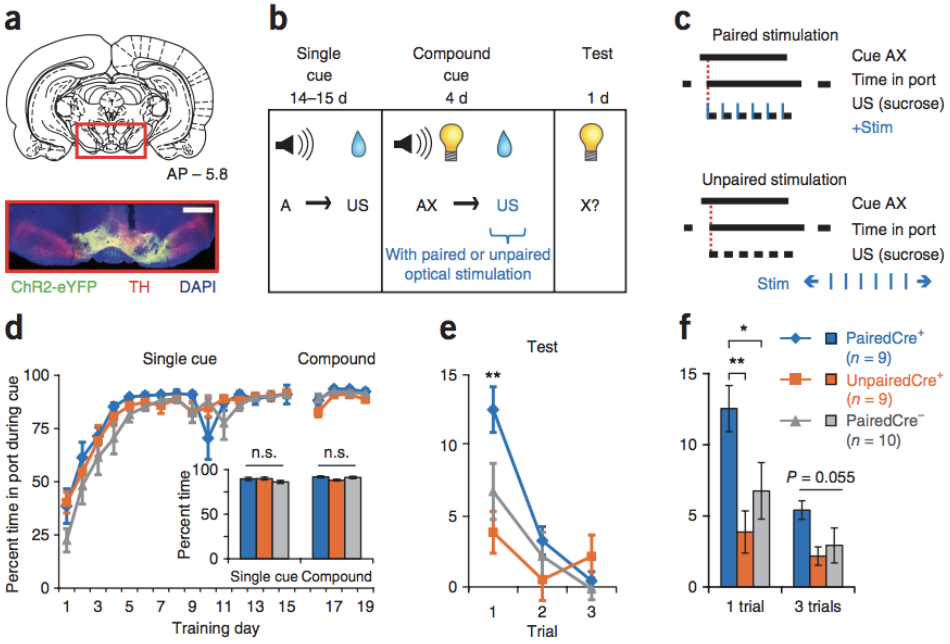


Figure 6: (a) Histology of virus injection and optical fiber placement in VTA. (b) Experimental design as in fig 5 with optogenetics. (c) During training, optical stimulation was synchronized with sucrose delivery in Paired condition, but not in Unpaired condition. (d,e,f) same as (c,d,e) in fig 5 for Paired (blue) and Unpaired (orange) conditions, as well as a control group who received the Paired stimulation protocol but was not genetically modified for the optogenetic stimulation to have an effect

f) What does the data in figure 6 suggest about the role of dopamine signaling in learning? Explain your answer in terms of the Rescorla-Wagner rule.

g) Do these experiments (fig 4 and fig 6) conflict? If so, how?

h) Given these two sets of experiments, what would you conclude about the role of dopamine in the brain?

3 Innate mechanisms of aggression and mating behaviour

In addition to learning from experience, animals can perform many types of behaviour without any learning at all. In the experiments discussed in this question, researchers take advantage of highly stereotyped innate behaviours of mice to dissect the circuitry involved in mating and fighting. Their paradigm is simple: introduce an intruder mouse to the home cage of another mouse. If the intruder

is male, then the resident mouse will attack, but if the intruder is female, then the resident mouse will attempt to mate instead.

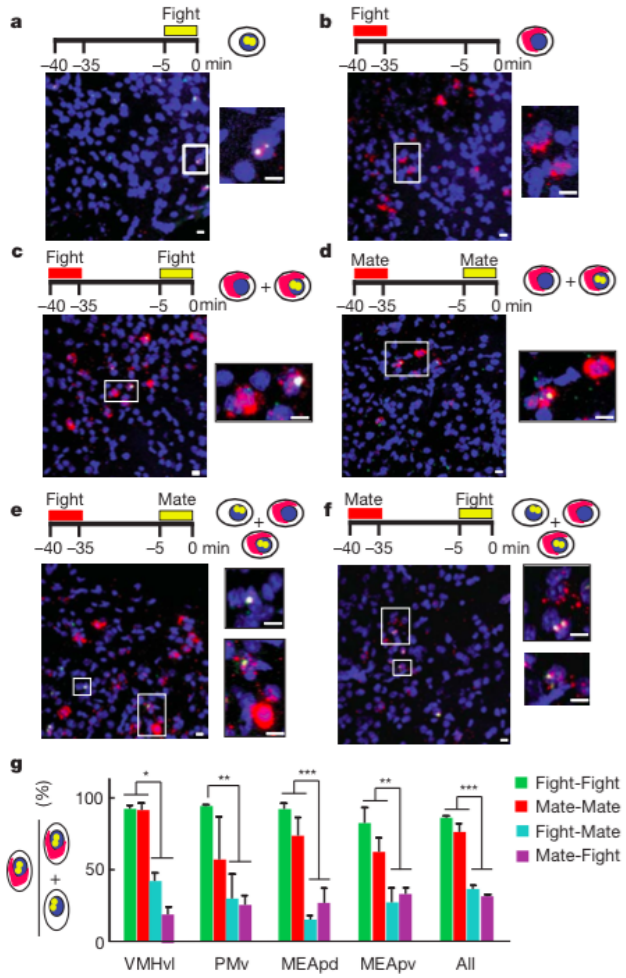


Figure 7: Example images of brain slices after different combinations of behaviour. Red and yellow bars and schematics illustrate the color and localisation of the staining relating to each type of activity. Yellow nuclear staining relates to the most recent behaviour (5 minutes before sacrifice) and red cytoplasmic staining indicates activity in the behaviour 40 minutes before sacrifice. Panel (g) summarises their finding for 5 regions for each behavioural pairing (i.e. fight-mate means that a fighting behaviour took place 30 minutes before sacrifice while a mating behaviour occurred 5 minutes before sacrifice). This is expressed as a ratio of the cells active in both behaviours to those active in both plus those active only in the second behaviour.

To find target regions that might be involved in aggression, they applied a technique called cellular compartment analysis of temporal activity by fluorescence in situ hybridization (catFISH). This

method takes advantage of the localisation of immediate early gene (IEG) expression (genes that are selectively upregulated following neuronal activity) to screen for activity relating to two behaviours simultaneously. cFos RNA and protein have different spatial and temporal expression profiles. Neuronal activity less than 5 minutes prior to fixing the brain can be detected through cFOS RNA in the nucleus, while activity around 30 mins prior will be seen as cFOS protein expression in the cytosol. By timing their behavioural tests appropriately, they distinguish between activity relating to each. Their results are shown in figure 7.

a) What can we infer from the expression patterns in figure 7?

b) Given these results, why do you think the experimenters then chose to focus on the ventrolateral part of the ventromedial hypothalamus (VMHvl) (rather than e.g. AII) to study aggressive behaviour?

c) What are the limitations of this catFISH-based approach?

d) discuss the advantages and disadvantages of using innate behaviour to study the brain.

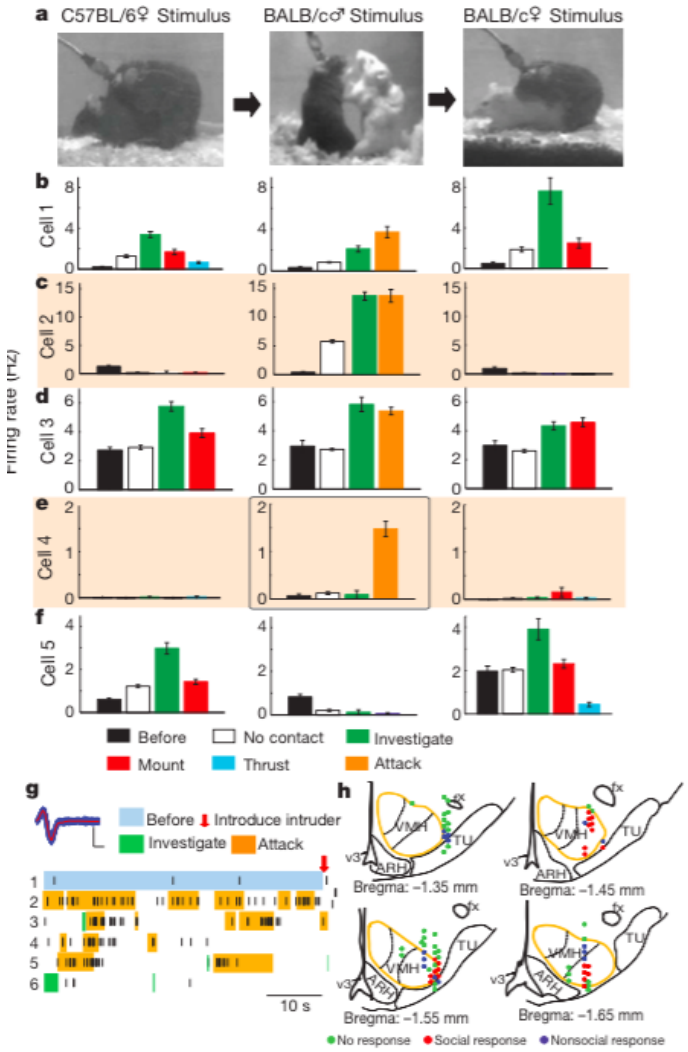


Figure 8: Images of each behaviour exhibited by a C57BL6 mouse to a female C57BL6 mouse (left), a male (middle) and female (right) BALBc mouse. Example recordings in (b) - (f) show the average firing rates binned according to different sub-stages of each behaviour. (g) shows the raster plot used to form middle panel in (e). The location of each probe recording site is shown in (h).

To build on these preliminary results, the authors then made probe recordings to better understand the contribution of individual neurons to each behaviour.

e) Describe the different types of responses shown in figure 8. What do the firing patterns suggest about the role of neurons in relation to mating and aggressive behaviours? Is there a qualitative difference in terms of how neurons in the VMHvl respond in the two behaviours? If so, explain.

Figure 9 shows how the population activity changes when a male or female mouse is introduced to the home cage.

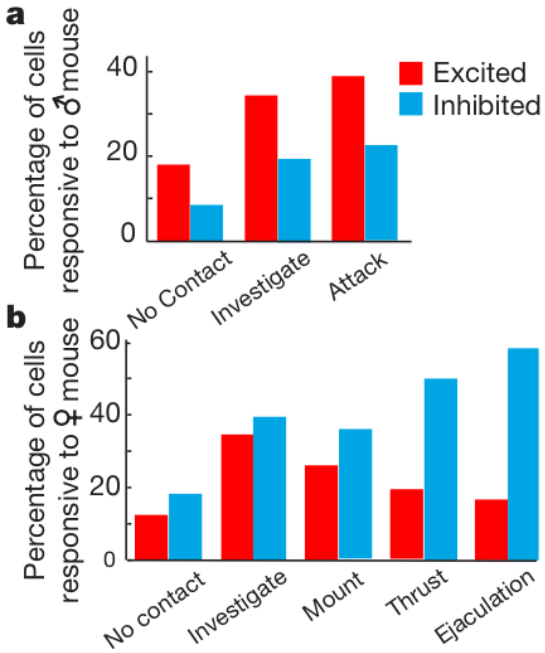


Figure 9: Summary of neuronal responses in VMHvl to the presentation of a male (top) and female (bottom) mouse.

They go on to demonstrate that optogenetic stimulation of the VMHvl will cause mice to attack inanimate objects, such as a surgical glove.

f) You decide to reproduce these experiments yourself, because you want to show your friends that you can make your pet mouse, Maureen, attack things on command. Incidentally you notice that the amount of stimulation required to induce aggressive behaviour is much greater if you do this during an ongoing mating behaviour - mice aren't that into BDSM. You have a hunch that there might be two competing circuits, one for mating and another for fighting. Devise an experiment that would test this (hint: consider figures 7 and 9)