Neural network model of amygdalar memory engram formation and function

Summary: The past decade has seen an explosion of experimental research into the formation and function of *memory engrams*, distributed sub-populations of neurons that undergo enduring changes in response to learning and that are both sufficient and necessary to retrieve a learnt association. However, theoretical understanding has lagged behind experimental work, leaving important questions unanswered, in particular, what computational principles determine how neurons collectively coordinate to encode an experience in an engram or how engrams interact with one another to control behavior. Here, we provide a novel interpretation connecting engrams in associative learning to a Bayesian behavioral account of learning. We then implement our insights in a rate-based neural network model and use it to explain key experimental findings including (1) how extinguishing one conditioned stimulus can affect responses to an unrelated conditioned stimulus, and (2) when and how unlearning or new learning occurs if reward contingencies change. Our model relies on key circuit-level findings, specifically excitability, competition, and dopamine-controlled synaptic plasticity. Our minimal model is consistent with both neuroanatomy and biologically plausible learning mechanisms.

Background: Memory engrams are commonly studied using associative learning tasks e.g. auditory fear conditioning. In general, neurons display endogenous time-varying excitability. During conditioning, neurons compete for recruitment to the resulting engram, with excitable neurons being preferentially recruited. Learning increases AMPA/NMDA ratio and spine density in engram neurons. After conditioning, activation of the engram induces recall. Engram size differs between regions, but is consistent within regions and is sparse ($\sim$10% of the population). Two experimental findings are of particular interest. First, Rashid, 2016 showed that if mice undergo separate fear conditioning to two auditory tones, the second fear response is acquired more quickly and extinguishing one fear response partially extinguishes the other fear response if the conditionings are separated by 6 hours, but not 24 hours. We call this phenomenon *entangled engrams*. Second, there are conflicting accounts about learning a changed stimulus-reward pair (e.g. during fear extinction); whether it is the result of weakening a previously formed engram (unlearning), or the creation of a new engram (new learning). See Josselyn 2020 for references.

Computational Principles: As animals move through the world, they encounter a sequence of stimuli and rewards or punishments, which we assume are generated by an underlying cause (*latent state*). Therefore, the animal needs to learn an association between a stimulus and ensuing reward based on a belief on the current latent state. Critically, the animal needs to decide if and when the latent state changes, and whether different stimuli result from the same or different latent states. Gershman 2015 showed that a model with separate parameters per latent state can explain diverse behavioral findings when combined with a Bayesian nonparametric model for latent state inference. Here, we connect this Bayesian behavioral account with experimental engram evidence by arguing that an engram (in the associative learning context) is the neural representation of a latent state, and creation of a new engram corresponds to addition of a latent state to a reservoir. This view explains the roles of excitability and competition: excitability can initiate the creation of a new latent state's neural representation, and competition promotes a separation between different latent states' representations.

Model: Our model (Fig. 1) has five nuclei: the conditioned stimulus (CS, e.g. auditory input), the unconditioned stimulus (US, e.g. shock), lateral amygdala (LA), central amygdala (CE, which controls the model's freezing response) and dopaminergic neurons (DA). CS projects to LA and LA projects to CE. DA computes the reward prediction error by summing inputs from CE and US with opposite signs. We require CS to LA synapses to be excitatory and LA to CE synapses to be some excitatory and some inhibitory. LA neurons with excitatory (inhibitory) connections to CE are called LA$^+$ (LA$^-$); DA is similarly divided into two subpopulations, with DA$^+$ (DA$^-$) firing if the prediction error (summed inputs from CE and US with opposite signs) is positive (negative). Synapses evolve according to Hebbian learning with weight decay, but are constrained such that their signs cannot change.

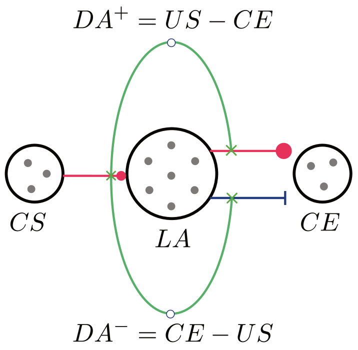


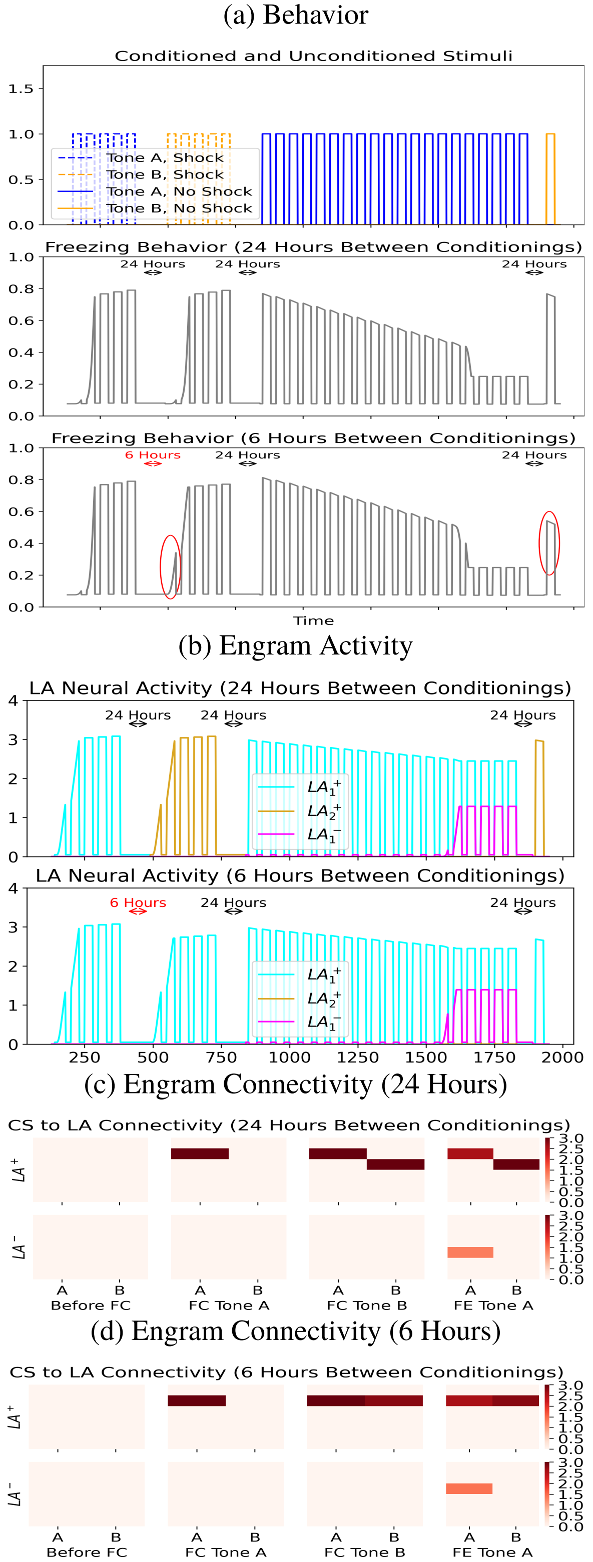
Figure 1

Excitability is implemented as a time-varying excitatory input to a random subset ($\sim$ 15%) of LA neurons, and competition is implemented as an inhibitory input to all LA neurons proportional to the summed rates of LA neurons. Dopamine gates Hebbian plasticity in the following way: DA^{+} (DA^{-}) enables Hebbian learning of CS to LA^{+} (CS to LA^{-}) and of LA^{+} to CE (LA^{-} to CE), and accelerates the decay of LA^{-} to CE (LA^{+} to CE) synapses. After learning, the engram is the population in LA that is now activated by the CS.

Selected results: We concentrate on the experiment of Rashid, 2016 (Fig. 2a, top) because it nicely illustrates our model's properties. We fear condition Tone A by pairing it with an aversive shock. We then fear condition Tone B by pairing it with an aversive shock after either 24 hours (Case 1) or after 6 hours (Case 2). 24 hours later, we extinguish the fear response to Tone A by repeating it without a shock, and after another 24 hours, we test the behavioral response to Tone B. In Case 1, when Tone A is conditioned, freezing behavior rapidly increases, and when Tone B is conditioned, freezing behavior increases at the same pace (Fig. 2a, middle). As Tone A is extinguished, behavioral freezing slowly declines, but Tone A fear extinction has no effect on fear retrieval to Tone B a day later. In contrast, in Case 2, when Tone B is conditioned, freezing occurs more quickly (Fig. 2a, bottom), and when Tone A is extinguished, Tone B fear recall is reduced.

Excitability and competition are responsible for these differences. In Case 1, fear conditioning for Tone A is learnt using one subset of LA^{+} neurons (Fig. 2b, top) (teal), and fear conditioning for Tone B is learnt using a different subset of LA^{+} neurons (gold) due to excitability changing after 24 hours. This expresses the model's belief that Tone A and Tone B originate from different underlying latent states due to the temporal separation between them. Extinguishing Tone A weakens the Tone A fear engram (teal) and creates a competing Tone A extinction engram (pink) in LA^{-}. However, because the Tone B fear engram is distinct, its firing is unaffected after Tone A extinction (gold). In contrast, in Case 2, fear conditioning for Tone A and Tone B uses the same population of LA^{+} neurons (Fig. 2b, bottom) (teal) because excitability has not yet appreciably changed. This expresses the belief that Tone A and Tone B originate from the same underlying latent state. Both CSs are consequently assigned to the same engram, and learning Tone B is faster because synapses from the shared fear engram to CE were strengthened during Tone A fear conditioning. Extinguishing Tone A thus results in a weakening of the shared fear engram (teal), partially extinguishing Tone B's fear recall. For confirmation, we visualize the strength of synapses from Tone A and Tone B to LA, distinguishing between LA neurons that excite CE (LA^{+}) and those that inhibit CE (LA^{-}). In Case 1, Tone B fear conditioning excites a separate LA^{+} engram than Tone A does (Fig. 2c), meaning Tone A fear extinction does not weaken Tone B's path from LA^{+} to CE. In Case 2, Tone B fear conditioning excites the same LA^{+} engram as Tone A (Fig. 2d), meaning Tone A fear extinction weakens Tone B's path from LA^{+} to CE.

Figure 2



(a) Behavior

(b) Engram Activity

(c) Engram Connectivity (24 Hours)

(d) Engram Connectivity (6 Hours)

Discussion: Our model critically relies on heterogeneity in dopaminergic neurons, both in response to prediction error and the asymmetric effects they have on plasticity at specific synapses. The detailed model allows for predictions such as the possibility of generating an extinction engram without fear conditioning, by activating the appropriate DA population, that would delay subsequent fear learning.