

What's Next if Reward is Enough? Insights for AGI from Animal Reinforcement Learning

Shreya Rajagopal

University of Michigan

500 S State St

Ann Arbor, MI 48109, USA

SHREYARA@UMICH.EDU

Editor: Patrick Hammer

Abstract

There has been considerable recent interest in the "The Reward is Enough" hypothesis, which is the idea that agents can develop general intelligence even with simple reward functions, provided the environment they operate in is sufficiently complex. While this is an interesting framework to approach the AGI problem, it also brings forth new questions - what kind of RL algorithm should the agent use? What should the reward function look like? How can it quickly generalize its learning to new tasks? This paper looks to animal reinforcement learning - both individual and social - to address these questions and more. It evaluates existing computational models and neural substrates of Pavlovian conditioning, reward-based action selection, intrinsic motivation, attention-based task representations, social learning and meta-learning in animals and discusses how insights from these findings can influence the development of animal-level AGI within an RL framework.

Keywords: Animal Learning; Artificial General Intelligence; Reinforcement Learning; Social Learning; Meta-Learning

1. Introduction

On the face of it, the field of AI seems to be making progress in leaps and bounds. Grand declarations of AI outperforming humans in a host of tasks have propagated the notion that we are not too far from what had purely seemed to be science fiction thus far - the day AI will finally take over. The trouble with this notion is that while AI does seem to be doing exceptionally well at the specific tasks it is designed for (hereafter referred to as narrow AI), we are far from achieving general artificial intelligence wherein the same system can perform well in a range of different tasks and learn to adapt to new tasks in unfamiliar environments (Fjelland, 2020).

In his 1988 article aptly titled "The prospects for Building Truly Intelligent Machines", David Waltz, a former professor of computer science at Columbia University, made an almost prophetic observation: "It seems certain that we will be able to build hardware that is a match for human computational power for an affordable price within the next thirty years or so... Having hardware with brain-level power will not in itself, however, lead to human-level intelligent systems, since the architecture and programs for such systems also present unprecedented obstacles" (Waltz, 1988). This is precisely where the field of AI finds itself now. On the one hand, we've made use of increasingly cheaper computational power and large amounts of training data to achieve great success in narrow AI. The last few years have seen incredible progress in the field of generative AI - AI that generates novel content as opposed to analyzing existing content. We have seen the evolution

of transformer architecture from the basic transformer model (Vaswani et al., 2023) to Generative Pre-trained Models (GPT), all the way to ChatGPT and now GPT-4 (OpenAI, 2022, 2023). With its ability to quickly understand and generate human-like text, ChatGPT has enabled industries like customer service, healthcare, and education to automate and improve efficiency in communication, data analysis, and decision-making processes. GPT-4 has the added advantage of taking images and text as input to give a more informed reply. GPT-4 has demonstrated exceptional performance on several academic benchmarks, including passing a simulated bar exam in the top 10% of all candidates (OpenAI, 2023). Dall-E – a text-to-image generator based on the GPT-3 model – can generate art in specific styles and with specific subject matter (OpenAI, 2021). Flamingo – a family of visual language models – can take in multimodal language and image inputs and perform several tasks that integrate these modalities including visual question-answering tasks where it answers a question by interpreting an image and captioning tasks where it provides a description of an input image (Alayrac et al., 2022).

Convolutional neural networks have long been outperforming humans in object recognition, and are now extending this success to diagnostic and research domains (Buetti-Dinh et al., 2019; Ho-Phuoc, 2019; Tschandl et al., 2019). One of the most prominent successes of AI that furthered the idea of being close to achieving AGI was that of AlphaGo. DeepMind’s AlphaGo, built using a combination of deep neural networks and reinforcement learning, defeated the European Champion 5 - 0 in the game of Go, considered to be particularly challenging for AI given the massive search space of possible moves (Silver et al., 2016).

On the other hand, though, most of these wins for narrow AI haven’t stemmed from ‘truly intelligent systems that can think for themselves and actually understand the task they are performing. While there is yet to be a clear consensus on what abilities a system needs to possess AGI, it can broadly be thought of as one that can perform well on multiple tasks and can adapt to environments it was not explicitly trained on. For example, while AGI need not necessarily match up to, or even have the same attributes as human intelligence, it remains a good example of what it means for intelligence to be general. There are several different approaches to building AGI systems at present. One recent approach puts forth the idea that formulating the goals an agent is supposed to achieve as a reinforcement learning problem will naturally bring about complex intelligence in the agent in pursuit of these goals. In a paper titled “Reward is Enough”, some of the most influential figures in the field of reinforcement learning today argue that the process of reward maximization is enough to fully reproduce intelligence artificially (Silver et al., 2021). Using examples of both natural intelligence as represented by a squirrel trying to store acorns for the winter, and artificial intelligence in terms of a kitchen robot, they demonstrate how all attributes of intelligence can be thought of as subserving the pursuit of reward. The goals of a squirrel can be thought of as rather straightforward - survive, reproduce and minimize hunger. However, to achieve these goals, the squirrel has to learn to navigate a complex environment using several attributes of intelligence, often in tandem - perception, prior and learned knowledge, motor skills, and even social knowledge. The authors suggest that even simple goals in a reinforcement learning (RL) framework within a challenging environment allow general intelligence to emerge in an agent. They argue it is primarily the complexity of the environment that brings about intelligence – even if the goal is just to find food, the squirrel will need to develop strategies to look for hidden food with limited energy resources, compete with and deceive other squirrels, and so forth.

Briefly, an RL framework involves an agent trying to achieve a goal in the environment in which it exists. The environment poses the problem, while the agent solves it, subject to being bound by

certain constraints. The way the agent solves the problem is by trying to maximize the cumulative reward it receives from the environment in pursuit of its goal. The environment doles out rewards to the agent based on a reward function, and the instantaneous reward the agent receives at each time step is termed the reward signal. This reward might manifest in actually receiving a physical reward from the environment, or even from a sense of internal reward (intrinsic motivation) (Singh, Lewis, and Barto, 2009).

The idea of reward being enough to explain all of intelligence does sound rather simplistic on the face of it and has also quickly garnered considerable criticism in the short time since it was published. The indignation of those in opposition is best summarized by the use of the same title across articles - “Reward is not enough” (Byrnes, 2021; Roitblat, 2021; Vamplew et al., 2021). Specifically, Byrnes (2021) suggests that simple reward functions cannot always capture complex goals and that such goals might best be captured by multiple reward functions with occasionally conflicting objectives. Roitblat (2021) argues that an artificial agent’s dependence on humans to select the optimal reward function for a given problem implies that the agent is not truly achieving general intelligence since identifying an appropriate goal itself is a key aspect of such intelligence. Vamplew et al. (2021) echo Byrnes’ suggestion for the need for multi-objective reward representations. They emphasize that a single scalar reward cannot capture how different aspects of a complex task interact with each other to achieve a goal, and suggest the need for a vector reward representation instead. Especially in tasks that involve multiple conflicting objectives, a scalar reward will only maximize the reward associated with one of these objectives or a scalarized combination of them with a specific, predetermined weighting.

Additionally, if an agent were to develop general intelligence within an RL framework, there are several factors researchers must take into consideration. For one, the ‘reward is enough’ hypothesis requires the agent to be subjected to one or more sufficiently complex environments in order to develop several attributes of intelligence. What would these environments look like for AGI systems - what challenges must they present the agent with, what objects must they contain, should they also include other agents, and so on. Secondly, the researcher has to impose certain constraints on the agent. In a biological agent, these constraints might come in the form of the number of neurons and interconnections that exist. Given different embodiments of cognition across different systems of natural and intelligent systems, careful thought would have to go into what limitations and abilities an agent has - would they have to match human/non-human animal abilities? Would these constraints be different? Finally, there would also be a need to identify a sufficiently expressive reward function that both ensures the reward signal is not too sparse, in order to encourage learning long before being close to achieving the final goal and also somehow accommodates multiple sub-goals. There is also great latitude in defining reward functions in reinforcement learning problems - while they can incorporate the agent’s interactions with the external environment, they can also take into account the agent’s internal environment, with the latter being particularly useful to aid learning when the former is sparse (Barto; Pathak et al., 2017). Additionally, AGI researchers would also have to select the reinforcement learning algorithm that learns optimally from reward.

While the reward hypothesis provides an interesting and meaningful framework for working towards artificial general intelligence, we will need to look elsewhere to provide meaningful intuitions for addressing the questions it has brought forth. Animals can be thought of as possessing general intelligence with their abilities to perform several different tasks and adapt to changes in their environment. Given that evolution has honed the way animals interact with their environments to meet multiple concurrent goals like survival and reproduction, understanding how animals learn

from reward is a meaningful initial step in formulating the AGI problem in a reinforcement learning framework. In fact, animal learning and decision-making has been densely explored within a reinforcement learning framework and neural correlates of RL have been found in animal (and human) brains (Maia, 2009; Niv, 2009). These findings and their implications for AGI will be discussed in detail in this paper.

This paper begins by exploring how animal learning fits into an RL framework. It first evaluates RL-based computational models of classical conditioning in animals and highlights what these models might add to AGI development. It then explores reward-based action selection in animals and describes competing RL-based theories for how this is achieved, also commenting on how the neuroscientific findings underlying these theories would be useful for AGI. It then describes the concept of intrinsic motivation in animals, wherein animals obtain rewards from within themselves and not just from the environment. The paper then delves into the role of attention in making sense of high-dimensional data, and how animals form optimal task representations. It then describes social learning in animals within a reward framework, and how this could be incorporated into AGI systems. Finally, it discusses the role of meta-RL in animal learning, and how animals quickly adapt to new tasks based on what they have already learned. Across all sections, the paper provides ideas to translate these animal findings into building AGI systems. The paper concludes with a summary of meaningful insights discussed throughout the paper.

2. Classical Conditioning within an RL Framework

This section reviews computational two RL-based models of classical conditioning in animals, and comments on the strengths and shortcomings of these models. These two models are based on two competing theories of how animals make predictions about their environment – Model-Based and Model-Free approaches. It also provides some directions for how AGI research can benefit from the findings derived from these models.

2.1 Computational Models of Classical Conditioning in Animals

Cognitive psychology broadly divides the way animals learn from reward into two categories - Pavlovian learning and instrumental conditioning. Pavlovian learning, also termed classical conditioning, involves the animal learning to associate reward with a previously unrewarded stimulus by its repeated association with a naturally rewarded stimulus. Instrumental conditioning, on the other hand, involves the animal actively selecting its actions to maximize the possibility of reward and minimize punishments. Thus the former can be thought of as learning to predict rewards and the latter as learning to select the most rewarding actions.

To understand Pavlovian learning, it is first important to realize that animals biologically associate reward with some stimuli like food and social interactions. In some sense, this can be thought of as prior knowledge attained over the course of evolution to maximize an organism's fitness. In classical conditioning, these naturally rewarded stimuli are called unconditioned stimuli (US). Other stimuli, like a musical tone or a flash of light, are neutral stimuli unassociated with reward or punishment. If these neutral stimuli are repeatedly presented to an animal alongside a naturally rewarded stimulus, the animal begins to associate these previously neutral stimuli with reward. This no longer neutral stimulus is termed a conditioned stimulus (CS). The way to determine whether a neutral stimulus has turned into a CS is by measuring the animal's response to the neutral stimulus presented in isolation - if the animal generates the same response as it would for the US the

neutral stimulus was paired with, the animal is said to have acquired a conditioned response (CR) to a previously neutral stimulus.

There are a host of experimental phenomena associated with Pavlovian learning that a single computational model has yet to explain. An example is the blocking phenomenon (Jones, Gray, and Hemsley, 1990). Blocking experiments involve two training stages. In the first stage, a neutral stimulus A is repeatedly presented alongside the US till it acquires a CR. In the second stage, this CS A, and another neutral stimulus B are simultaneously presented with the US. In this scenario, stimulus B never acquired a conditioned response. Take for example an experiment where a flash of light repeatedly preceded a drop of juice given to a monkey. Over time the monkey will begin anticipating juice with flashes of light. However, if the flash of light is then paired with a musical tone, and followed by juice delivery repeatedly, the monkey never grows to associate the musical tone with juice as acquiring a conditioned response to the tone is blocked by the flash of light. This is just one of the several phenomena this kind of animal learning encompasses, and some are easier explained intuitively than others (Alonso and Schmajuk, 2012).

The temporal difference model of classical conditioning uses a model-free RL framework to look at Pavlovian learning phenomena (Ludvig, Sutton, and Kehoe, 2012). A *model-free framework* suggests that animals make predictions about their environment simply based on cached values of reward associated with each state they find themselves in, which they have learned by repeated associations over several trials. This is in opposition to a *model-based framework* where the animal truly understands the environment and thinks about what prediction would be the most meaningful given a particular state. A temporal difference model allows an animal to update its expectations of reward from the stimuli it sees at each time point, based on how the combination of its present reward and future reward expectations at present differ from its previous expectations of reward. The model is described by Equations 1 through 3.

$$V_t(X) = W_t^T X = W_t(i)x(i) \quad [1]$$

Equation 1 describes how the value of the state X the animal is in, at time t, is computed as the weighted sum of all CS-s present at time t, with the weights depicting the strength of the association the animal has formed between each stimulus and the US, and the value depicting the animal's expectation of the cumulative reward it will receive through the course of the trial, at that particular point in time. A state simply comprises all CS-s an animal is exposed to at a particular time point, as depicted by $x(i)$.

$$\delta_t = r_t + \gamma V_t(x_t) - V_t(x_{t-1}) \quad [2]$$

Equation 2 shows how the temporal difference at time t - δ_t - is computed as the difference between the animal's expectations of cumulative reward in the previous state - $V_t(x_{t-1})$, and the animal's new expectations of reward at the current timestep. The animal's new expectations are computed as a combination of the instantaneous reward the animal might have received at the current timestep - r_t and it's new expectations of cumulative reward in the new state it has attained - $V_t(x_t)$ - as discounted by a factor of γ .

$$W_{t+1} = W_t + \alpha \delta_t x_t \quad [3]$$

The animal's associative weights of the stimuli present at time t, with the expectation of reward, are then updated based on this temporal difference. If the animal's expectations of the world at the

previous timestep were greater than the new estimate it formed at the current timestep, based on the instantaneous reward received, the temporal difference would be negative, and the animal would reduce associative weights between the stimuli present in the trial at the time, and reward. On the contrary, if the world turned out to be better than expected, it would increase associative weights.

As this model has a temporal element within each trial of stimulus presentation, it was presented with a need to also define what stimulus representations looked like in the animal's brain for the duration of stimulus presentation. Did the stimulus representation sustain at the same intensity for the whole duration of its presence, or did it decay with time? To address this, they modeled classical conditioning phenomena using 3 kinds of temporal stimulus representations - the presence, complete serial compound (CSC), and the microstimuli representations as seen in Figure 1. The presence representation simply represented the stimulus with the same strength for the entire duration it was present. The CSC representation, on the other hand, consisted of the same number of stimulus elements as the number of time points the stimulus was present for, with only one element active at a time. Finally, the microstimuli representation consisted of a series of temporally mean-shifted, overlapping Gaussians of increasing variance for the duration the stimulus was present.

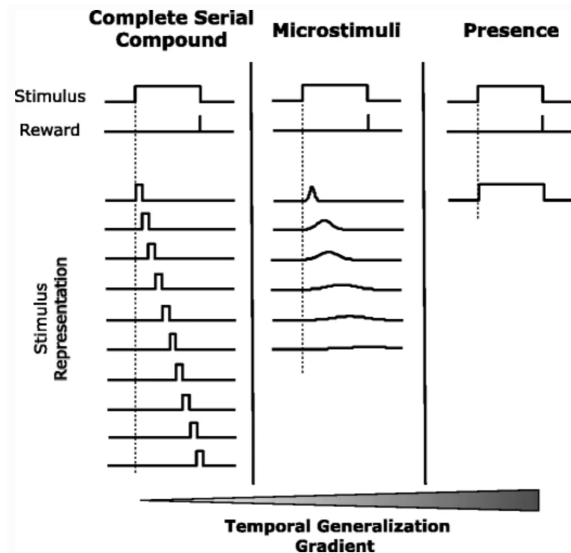


Figure 1: Types of Stimulus Representations–Reproduced from Ludvig et al. (2012)

Across all phenomena this paper modeled, the microstimuli representation was shown to most closely replicate results from empirical data. The authors suggested that the overlap of the elements of the representation – the individual Gaussians – in this representation allowed for learning to propagate faster. Interestingly, there appears to be some evidence for the microstimuli representation being employed in monkey brains. Electrophysiological recordings from the putamen of monkeys exposed to aversive, neutral, and rewarded cues yielded the same pattern across all 3 conditions – distinct clusters of neurons representing the distinct mean-shifted Gaussian elements of the microstimuli representation, with increasing variance (Adler et al., 2012; Gershman, Moustafa, and Ludvig, 2014). Additionally, there is mounting neural evidence in opposition to the CSC representation. With dopamine responses having been found to be a proxy for temporal difference computations in the primate brain, the way the dopamine responses change in different states when

paired with a reward has not been seen to be consistent with the CSC representation (Daw, Courville, and Touretzky, 2006; Ludvig, Sutton, and Kehoe, 2008).

The significant contribution of the TD model over previous models of classical conditioning like the Rescorla-Wagner (RW) model Rescorla and Wagner (1972) was its incorporation of the idea of time within a trial. By doing so, it was able to explain empirical findings from modulating the inter-stimulus intervals across several key conditioning phenomena.

However, it is entirely possible that unlike what the TD model assumes, animals do form models of their world and understand their environment beyond just forming stimulus-reward associations. Reinforcement learning models of this kind are called model-based approaches. This form of learning, wherein the animal attributes the different states associated with a task to an underlying task structure and tries to acquire the structure to guide their predictions of reward is called structure learning (Gershman, Norman, and Niv, 2015). The latent cause theory of classical conditioning is a prominent model-based approach to animals learning to predict reward (Gershman and Niv, 2012). This theory suggests that animals attribute experimental trials to some underlying latent cause. More specifically, if the same set of CS-s and US appear together repeatedly, they might attribute that state to being brought about by a particular cause. The way they are thought to attribute a cause to data they experience is described by Equation 4. This equation suggests that the probability of attributing a certain cause to the state (set of stimuli) they are subjected to is directly proportional to the product of the probability that that particular state could be generated by that particular cause, and how frequent that particular cause is as per their model of the world.

$$P(Cause|Data) \propto P(Data|Cause).P(Cause) \text{ [4]}$$

This model has two key assumptions. Firstly, the animal doesn't know beforehand how many latent causes exist to give rise to all the states in its environment. Secondly, the animal is thought to attribute each trial as originating from a single latent cause. The probability of an animal attributing a trial to a cause K is described by Equations 5 and 6. In the equations, a represents the animal's prior belief about the number of latent causes that exist with $a=0$ signifying a single latent cause, and $a = \text{infinity}$ signifying that every single trial is generated by a distinct latent cause. N_k is the number of previous trials already attributed to the cause k , and t is the number of the current trial. Equation 5 describes the probability of the current trial being assigned to cause K, if K has already been attributed to at least one previous trial, while equation 6 describes the probability of the trial being attributed to K if it is a new cause.

$$K = \text{Old Cause} : P(C_t = K|C_{1:t-1}) = N_k/(t - 1 - a) \text{ [5]}$$

$$K = \text{New Cause} : P(C_t = K|C_{1:t-1}) = a/(t - 1 - a) \text{ [6]}$$

Assuming that the animal had its own model of the world, and describing it as in the equations above allows the latent cause theory of classical conditioning to explain phenomena from several classes of conditioning effects that the RW and Temporal Difference model struggle with including Discrimination, Recovery, and some Pre-exposure effects as well (Gershman and Niv, 2012). However, since the latent cause theory fails to incorporate the concept of time within a trial, and consequently fails to comment on stimulus representations over time, and the impact of varying the inter-stimulus interval between CS-s and the US.

2.2 Takeaways for AGI from models of Classical Conditioning

While neither model discussed successfully explains all empirical phenomena associated with classical conditioning in animals, the development of these models, and understanding their successes and shortcomings provide some interesting insights to AGI researchers. First and foremost, there does seem to be clear evidence of non-human animals using an RL framework to learn and make predictions about their environment. This further highlights that AGI within an RL framework could benefit from a study of animal learning. More importantly, the use of a model-based approach to understanding classical conditioning has been able to explain a set of Pavlovian learning phenomena that older model-free approaches had struggled to explain. Although the latent cause theory had its own shortcomings, a large proportion of these could be attributed to the fact that the theory did not accommodate the concept of time within a trial. Taken together, these findings suggest that not only are animals learning to make predictions based on reward but that they might even be using these stimulus-reward associations to build a model of the world they are in. With non-human animals being some of the simplest general intelligence systems AGI can aspire to, this suggests that some combination of model-based and model-free RL might be useful in developing general intelligence.

Another useful takeaway, particularly from the Temporal Difference model would be to think of how AGI systems should represent the stimuli they are subjected to, in their own internal systems. This is referred to as the feature representation learning problem in machine learning. Using different types of internal representations of stimuli across time dramatically altered the model's ability to replicate empirical phenomena, with the microstimuli representation coming the closest in this replication, and the CSC representation being better than the presence representation. These findings make two major suggestions: 1) Animals likely distribute their representation of the stimuli they observe over several different elements (as seen in CSC and microstimuli representations) 2) The representative elements they use likely overlap temporally to some degree (as seen in the microstimuli representation). The authors also suggested that it is this overlap that allows for faster learning. This is because, at any given time point, only a single element of the CSC will be active and consequently be subjected to weight changes, while several elements of the microstimuli representation would be simultaneously active at a given time point. From these findings, it appears that the process of evolution might have arrived at a distributed, temporally overlapping feature representation of stimuli.

Finally, the fact that Pavlovian learning is even possible in the animal world is reliant on the fact that animals naturally associate reward with some stimuli like food and social interactions. In other words, animals seem to rely on prior beliefs about the world to make sense of new experiences and formulate new beliefs. Even the model-based approach to classical conditioning relied on several assumptions about how animals understand the world, both in terms of the number of latent causes they think exist in their environment and in attributing only one cause to each trial they experienced. These ideas bring forth questions about prior beliefs being factored into AGI. If AGI were to take inspiration from animal learning, what stimuli would be naturally rewarded for an AGI, and what assumptions would we make about how it viewed the world? These remain open questions for the field of AGI.

While these models both contribute interesting insights and pose meaningful questions for AGI research, they also present a rather pessimistic future for achieving AGI purely within the reinforcement learning framework. Even with several decades of animal experiments providing

insight into how animals learned to predict reward and many groups of computational neuroscientists and cognitive psychologists trying to model classical conditioning in animals, we are yet to arrive at a comprehensive model that explains at least the large majority of associated empirical phenomena. If it isn't even straightforward to decipher existing general intelligence systems within the reinforcement learning framework, to build AGI from scratch simply by learning from reward seems rather far off, even more so if we want the system to be interpretable to humans.

3. Reward-Based Action Selection in Animals

This section discusses instrumental conditioning within an RL framework, wherein reward and punishment are used to modify an animal's behavior in terms of the actions it chooses. Unlike in classical conditioning where the animal can only predict the reward but not influence it, in instrumental conditioning, selecting the most optimal action impacts the reward they receive. This section describes two kinds of RL methods that animals might employ – Model-based and Model-Free RL – and the factors that influence which of these methods are selected. It also provides insights for AGI researchers to navigate the use of these RL methods.

3.1 Model-Free vs. Model-Based Reinforcement Learning in Animals

Animals select actions that maximize reward and minimize punishment in one of 2 ways – Habitual Learning and Goal-Based Learning (Balleine and O'Doherty, 2010; Doll, Simon, and Daw, 2012). The use of a post-trainer reinforcer devaluation in animal learning experiments will help make this distinction clear - Consider a typical experiment where a rat has to perform a sequence of actions to get to the reward (Daw, Niv, and Dayan, 2005). It first has to press a lever, and then enter a food magazine where it will be presented with a food pellet. After being trained on several trials of this task, the rat learns to perform these sequential actions to obtain reward. Now, following training, the value of the outcome is somehow reduced. This can be done by either poisoning some of the pellets causing the rat to get sick, or pre-feeding the animal. If the animal continues to perform the series of sequential actions just as it previously had, it indicates an insensitivity to the devaluation which suggests the rat is acting purely from habit. Such learning is called *habitual learning*.

Now, consider the case where the rat realizes that it no longer wants the food pellets and decides actively to stop pressing the lever and entering the magazine or at least change the frequency of these actions. In this case, the rat seems to have a model of the world it is in, and how changing its actions are going to influence the outcome it receives. It is also able to have a concept of a goal, wherein its goal changes when it is no longer hungry or worried about the pellets being poisoned. This type of learning, which is sensitive to outcome devaluation, is called *goal-directed learning*.

While these distinctions have long existed in the realm of cognitive psychology, they garnered recent interest in terms of using a reinforcement learning framework to model these action-selection methods. In this context, habitual learning can be thought of as *model-free reinforcement learning* because it estimates the value of each possible state-action pair over the course of several training trials, where the value refers to reward predictions of long-run outcomes (Daw, Niv, and Dayan, 2005). The animal retains these values by use of a cached storage and any changes to the reward or environment would only reflect in the cached values after further extensive training. This method of learning is highly computationally efficient but leaves the animal rather inflexible to immediate changes in the environment or its goals.

On the other hand, goal-directed learning is best explained by a *model-based reinforcement learning* framework, where an animal builds a model of its world to truly understand how its actions impact the outcomes it receives, rather than relying on cached values from associations formed over several trials (Doll, Simon, and Daw, 2012).

But which of these do animals use to select actions? There is considerable evidence for dopamine neurons encoding reward prediction error, specifically as defined by the temporal difference model (Montague et al., 1995; Schultz, Apicella, and Ljungberg, 1993). A TD reward prediction error changes when the animal changes its expectation of reward. The findings from an experiment by Schultz, Apicella, and Ljungberg (1993), which obtained recordings from midbrain dopamine neurons of monkeys help highlight this. The task was a simple choice task, where the monkey had to press one of two levers following a visual cue. The left lever led to the delivery of juice, while the right lever did not elicit any reward.

In the initial phases of the task, the monkey behaved erratically, but the dopamine neurons responded strongly when the juice reward was received. However, over time, as the monkey learned the association and focused its attention on the left lever, the dopamine neurons were silent after juice delivery. They did begin to strongly respond to the onset of the visual cue instead. This experiment shows how the responsibility for reward is propagated backward in time with training, which is exactly what the class of TD models does Glimcher (2011). Once the animal grows to expect a reward with the left lever press, receiving the reward following that action is no longer contrary to its expectations, as opposed to the beginning where it had no idea the left lever press would yield a reward.

The strong evidence for dopamine neurons coding for TD reward prediction error in animals suggests the animal brain is making use of a model-free approach. However, animals are also sensitive to outcome devaluation, subject to the amount of prior training. In the lever press followed by the magazine entry experiment, rats remained sensitive to outcome devaluation in the action more proximal to the reward - magazine entry - even with extensive prior training. This meant that they decreased their frequency of magazine entry on the fly when the outcome was devalued, even when they had been overtrained on entering the magazine (Holland, 2004). The even more interesting finding was what happened to the action more distal to the reward - lever pressing. While rats remained sensitive to outcome devaluation in their lever-pressing responses following moderate prior training, they lost this sensitivity following extensive training.

These findings seem to suggest not only that the brain has multiple controllers for action selection, but that it has some means to decide when to use which. Daw, Niv, and Dayan (2005) suggested that the way the brain makes this decision is based on the uncertainty each controller has about its value estimates. They model habitual learning as a cache system, and goal-directed learning as a tree, wherein the animal needs to build a tree with the states it experiences as the nodes, and the possible actions being branches out of the node leading to new states. They suggest that the uncertainty in the model-based approaches comes from having to conduct a tree search to determine which tree best fits the world they are subjected to. The uncertainty in the model-based approach comes from the idea of bootstrapping: Using value estimates of future states as reliable stand-ins for predecessor states. The final decision about which controller to use is thought to be decided on the basis of which one is more likely to be more accurate by being less uncertain.

Daw, Niv, and Dayan (2005) argue that neural substrates of the model-free approach - i.e. the dorsal striatum and dopaminergic afferents, and the model-based approach - i.e the prefrontal cortex and associated regions compete to influence the animal's action selection. A review of the neural

substrates of habitual and goal-directed learning in both rodents and humans seem to suggest that there is also evidence for these two controllers cooperating and functioning in synergy (Balleine and O’Doherty, 2010). However, they point out that it remains unclear what conditions cause these controllers to choose to cooperate or interfere with each other.

3.2 Cooperation between Model-Free and Model-Based Controllers – Insights for AGI

There is evidence to suggest that there is a significant overlap between the cortico-striatal networks involved in habitual and goal-directed learning in both rodents and human beings Balleine and O’Doherty (2010), and no study to date has been able to find completely distinct functional networks for the two kinds of action-selection mechanisms in animals. This seems to suggest that the same underlying structure and set of functional networks are somehow able to support both model-based and model-free approaches. This is an interesting insight for building AGI systems as it appears that the ability to acquire distinct methods for action selection is more of a functional trait than a structural trait, as the same underlying architecture seems able to support both modes of learning.

However, this overlap posed a new problem for animal learning research. Dopamine neurons have consistently been shown to encode for temporal difference-based reward prediction error, which has long been associated with a model-free approach. This is because computing the temporal difference RPE between the animal’s prior estimate of long-run reward and the new estimate is used over the course of several training trials to create a tabular storage or cache of the estimated value of taking each possible action in each state. A model-free approach consequently has no need to store state-transition information or the reward function - all it stores is a precomputed set of value estimates. If structures for goal-directed (model-based) learning do overlap with that of habitual learning, model-based learning would also have to somehow make use of the same TD RPE that the dopamine neurons encode to build a model of the world. A model-based approach involves storing only the state transition function, which indicates what actions lead to what consequent states, i.e how to transition from one state to the next, and the reward function. At each given state, it iteratively computes the best course of action.

Two interesting findings about the dorsolateral cortex and its neighboring dorsomedial cortex spurred some research in this domain. Firstly, it was seen that rats with lesions in their dorsolateral cortex remained sensitive to outcome devaluation, but were no longer able to form habitual associations (Yin, Knowlton, and Balleine, 2004). This suggests that they are able to perform model-free learning but not model-based learning. On the other hand, lesions of the dorsomedial cortex resulted in rats losing their sensitivity to outcome devaluation but retained their ability to form habitual associations (Yin et al., 2005). These are regions of the same structure and the way dopamine neurons encode RPE does not differ across the two. This conundrum led to the idea of successor representation being employed for goal-directed learning.

Successor representation can be thought to fall somewhere in-between purely model-free and model-based approaches (Momennejad et al., 2017). In this model, instead of storing the state-transition matrix as in a model-based approach, the animal is thought to store long-term predictions about the states it is likely to visit in the future. More specifically, based on its starting state, it stores the probable frequency of having to visit each successor state in the future. These predictions have been shown to be possible using the temporal difference weight updation method. Combined with the reward function, this set of predictions about successor states allows for the animal to adapt to reward devaluation more cost-effectively than a purely model-based approach where it has to

iteratively compute the value of each action on the go (Russek et al., 2017). This idea neatly fits with the findings about the dorsolateral and dorsomedial striatum, with the former being thought to cache tabular estimates of value, and the latter being thought to store successor representations. If one of the controllers is impaired, the animal has to rely on the stored set of values from the other. Interestingly, the successor representation has also been demonstrated in human reinforcement learning (Momennejad et al., 2017).

Another theory about how model-free and model-based approaches interact was inspired by Sutton’s DYNA architecture (Sutton, 1991). This model suggests a means for model-free and model-based reinforcement learning controllers to operate cooperatively. Within this architecture, the model-free controller learns from the environment while it can, and then learns from the simulations generated by the model of the world constructed by the model-based controller. Within this architecture, the model-free controller doesn’t distinguish between real and simulated experiences and learns similarly from both. Experiments of human reinforcement learning have also suggested such cooperation between the two controllers (Gershman, Markman, and Otto, 2014). Rats have also been shown to reactivate task memories both in sleep and in awake states right after tasks Foster and Wilson (2006).

It is apparent that much like Pavlovian learning, understanding how animals learn to select the best actions to maximize reward is very much a work in progress. However, the main idea that springs forth is one that encourages the use of both model-free and model-based reinforcement learning within the same agent, ideally with them working in cooperation. Model-free approaches are extremely computationally cost-effective and result in fast action selection. This is particularly useful in an environment where the agent’s goals and the value of the reward remain fairly static. On the other hand, a model-based approach allows for the agent to quickly modify its behavior to suit the new circumstances as opposed to having to be trained extensively in the new conditions to acquire an appropriate response. While this approach is computationally more involved, it allows the animal to gauge what is going on in its world and learn to adapt. Even in the context of AGI, there might be circumstances in which an immediate, quick albeit relatively inflexible choice is preferred over a slower, more carefully thought-out response. For the same agent, there might be alternative circumstances wherein a more carefully thought-out a solution tailored to the current state is required. In some sense, this can be thought of in the context of Kahneman’s two systems - system 1 providing rapid, intuitive responses with lower degrees of accuracy and energy costs, and system 2 providing more deliberate and consequently more accurate responses at higher energy costs to the system (Kahneman, 2011).

While the question about how these two systems cooperate in animals remains largely open, steps in this direction with regard to successor representations and the DYNA architecture might provide some initial directions for building agents that can aspire to achieve at least animal-level general intelligence. The idea of animals learning from simulations in addition to learning from real experiences - Experience Replay - has already made it to models of computational reinforcement learning (Adam, Busoniu, and Babuska, 2012; de Bruin et al., 2015; Zhang and Sutton, 2018). Experience replay might help explain how animals are able to learn about the world from relatively small amounts of data, and this has indeed also been observed in computational RL agents that are allowed to learn from simulations of past experiences.

4. Intrinsic Motivation

Intrinsic motivation is an idea wherein the reward for performing certain actions comes from within the organism, rather than from the external environment. In human beings, it can be thought of as a lifelong growth function that is based on the basic psychological needs of *autonomy* (one’s behavior being authentic to the self as opposed to being externally controlled) and *competency* (mastery of a certain skill) (Deci and Ryan, 1985). It generally presents in the form of exploratory behavior or play, where animals interact with novel objects and explore new environments in the absence of any observable external incentive (Pellis and Burghardt, 2017). For example, rhesus monkeys have been shown to engage in mathematical puzzles for long hours in the absence of any external reward (Harlow, 1950). In humans, levels of intrinsic motivation have been shown to predict enhanced learning and creativity (Di Domenico and Ryan, 2017). Since it has been posited that intrinsic motivation in animals contributes significantly to their ability to learn quickly from small samples, it is likely that understanding the mechanisms underlying it holds meaningful insights for AGI.

To begin with, it is important to differentiate between intrinsic and extrinsic rewards since all reward signals are generated within the animal. Intrinsic rewards are rewards that aren’t tangible, like the pleasure of completing a crossword puzzle or learning to play a song on the guitar. Extrinsic rewards are tangible rewards from the external environment like food or money. Interestingly, intrinsic rewards aren’t independent of extrinsic rewards. For example, a study found that people always picked easier puzzles over hard ones if they were equally rewarded, but always picked harder puzzles when neither was rewarded Danner and Lonky (1981); Shapira (1976). Additionally, outcome-focused rewards often undermine intrinsic motivation. Children of controlling parents were seen to develop more outcome-focused over more learning-focused goals (Gurland and Grolnick, 2003). On the other hand, behavior-focused extrinsic rewards that encourage certain behaviors are likely to aid intrinsic motivation. For example, rewarding the hours spent engaging with study material over the grade received is more likely to encourage already intrinsically motivated learning (Niemic and Ryan, 2009). Taken together, these findings suggest the need to clearly define intrinsic and extrinsic rewards in computational reinforcement learning, and determine how they interact with each other based on the context.

Intrinsic motivation in primates and humans has been associated with the dopaminergic system. Specifically, there are two functional types of dopamine neurons - *value-encoding neurons* and *salience-encoding neurons*. Value-encoding neurons are excited by positive and inhibited by aversive rewards. On the other hand, salience-encoding neurons are excited by unexpected rewards or events of either valence. Value-encoding dopamine neurons were shown to be activated by motivating verbal feedback, that aids intrinsic motivation, even in the absence of extrinsic rewards (Tricomi and DePasque, 2016). Additionally, novel tasks individuals deemed ‘interesting’ resulted in the activation of the salience-encoding neurons (DeYoung, 2013). It appears that the dopaminergic system encodes intrinsic motivation in both how enjoyable an activity feels, and how new and unexpected it is. This suggests some ideas for the parameters an intrinsic reward function should encode for in computational RL.

Computational RL has begun looking for ways to incorporate the concept of intrinsic motivation. The major adjustment in the typical RL architecture has been to split the agent’s environment into an external and internal environment, with the agent and the internal environment together constituting the animal (Barto, 2013). The internal environment houses components and processes that take place within the animal, while the external environment contains everything that takes place outside

the animal. However, it must be noted that while the rewards themselves might be present in the external or internal environment, all *reward signals* are generated within the animal. This is akin to how rewards for biological animals are events or objects, and reward signals are how dopamine neurons encode this information. Such an architectural framework allows for interactions between external rewards and the animal’s intrinsic motivation, with external rewards generating internal reward signals based on the saliency of the reward as determined by the internal environment of the animal.

However, there has also been an argument against trying to differentiate between intrinsic and extrinsic motivation (Singh et al., 2010). The authors argue that both kinds of motivation are emergent properties of the primary reward function, and the difference between them is simply a matter of the degree to which something helps an animal attain its evolutionary objective. They claim that there are no clear features that distinguish between the kinds of motivation, and they emerge from an optimal reward function selected following an extensive search over reward functions. Other Attempts in modifying the reward function to accommodate intrinsic reward have also broadly resulted in equipping the RL agent with life skills that make the acquisition of new tasks considerably faster than if it had started from scratch (Houthoofd et al., 2017; Mohamed and Rezende, 2015). For example, one set of manipulations to the reward function is as follows: 1. Rewarding the exploration of novel states, 2. Encouraging the agent to find ways to reduce its error in predicting the consequences of its own actions (Pathak et al., 2017).

Additionally, there has been work using intrinsic memory to allow RL agents to develop skills, and store these skills to be used across different kinds of tasks Chentanez, Barto, and Singh (2004). In this work, the internal environment was built to represent the agent’s motivational system and was not redefined for different problems, unlike the regular RL practice. Intrinsic rewards were generated when unexpected salient events occurred. However, since expectancy is one of the few parameters that evoke intrinsic reward, this internal environment holds the potential to become more comprehensive.

Finally, intrinsically motivated goal exploratory processes (IMGEP) are another new algorithmic approach that explores skill discovery by *autotelic* agents. Autotelic agents are agents that generate their own goals and learning signals. In this approach, intrinsic rewards guide both the creation and selection of goals. Once the agent forms the goals, it explores the goals, and the information learned during this exploration also serves to find solutions for other goals. This approach has been demonstrated on a humanoid robot that explores multiple spaces of goals with several hundred continuous dimensions and distractors Forestier et al. (2022). In the absence of a specific goal, the robot continues diversifying its skills which can later be employed in a complex goal.

These results suggest that the incorporation of intrinsic motivation might also be an important factor to keep in mind while building AGI systems that can develop life skills and remain life-long learners even in the absence of immediate goals or external rewards.

5. Forming Optimal Task-State Representations

A major part of reinforcement learning is designing the correct state space for the task at hand. This is crucial because tasks are not confined to a single task-state representation, and different state spaces can lead to different optimal policies. Additionally, the number of states must also be carefully chosen because too many states can slow down learning. In computational RL, state

spaces for the most part are assumed, and designed by experts. However, animals have to form their own state representations given a particular task.

This is particularly difficult given the different streams of constant sensory input animals experience during a task. Identifying task-relevant details in high-dimensional data is key to forming optimal task-state representations. Niv (2019) found that human subjects updated values of only the task-relevant dimension of the presented stimuli (one of shape, color, or texture) and not of the stimuli as a whole when only the relevant feature was predictive of reward. Increased attention was directed to the task-relevant dimension during the course of the experiment. This suggests that the brain employs attention-weighted reinforcement learning to perform a dimensionality reduction of the multiple streams of information coming its way to associate the most relevant features with reward. Leong et al. (2017) further demonstrated that RL signals in the brain are attention-filtered by fitting an eye-tracking and fMRI-derived measure of attention to computational models of participants' choices in an RL task.

Recent work has drawn inspiration from neuroscience to develop attention-weighted computational reinforcement learning algorithms that incorporate self-attention into a deep RL algorithm Bramlage and Cortese (2022). This attention-weighted deep RL network achieves a better success than comparable machine learning models in tasks with noisy, high-dimensional data.

6. Social Learning in the Reinforcement Learning Framework

A discussion about AGI drawing inspiration from animal learning would be incomplete without mentioning how animals learn from other members of their group, in addition to learning from their own direct experiences. Social learning occurs when groups of animals in a community learn about their environment and pass that information on to other members. In this mode of learning, there is usually no explicit teaching taking place, and the animal learns purely from watching the actions of others. This is extremely beneficial as members of a group can rely on others to guide their actions and behaviors when the cues from the environment aren't clear enough for them to arrive at a choice of their own accord. This could occur if the information they are receiving is too noisy, or if the cost of acquiring additional information is too high, be it in terms of the animal's energy or the risk it is subjected to Olsson, Knapska, and Lindström (2020). Even humans, arguably the smartest of all animals, cannot individually acquire all the knowledge they need to survive in their complex worlds and adapt quickly to changing environments, which is exactly why animals evolved to develop this incredible ability to learn from others and tweak it for the specific task at hand (Boyd, Richerson, and Henrich, 2011).

While AGI holds to benefit a lot from an incorporation of social learning, there is an underlying belief that this would involve the addition of special-purpose modules fundamentally different from those involved in individual learning, and that the learning mechanisms involved would also differ (Heyes, 2012). However, social learning has been shown to be possible even in typically asocial animals suggesting that such specialized modules might not be required. The octopus, which is largely a solitary animal, when provided with two colored objects repeatedly attacked the same colored object it had observed another octopus attack, in an experimental task (Fiorito and Scotto, 1992). This suggests that an animal that isn't designed to learn socially still possesses the architecture, and the learning algorithm to watch and learn if needed.

In a paper titled 'What's so social about social learning?' Cecilia Heyes, a professor of Psychology at the University of Oxford, made the striking claim that all of social learning can

be explained within the reinforcement learning framework we have discussed so far, for learning from direct experiences. She does so by highlighting the parallels between the two types of learning. For instance, she draws a parallel between stimulus-stimulus associations in Pavlovian learning and the phenomena of observational conditioning in the social context. An example of observational conditioning is monkeys who watch conspecifics be scared of snakes and also develop a fear of snakes from pure observation (Cook et al., 1985). Another parallel she draws is between instrumental conditioning where the animal selects an action to maximize its reward and imitation behaviors. An example of a human performing imitation could be seeing your friend leaving their shoes outside prior to entering their house, and you automatically also leaving your shoes outside without explicitly being told/rewarded to do so.

Learning from rewards being offered to other members of the group is called vicarious reinforcement learning, and there is a great deal of evidence for this happening in the animal world. Take for example bumblebees; bumblebees were made to observe conspecifics foraging through a screen, and the conspecifics had picked one of two flower colors to draw nectar from. Following this screening, the observers foraged alone and instantly copies the color preferences of their conspecifics as opposed to controls who foraged indiscriminately (Leadbeater and Dawson, 2017). The authors attributed this to second-order conditioning. Second-order conditioning is a Pavlovian phenomenon that takes place in two stages. In the first stage, a neutral stimulus (CS1) is paired with a naturally rewarded stimulus (US) till it acquires a conditioned response. In the second stage, a second neutral stimulus (CS2) is paired with the conditioned stimulus from the previous stage (CS1) and the animal acquires a conditioned response to the new neutral stimulus as well. In this example, the authors suggest that the nectar drawn from the flowers is the US, the other bees are CS1 and the flower color is CS2.

Vicarious reinforcement learning has also been seen in rhesus macaques who perform different categories of tasks, such that their performance in the tasks elicits a reward for one of three receivers themselves, a conspecific, or an inanimate object. While their performance in the tasks is the best when they are the ones to receive a reward, they perform significantly better in tasks that reward the conspecific than the one with the inanimate receiver (Chang, Winecoff, and Platt, 2011). These findings are increasingly being explained by the use of an associative learning framework, the same as the kind we observed in model-free reinforcement learning from direct experiences (Dawson et al., 2013; Lind, Ghirlanda, and Enquist, 2019). Additionally, these computational theories of social learning in an associative framework are finding neural credibility with research suggesting that networks for social and direct learning largely overlap, both in rodents and humans (Olsson, Knapska, and Lindström, 2020).

6.1 Insights for AGI from Social Learning in Animals

The world of artificial intelligence seems to have largely thought of intelligence in a social vacuum. However, social learning holds a lot of promise for AGI, especially in the context of multiple AGI systems interacting with each other. Within a reinforcement learning framework, multi-agent reinforcement learning is a setup wherein multiple RL agents coexist within a shared environment. These agents might cooperate or compete with each other, or engage in a mix of the two behaviors based on their goals. Social learning in AGI ‘societies’ can confer similar benefits to social learning in animal societies - learning from experts to quickly learn the optimal policy and avoiding costly exploration in the process.

There have been some initial attempts to incorporate social learning in multi-agent systems. Most of these attempts employ imitation learning, wherein new agents learn from ‘expert’ and more experienced agents. This can be thought of as a close parallel to vicarious reinforcement learning or learning by observation in animals. Most attempts to incorporate imitation learning involve the developer externally designing expert agents, and providing structured demonstration data for the tasks the new agents need to learn (Reddy, Dragan, and Levine, 2019; Ross, Gordon, and Bagnell, 2011). This results in a lack of generalizability, and policies that can never be better than the experts. However, humans and non-human animals are well equipped to identify experts and learn from them without external help. Another major issue with imitation learning in RL is that demonstration data is usually provided in the first person to the agent, as states it should have visited and actions it should have performed. However, animals learn from observing other animals perform tasks, by viewing states and actions in the third person. (Stadie, Abbeel, and Sutskever, 2019) use the demonstration of a teacher from a different viewpoint, combined with unsupervised teaching methods with no link between the teacher and student to make imitation learning a bit more biologically plausible. Another major difference is that imitation learning usually designates a single high-quality expert to a new agent, while animals learn from multiple other animals with varying degrees of expertise. Cheng, Kolobov, and Agarwal (2020) teach agents to combine information across multiple poor experts to speed up policy optimization in a novel agent.

Ndousse et al. (2021) comes the closest to mimicking vicarious learning in animals, combining an increased cost of individual exploration, the introduction of social prestige cues, and forcing agents to model other agents’ policies to incorporate social learning in multi-agent RL systems. Agents trained this way don’t simply copy experts’ policies but even improve upon them and learn to outperform experts. Additionally, these agents train in an interleaved manner, both by themselves and with an expert, decreasing dependence on the expert. Finally, the experts are not incentivized to teach other agents and they consequently do not engage in formal demonstrations.

From these approaches, it is apparent that incorporating social learning strategies into agents helps them learn an optimal policy more quickly and learn from far fewer samples. It also helps in generalizing what the agent learns to novel environments it has not previously encountered. It would be worthwhile to look deeper into animal social behavior to understand how experts are designated, and what the balance between social and solo learning looks like.

7. Meta-Reinforcement Learning in Animals – Learning to Learn

Animals can quickly adapt to novel tasks and environments that they haven’t been previously exposed to. This ability is of great interest for viewing AGI within a reinforcement learning framework, to allow for more general RL instead of separately trained RL algorithms for each minor variation in the task or environment. Interest in how animals learn how to learn has spurred recent neuroscientific research into meta-reinforcement learning in the brain. Meta-reinforcement learning occurs when humans and non-human animals map existing knowledge gleaned from prior experience onto the specifics of a new situation. This involves extracting the common structure across similar tasks abstracted away from sensorimotor specifics, storing this abstract structure, and applying it with task-specific modifications when presented with a related task. This makes learning a somewhat familiar task much faster than starting from scratch.

Zhou et al. (2021) found a neural signature of such an abstracted representation across a succession of odor-sequence problems in rats. They found that with each successive task with a similar structure, neuronal ensembles in the orbitofrontal cortex converged into a low-dimensional neural code that represented the task structure. They also found that the rate at which this neural code evolved increased with each new, similar task. (Samborska et al., 2022) trained mice on a reinforcement-learning paradigm, with the same abstract structure involving reversal-learning across tasks, but with different sensorimotor specifics across tasks. Specifically, they trained mice to poke their nose at one of two ports that had a high probability of reward and reversed the reward contingencies right after they learned which button was rewarded. The mice then had to learn to reverse their nose pokes. Across tasks, the physical locations of the two buttons varied. Interestingly, they found that although the physical layouts differed significantly across tasks, mice became quicker at reversing their preferences across tasks. Parallely, they found that the same set of neurons in the prefrontal cortex of the mice encoded the abstracted task structure across all task instantiations, while different sets of neurons encoded different instantiations in the hippocampus. They concluded that mice stored their policy integrated over multiple tasks in the prefrontal cortex, while the hippocampus stored concrete task representations. (Wang et al., 2018) further evaluated the role of the prefrontal cortex in meta-learning in animals. The dopaminergic system has long been thought of as the reinforcement learning system in the animal brain. However, it has been seen that even the prefrontal cortex has all elements typically associated with an RL system – prefrontal cortex neurons dynamically encode conversions from previous choices and corresponding rewards to values of states and actions, which then influence the animals’ new choices. (Wang et al., 2018) demonstrate that the dopaminergic RL system fine-tunes the weights of the prefrontal cortex modeled as a recurrent neural network, by broadcasting reward prediction error signals to it, leading the prefrontal cortex to store task abstractions.

Computational RL has begun attempts to incorporate this kind of meta-learning to help generalize learning across tasks and environments. Rohani, Hedayatian, and Baghshah (2022) built a Brain-Inspired Meta-Reinforcement Learning (BIMRL) deep RL system, where each layer of the network corresponds to a part of the prefrontal cortex. They also incorporate a memory module that represents the hippocampus, which stores memories both within each episode and across tasks. Within the memory module, task-agnostic internal reward guides exploration in novel environments. They find that incorporating these neuroscientific modules helps generalize learning across tasks, and suggest adding a long-term memory module for even better results. Other approaches incorporating meta-RL have also found success in online adaptation to rapidly changing environments and crippled body parts (Nagabandi et al., 2019), learning to explore more efficiently based on previous task experience (Gupta et al., 2018) and extracting task abstractions across similar tasks (Rakelly et al., 2019). From neuroscientific insights and implementation into computational RL, it appears that meta-reinforcement learning could help with generality across tasks in AGI viewed within a reinforcement learning framework.

8. Conclusion

The investigation into animal reinforcement learning offers rich insights for advancing AGI, with each area of study contributing specific and actionable guidance. The key insights are summarized as follows:

1. **Reward-Based Action Selection in Animals:** The balance between *model-based* (cognitive, planning-driven) and *model-free* (habitual, reflexive) reinforcement learning in animal behavior suggests AGI systems should be versatile in their learning approach. This balance could enable AGI to optimize decision-making, choosing between rapid, less energy-intensive responses and more deliberate, thoughtful actions. Additionally, insights into the contextual switching between these learning methods in animals can inform the development of AGI systems that can dynamically adjust their learning strategy based on the task complexity and environmental variables.
2. **Intrinsic Motivation:** The study of intrinsic motivation in animals, particularly its link to the dopaminergic system highlights the importance of non-tangible rewards in learning processes. AGI systems might benefit from algorithms that simulate intrinsic rewards, encouraging self-driven exploration and persistent engagement with tasks, even in the absence of external rewards. This approach could lead to AGI systems that are more autonomous and capable of self-improvement, driven by internal metrics of success and curiosity.
3. **Forming Optimal Task-State Representations:** Animals' ability to filter out irrelevant sensory information emphasizes the need for AGI systems to employ attention mechanisms. This can lead to more efficient data processing, where the system focuses on critical information relevant to the task at hand, thereby improving learning speed and effectiveness. Implementing attention-based mechanisms in AGI could result in systems that are not only better at learning from their environment but also more efficient in terms of computational resources.
4. **Social Learning in the Reinforcement Learning Framework:** Observational learning in animals suggests that AGI systems in multi-agent scenarios could benefit significantly from mechanisms that allow learning through observation of peers. This could reduce the need for individual trial-and-error learning, speeding up the learning process. Implementing social learning strategies in AGI could enable systems to quickly adapt to new environments or tasks by leveraging the knowledge and experiences of other agents, fostering collaborative learning and problem-solving.
5. **Meta-Reinforcement Learning in Animals – Learning to Learn:** Animals' ability to apply knowledge from previous tasks to new, similar tasks suggests that AGI should be capable of meta-learning. This would enable AGI systems to abstract commonalities across different tasks and apply this knowledge to new situations, reducing the need for learning from scratch. Incorporating meta-learning into AGI could result in systems that are not only quick to adapt to new challenges but also capable of evolving their learning strategies based on accumulated experiences.

In conclusion, the study of animal reinforcement learning mechanisms provides a wealth of knowledge that can be directly applied to the development of AGI systems within an RL framework. These insights range from enhancing predictive capabilities and decision-making processes to fostering intrinsic motivation, attention-focused learning, social learning dynamics, and meta-learning capabilities. Such advancements could lead to AGI systems that are not only efficient and adaptable but also capable of complex understanding and interaction within their environments.

Acknowledgements

I would like to thank Dr. John Laird for his contributions to developing the ideas, outline, and directions for this paper, and his insightful comments on the finished draft. I'm also grateful for Dr. Thad Polk's and Dr. Richard Lewis' thoughtful feedback on this paper. Finally, I also thank the Editor and the reviewers for their helpful feedback that helped shape this paper.

References

- Adam, S.; Busoniu, L.; and Babuska, R. 2012. Experience Replay for Real-Time Reinforcement Learning Control. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42(2):201–212. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).
- Adler, A.; Katabi, S.; Finkes, I.; Israel, Z.; Prut, Y.; and Bergman, H. 2012. Temporal Convergence of Dynamic Cell Assemblies in the Striato-Pallidal Network. *Journal of Neuroscience* 32(7):2473–2484. Publisher: Society for Neuroscience Section: Articles.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198 [cs].
- Alonso, E., and Schmajuk, N. 2012. Special issue on computational models of classical conditioning guest editors' introduction. *Learning & Behavior* 40(3):231–240.
- Balleine, B. W., and O'Doherty, J. P. 2010. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology* 35(1):48–69. Number: 1 Publisher: Nature Publishing Group.
- Barto, A. G. Adaptive Critics and the Basal Ganglia. 20.
- Barto, A. G. 2013. Intrinsic Motivation and Reinforcement Learning. In Baldassarre, G., and Mirolli, M., eds., *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin, Heidelberg: Springer. 17–47.
- Boyd, R.; Richerson, P. J.; and Henrich, J. 2011. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences* 108(supplement_2):10918–10925. Publisher: Proceedings of the National Academy of Sciences.
- Bramlage, L., and Cortese, A. 2022. Generalized attention-weighted reinforcement learning. *Neural Networks* 145:10–21.
- Buetti-Dinh, A.; Galli, V.; Bellenberg, S.; Ilie, O.; Herold, M.; Christel, S.; Boretska, M.; Pivkin, I. V.; Wilmes, P.; Sand, W.; Vera, M.; and Dopson, M. 2019. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports* 22:e00321.

- Byrnes, S. 2021. Reward Is Not Enough - LessWrong.
- Chang, S. W. C.; Winecoff, A. A.; and Platt, M. L. 2011. Vicarious reinforcement in rhesus macaques (macaca mulatta). *Frontiers in Neuroscience* 5:27.
- Cheng, C.-A.; Kolobov, A.; and Agarwal, A. 2020. Policy Improvement via Imitation of Multiple Oracles. arXiv:2007.00795 [cs, stat].
- Chentanez, N.; Barto, A.; and Singh, S. 2004. Intrinsically Motivated Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.
- Cook, M.; Mineka, S.; Wolkenstein, B.; and Laitsch, K. 1985. Observational conditioning of snake fear in unrelated rhesus monkeys. *Journal of Abnormal Psychology* 94(4):591–610. Place: US Publisher: American Psychological Association.
- Danner, F. W., and Lonky, E. 1981. A Cognitive-Developmental Approach to the Effects of Rewards on Intrinsic Motivation. *Child Development* 52(3):1043–1052. Publisher: [Wiley, Society for Research in Child Development].
- Daw, N. D.; Courville, A. C.; and Touretzky, D. S. 2006. Representation and Timing in Theories of the Dopamine System. *Neural Computation* 18(7):1637–1677.
- Daw, N. D.; Niv, Y.; and Dayan, P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8(12):1704–1711. Number: 12 Publisher: Nature Publishing Group.
- Dawson, E. H.; Avarguès-Weber, A.; Chittka, L.; and Leadbeater, E. 2013. Learning by Observation Emerges from Simple Associations in an Insect Model. *Current Biology* 23(8):727–730.
- de Bruin, T.; Tuyls, K.; Kober, J.; and Babuska, R. 2015. The importance of experience replay database composition in deep reinforcement learning. 9.
- Deci, E. L., and Ryan, R. M. 1985. Conceptualizations of Intrinsic Motivation and Self-Determination. In Deci, E. L., and Ryan, R. M., eds., *Intrinsic Motivation and Self-Determination in Human Behavior*, Perspectives in Social Psychology. Boston, MA: Springer US. 11–40.
- DeYoung, C. G. 2013. The neuromodulator of exploration: A unifying theory of the role of dopamine in personality. *Frontiers in Human Neuroscience* 7. Place: Switzerland Publisher: Frontiers Media S.A.
- Di Domenico, S. I., and Ryan, R. M. 2017. The Emerging Neuroscience of Intrinsic Motivation: A New Frontier in Self-Determination Research. *Frontiers in Human Neuroscience* 11:145.
- Doll, B. B.; Simon, D. A.; and Daw, N. D. 2012. The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology* 22(6):1075–1081.
- Fiorito, G., and Scotto, P. 1992. Observational Learning in *Octopus vulgaris*. *Science* 256(5056):545–547. Publisher: American Association for the Advancement of Science.
- Fjelland, R. 2020. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications* 7(1):1–9. Number: 1 Publisher: Palgrave.

- Forestier, S.; Portelas, R.; Mollard, Y.; and Oudeyer, P.-Y. 2022. Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning. *arXiv:1708.02190 [cs]*.
- Foster, D. J., and Wilson, M. A. 2006. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440(7084):680–683.
- Gershman, S. J., and Niv, Y. 2012. Exploring a latent cause theory of classical conditioning. *Learning & Behavior* 40(3):255–268.
- Gershman, S. J.; Markman, A. B.; and Otto, A. R. 2014. Retrospective revaluation in sequential decision making: a tale of two systems. *Journal of Experimental Psychology. General* 143(1):182–194.
- Gershman, S. J.; Moustafa, A. A.; and Ludvig, E. A. 2014. Time representation in reinforcement learning models of the basal ganglia. *Frontiers in Computational Neuroscience* 7:194.
- Gershman, S. J.; Norman, K. A.; and Niv, Y. 2015. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences* 5:43–50.
- Glimcher, P. W. 2011. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl 3:15647–15654.
- Gupta, A.; Mendonca, R.; Liu, Y.; Abbeel, P.; and Levine, S. 2018. Meta-Reinforcement Learning of Structured Exploration Strategies. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Gurland, S. T., and Grolnick, W. S. 2003. Children’s Expectancies and Perceptions of Adults: Effects on Rapport. *Child Development* 74:1212–1224. Place: United Kingdom Publisher: Blackwell Publishing.
- Harlow, H. F. 1950. Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of Comparative and Physiological Psychology* 43:289–294. Place: US Publisher: American Psychological Association.
- Heyes, C. 2012. What’s social about social learning? *Journal of Comparative Psychology* 126(2):193–202. Place: US Publisher: American Psychological Association.
- Ho-Phuoc, T. 2019. CIFAR10 to Compare Visual Recognition Performance between Deep Neural Networks and Humans. *arXiv:1811.07270 [cs]*. *arXiv: 1811.07270*.
- Holland, P. C. 2004. Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology. Animal Behavior Processes* 30(2):104–117.
- Houthooft, R.; Chen, X.; Duan, Y.; Schulman, J.; De Turck, F.; and Abbeel, P. 2017. VIME: Variational Information Maximizing Exploration. *arXiv:1605.09674 [cs, stat]*. *arXiv: 1605.09674*.
- Jones, S. H.; Gray, J. A.; and Hemsley, D. R. 1990. The Kamin blocking effect, incidental learning and psychoticism. *British Journal of Psychology (London, England: 1953)* 81 (Pt 1):95–109.

- Kahneman, D. 2011. *Thinking, fast and slow*. Thinking, fast and slow. New York, NY, US: Farrar, Straus and Giroux. Pages: 499.
- Leadbeater, E., and Dawson, E. H. 2017. A social insect perspective on the evolution of social learning mechanisms. *Proceedings of the National Academy of Sciences* 114(30):7838–7845. Publisher: Proceedings of the National Academy of Sciences.
- Leong, Y. C.; Radulescu, A.; Daniel, R.; DeWoskin, V.; and Niv, Y. 2017. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron* 93(2):451–463.
- Lind, J.; Ghirlanda, S.; and Enquist, M. 2019. Social learning through associative processes: a computational theory. *Royal Society Open Science* 6(3):181777.
- Ludvig, E. A.; Sutton, R. S.; and Kehoe, E. J. 2008. Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System. *Neural Computation* 20(12):3034–3054.
- Ludvig, E. A.; Sutton, R. S.; and Kehoe, E. J. 2012. Evaluating the TD model of classical conditioning. *Learning & Behavior* 40(3):305–319.
- Maia, T. V. 2009. Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience* 9(4):343–364.
- Mohamed, S., and Rezende, D. J. 2015. Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning. *arXiv:1509.08731 [cs, stat]*. arXiv: 1509.08731.
- Momennejad, I.; Russek, E. M.; Cheong, J. H.; Botvinick, M. M.; Daw, N. D.; and Gershman, S. J. 2017. The successor representation in human reinforcement learning. *Nature Human Behaviour* 1(9):680–692. Number: 9 Publisher: Nature Publishing Group.
- Montague, P. R.; Dayan, P.; Person, C.; and Sejnowski, T. J. 1995. Bee foraging in uncertain environments using predictive hebbian learning. *Nature* 377(6551):725–728. Number: 6551 Publisher: Nature Publishing Group.
- Nagabandi, A.; Clavera, I.; Liu, S.; Fearing, R. S.; Abbeel, P.; Levine, S.; and Finn, C. 2019. Learning to Adapt in Dynamic, Real-World Environments Through Meta-Reinforcement Learning. *arXiv:1803.11347 [cs, stat]*.
- Ndousse, K.; Eck, D.; Levine, S.; and Jaques, N. 2021. Emergent Social Learning via Multi-agent Reinforcement Learning. *arXiv:2010.00581 [cs, stat]*.
- Niemiec, C. P., and Ryan, R. M. 2009. Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education* 7(2):133–144. Publisher: SAGE Publications.
- Niv, Y. 2009. Reinforcement learning in the brain. *Journal of Mathematical Psychology* 53(3):139–154.
- Niv, Y. 2019. Learning task-state representations. *Nature Neuroscience* 22(10):1544–1553.

- Olsson, A.; Knapska, E.; and Lindström, B. 2020. The neural and computational systems of social learning. *Nature Reviews Neuroscience* 21(4):197–212. Number: 4 Publisher: Nature Publishing Group.
- OpenAI. 2021. DALL·E: Creating images from text. <https://openai.com/research/dall-e>.
- OpenAI. 2022. Introducing ChatGPT. <https://openai.com/blog/chatgpt>.
- OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs].
- Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-Driven Exploration by Self-Supervised Prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 488–489. Honolulu, HI, USA: IEEE.
- Pellis, S. M., and Burghardt, G. M. 2017. Play and exploration. In *APA handbook of comparative psychology: Basic concepts, methods, neural substrate, and behavior, Vol. 1*, APA handbooks in psychology®. Washington, DC, US: American Psychological Association. 699–722.
- Rakelly, K.; Zhou, A.; Finn, C.; Levine, S.; and Quillen, D. 2019. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *Proceedings of the 36th International Conference on Machine Learning*, 5331–5340. PMLR. ISSN: 2640-3498.
- Reddy, S.; Dragan, A. D.; and Levine, S. 2019. SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards. arXiv:1905.11108 [cs, stat].
- Rescorla, R., and Wagner, A. 1972. A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. *Classical Conditioning: Current Research and Theory*.
- Rohani, S. R. R.; Hedayatian, S.; and Baghshah, M. S. 2022. BIMRL: Brain Inspired Meta Reinforcement Learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9048–9053. arXiv:2210.16530 [cs].
- Roitblat, H. 2021. Building artificial intelligence: Reward is not enough. <https://bdechtalks.com/2021/07/07/ai-reward-is-not-enough-herbert-roitblat/>.
- Ross, S.; Gordon, G. J.; and Bagnell, J. A. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. arXiv:1011.0686 [cs, stat].
- Russek, E. M.; Momennejad, I.; Botvinick, M. M.; Gershman, S. J.; and Daw, N. D. 2017. Predictive representations can link model-based reinforcement learning to model-free mechanisms. Technical report, bioRxiv. Section: New Results Type: article.
- Samborska, V.; Butler, J. L.; Walton, M. E.; Behrens, T. E. J.; and Akam, T. 2022. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature Neuroscience* 25(10):1314–1326. Number: 10 Publisher: Nature Publishing Group.
- Schultz, W.; Apicella, P.; and Ljungberg, T. 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 13(3):900–913.

- Shapira, Z. 1976. Expectancy determinants of intrinsically motivated behavior. *Journal of Personality and Social Psychology* 34:1235–1244. Place: US Publisher: American Psychological Association.
- Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489. Number: 7587 Publisher: Nature Publishing Group.
- Silver, D.; Singh, S.; Precup, D.; and Sutton, R. S. 2021. Reward is enough. *Artificial Intelligence* 299:103535.
- Singh, S.; Lewis, R. L.; Barto, A. G.; and Sorg, J. 2010. Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. *IEEE Transactions on Autonomous Mental Development* 2(2):70–82. Conference Name: IEEE Transactions on Autonomous Mental Development.
- Singh, S.; Lewis, R.; and Barto, A. 2009. Where Do Rewards Come From? *Proceedings of the annual conference of the Cognitive Science Society* 2601–2606.
- Stadie, B. C.; Abbeel, P.; and Sutskever, I. 2019. Third-Person Imitation Learning. arXiv:1703.01703 [cs].
- Sutton, R. S. 1991. Dyna, an Integrated Architecture for Learning, Planning, and Reacting.
- Tricomi, E., and DePasque, S. 2016. The role of feedback in learning and motivation. *Advances in Motivation and Achievement* 19:175–202. Publisher: Emerald Group Publishing Ltd.
- Tschandl, P.; Rosendahl, C.; Akay, B. N.; Argenziano, G.; Blum, A.; Braun, R. P.; Cabo, H.; Gourhant, J.-Y.; Kreusch, J.; Lallas, A.; Lapins, J.; Marghoob, A.; Menzies, S.; Neuber, N. M.; Paoli, J.; Rabinovitz, H. S.; Rinner, C.; Scope, A.; Soyer, H. P.; Sinz, C.; Thomas, L.; Zalaudek, I.; and Kittler, H. 2019. Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA Dermatology* 155(1):58–65.
- Vamplew, P.; Smith, B. J.; Kallstrom, J.; Ramos, G.; Radulescu, R.; Roijers, D. M.; Hayes, C. F.; Heintz, F.; Mannion, P.; Libin, P. J. K.; Dazeley, R.; and Foale, C. 2021. Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). arXiv:2112.15422 [cs]. arXiv: 2112.15422.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762 [cs].
- Waltz, D. L. 1988. The Prospects for Building Truly Intelligent Machines. *Daedalus* 117(1):191–212. Publisher: The MIT Press.
- Wang, J. X.; Kurth-Nelson, Z.; Kumaran, D.; Tirumala, D.; Soyer, H.; Leibo, J. Z.; Hassabis, D.; and Botvinick, M. 2018. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience* 21(6):860–868. Number: 6 Publisher: Nature Publishing Group.

- Yin, H. H.; Ostlund, S. B.; Knowlton, B. J.; and Balleine, B. W. 2005. The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience* 22(2):513–523. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2005.04218.x>.
- Yin, H. H.; Knowlton, B. J.; and Balleine, B. W. 2004. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience* 19(1):181–189. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2004.03095.x>.
- Zhang, S., and Sutton, R. S. 2018. A Deeper Look at Experience Replay. *arXiv:1712.01275 [cs]*. arXiv: 1712.01275.
- Zhou, J.; Jia, C.; Montesinos-Cartagena, M.; Gardner, M. P. H.; Zong, W.; and Schoenbaum, G. 2021. Evolving schema representations in orbitofrontal ensembles during learning. *Nature* 590(7847):606–611. Number: 7847 Publisher: Nature Publishing Group.