DEINFORCEMENT LEARNING

Chaytan Inman
Undergraduate Student
University of Washington
Seattle
chaytan@uw.edu

Varun Ananth
Undergraduate Student
University of Washington
Seattle
vananth3@uw.edu

Marlene Grieskamp
Undergraduate Student
University of Washington
Seattle
markamp@uw.edu

Janna Hong
Undergraduate Student
University of Washington
Seattle
jannahg@uw.edu

ABSTRACT

Reinforcement learning (RL) in the field of machine learning shares a mathematical basis with dopaminergic responses to reward conditioning in the field of neuroscience. This can be most clearly seen in RL paradigms such as A2C (Niv 2009). Although mimicking this biological parallel has advanced more generalized computational learning (Tassa et. al 2018), current RL algorithms are incomplete models of the neuroscientific understanding of the brain. Mirroring dopamine pathways alone lacks one of biology's most potent signals: pain. Often in RL a lack of a reward signal, or a negative signal, is seen as analogous to punishment (Schultz et. al 1997). However, based on the biological mechanisms that induce pain in the body this cannot be the case. Incorporating pain into current RL models would not only allow algorithms to converge faster, but for behavior to become more complex and generalizable, if the stimulus of pain represents broadly avoidable behaviors. This paper first summarizes the broad contributions and connections between reinforcement learning and neuroscience, and then proposes several biologically inspired methods of incorporating pain into RL algorithms.

***K*eywords** Machine Learning · Reinforcement Learning · Neuroscience · Cognitive Science · Psychology

Introduction

Neuroscience and reinforcement learning have long benefitted from shared concepts of learning. Reinforcement learning took a turn in the 1980s when Richard Sutton, educated in both psychology and computer science, created an algorithm called 'temporal difference learning' (TD learning) to explain the response to violation of expectation. The difference between the expected reward and the actual reward defines the prediction error. According to TD learning, this is how learning occurs – through prediction error. In 1997, Wolfram Schultz described the striking resemblance between dopamine firing rates and the reward prediction error signal.

Schultz showed that when monkeys gain an unexpected reward, their brains responded with a burst of action potentials, which is commonly known as a phasic firing pattern. When the reward is expected, however, the phasic pattern is no longer evident. These experiments provided evidence that dopaminergic neurons followed spiking patterns that encoded TD error, or the difference between the expected firing and actual firing patterns.

At the same time, developments in reinforcement learning used TD error to update the weights of neurons in artificial neural networks. In this way, reinforcement learning has mirrored concepts from human-like learning to create novel optimization algorithms. Popular algorithms such as actor-critic reinforcement learning are still based on TD error.

Seeing how critical programmatic algorithms mirroring ancient biological algorithms of the brain have become relevant to the field of reinforcement learning, taking a deep dive into the learning mechanisms of the brain will help us understand how to advance reinforcement learning further. We start with a widely branded pleasure neurotransmitter: dopamine.

Dopaminergic Learning

Dopamine is a neurotransmitter with a reputation in popular culture for its reward and pleasure properties. Studies also reveal that the same dopamine regions are involved in pain, which may involve the reorganization of the components of reward circuitry. Approximately 90% of dopamine producing neurons are in two areas of the midbrain nuclei called the *substantia nigra pars compacta* (SNc) and the *mesolimbic ventral tegmental area* (VTA) (Arias-Carrión et al., 2014). These neurons project to the *nucleus accumbens* (NAc), which is the reward-related dopamine site. Blocking the dopamine pathway to NAc deprives the rewarding effects.

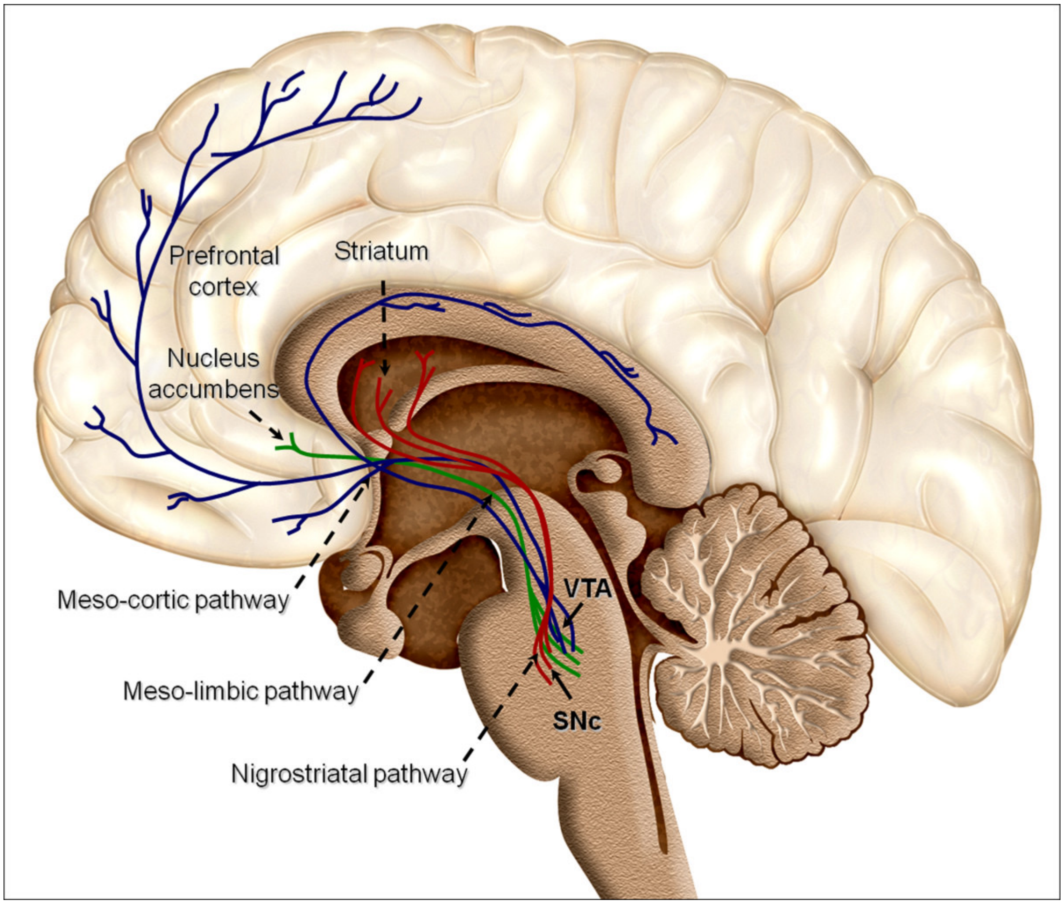

Figure 1: The mesolimbic pathway shows the dopaminergic neurons made in the ventral tegmental area projecting to the nucleus accumbens. (Arias-Carrión, O. et al., 2014)

In the early 1950s, scientists discovered the significance of the VTA region by observing the effects of electrical stimulation to certain regions of the brain in rats. When electrical stimulation in the VTA was followed by a certain task such as pressing on a lever, the rats performed that specific task repeatedly (Olds & Milner, 1954). The rats ended up pressing on the lever 2000 times per hour when they learned that this specific behavior reliably leads to an electrical stimulation. In this case, the action of pulling on the lever is followed by the electrical stimulation, thus the stimulus is the “reward”. The concept of reward is further described by Olds and Milner: “In its reinforcing capacity, a stimulus increases, decreases, or leaves unchanged the frequency of preceding responses, and accordingly it is called a reward, a punishment, or a neutral stimulus” (Olds & Milner, 1954, p. 419). Recent experiments in humans undergoing deep brain stimulation (DBS) for Parkinson’s show similar results. After participants learned that a certain task was followed by electrical stimulation of the SNc, an area with abundant dopaminergic neurons, they repeatedly performed that task, soon even without the stimulus (Perelman School of Medicine at the University of Pennsylvania, 2014).

Dopaminergic neurons (DA) can fire in two distinct patterns in response to varying stimuli: phasic and tonic activity. Phasic activity refers to a burst of action potentials firing in a short period of time, with a rate of up to 20Hz. In contrast, tonic activity indicates a steady firing rate of around 5 Hz. Tonic activity recorded in monkeys signals congruency between actual and expected reward, while phasic signal indicates a component of surprise. (Schultz et al.,

1997) This experiment showed that expected reward consistently decreases dopaminergic responses, leading it to a tonic firing activity. Likewise, expectation of reward reduces the phasic firing activity after the reward is given. DA can signal the difference between the expected reward and the reward it actually receives. The activation or inhibition of dopamine neurons causes positive or negative conditioning, respectively.

Reinforcement Learning

Reinforcement Learning (RL) is a branch of machine learning involved with choosing optimal policies, state value estimates, or both to optimize an agent to select the most rewarding action in a given environment. RL has been used by companies such as DeepMind to train humanoid and non-humanoid physical models to walk, run, jump, and play games (Tassa et. al 2018). It tends to perform best in scenarios where there are complex states and decisions are plentiful, not dissimilar to our own environment. The advantage actor-critic (A2C) RL model employs TD learning to adjust its reward predictions through time, analogous to how humans learn through classical and operant conditioning (Niv 2009). **We have chosen to study A2C because it contains the simplest implementation of advantage.** The importance of the "advantage" variable and its connections to reward prediction error (RPE) in the brain will be discussed later. First, an understanding of the separate actor and critic elements is needed.

The actor is a neural network which learns policy $\boldsymbol{\pi}$ parameterized by $\boldsymbol{\theta}$, that is a function of state **s**. A policy determines which action should be taken in a given state. Policies can be tuned to maximize reward (as is the case with A2C), or achieve a goal parameterized as a function of that reward. Actors in A2C are stochastic by nature, meaning they output a probability distribution for taking an action in the action space based on the current state (Geron 2019). This has high level parallels to how we as humans interact with our environment. We consider the "state" of the world around us and then "act" to maximize some reward tied to a goal. For example, when you decide to take action A to walk your dog, one might think of this as applying your internal policy ($\boldsymbol{\pi}$) to the state **s**, wherein your dog is barking at the door wanting to go for a walk. In many cases, the actor is a deep neural network with input dimensions being equivalent to the dimensionality of the state, and the output dimensionality being equivalent to that of the action space. Actions selected by the actor are taken in the environment, physical or simulated, and they will receive a scalar reward for taking that action. The actor is updated with the critic along the "advantage" variable, which will be discussed in a later section.

The critic network in A2C is more abstract. Its purpose is to approximate the function $V(s)$. This function returns the overall value of being in a state **s**, given a policy $\boldsymbol{\pi}$. It is equivalent to the expected return of starting in state "**s**" and following policy $\boldsymbol{\pi}$ thereafter (return in terms of discounted rewards) (Mnih et. al 2016).

$$V^{\pi}(s) = E_{\pi}\{R_t|s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma r(s_{t+k+1}, a_{t+k+1})|s_t = s\}$$

The "perfect" V function shown above is in most cases incomputable. Such is the case when we do not have access to the rewards for any given state, which we do not when we do not have a model of the environment. V(s) is, however, commonly estimated by another deep neural network: in our case the critic. The input of the critic is the state and the output is a scalar estimate of the value function.

Now that we have covered RL and dopaminergic learning, we will study their intersection and the importance of extending this synergy beyond reward-based learning.

Dopaminergic Influences on Actor-Critic Systems

Temporal Difference learning is the framework upon which critic updates are based. The loss of a critic is the advantage (**A**) squared; the advantage can be broken down to its constituent parts below:

$$A = r(s_t, a_t) + \gamma V(s_{t+1}) - V(s_t)$$

Assuming the agent just took an action and has moved to a new state: $r(s_t, a_t)$ is the reward given for taking that certain action in that certain state. Reward is often given by the environment after an action is taken and not model-intrinsic. The $\gamma V(s_{t+1})$ term represents the estimated future discounted rewards starting at the new state the agent has just entered. Finally, the $V(s_t)$ term is the estimated value function for the state the agent was just in before acting. Combining the first two terms, $r(s_t, a_t) + \gamma V(s_{t+1})$, we get the reward the agent received for taking an action in its previous state plus the discounted predicted value of the agent's current state. We call this the TD Target. Recall that the critic network is being trained to accurately estimate $V(s)$. To train the network with backpropagation, we need some ground truth to understand how well the model predicted the V function for this timestep. The TD Target contains the ground truth in the form of $r(s_t, a_t)$. $V(s_t)$ is subtracted from the TD target, and the difference between these two is the TD Error; it is, in essence, an "error" between the actual received rewards and the estimated rewards. Note that the $\gamma V(s_{t+1})$ term is partially nullified by the tail end of the $V(s_t)$ term, since this value function contains discounted value predictions for future states as well (Watabe-Uchida et. al 2017). Squaring the advantage allows us to treat negative and positive TD Error equivalently while preserving the differentiability of the function.

Because the calculation of advantage in actor-critic systems is based on TD Learning, A2C weight updates have correlation to dopaminergic learning in the brain. In the case of RPE, as more stimuli are experienced by the organism and dopaminergic learning occurs (see Dopaminergic Learning section), the RPE approaches zero. This means the organism has learned how to correctly predict the reward given its state (Schultz et. al 1997). Just as RPE approaches zero while an organism learns to predict reward, the advantage variable in an actor-critic system approaches zero as the critic learns to estimate V(s).

If no reward is present but reward is predicted, then dopamine activity is heavily depressed which causes updates to the organism's "value estimation function". However, an important distinction needs to be drawn between the lack of a reward and a punishment. In traditional RL, negative rewards stemming from a state-action pair are seen as "punishment" for a model. Initially, this seems accurate. Value updates propagate through a network, telling the agent that this state is not as valuable as it initially estimated. However, if we consider on a high level the effect of pain, we see that there is a large difference in how organisms react to pain versus negative reward. If a child puts their hand on a hot plate, they are unlikely to do it again, effective immediately. An A2C "child" agent would perhaps place its hand on the hot plate multiple times under the guise of "exploration". The result: slow model convergence, experienced by almost all modern reinforcement learning algorithms (Ghiassian 2020). Learning for these models can be made more efficient.

In psychology's operant conditioning, a response cost punishment is a "negative reward", or the removal of a positive stimulus. This is vastly different from aversive punishment, the addition of a negative stimulus. Consider a child who likes climbing trees. The parents may warn them of the dangers of falling and even punish them for climbing by taking away their video game privileges (response cost punishment), but the child will likely continue doing so until one day they break their leg from falling. The painful broken leg (an aversive punishment) is a much faster and stronger conditioning response than losing gaming privileges, not only because it is more disruptive to the child, but also because it is an immediate negative stimulus specific to the event of climbing. In other scenarios, such as doing homework and getting a problem wrong, aversive punishment is much less effective than response cost because it may simply discourage the child from attempting the homework in the first place. From these principles of psychology, it is clear that both forms of punishment are required for efficient learning.

We argue that while current learning algorithms have successfully incorporated pleasure-based, operant conditioning-like paradigms as described above, the neurological mechanisms by which humans learn from pain are fundamentally different. Their implementation has the potential to mirror human-like learning in RL.

Dopamine and Pain

The experience of pain teaches us to avoid actions that detrimentally impact our wellbeing. For the scope of this paper, we will ignore emotional pain. However, we can confidently say physical pain helps an organism maintain homeostasis by detecting the body's internal imbalances. Unlike the condition of reward which attracts an action towards the condition, the condition of pain triggers the action that escapes from the condition and seeks relief. In other words, dopamine in these regions is essential for motivation and the "reward" for pain relief (DosSantos et al., 2017).

How is reward and learning circuitry connected to pain? Contrary to previous literature, pain can excite areas of the brain neurons also involved in our reward circuitry. The heterogeneity of dopamine neurons means that this circuitry cannot be generalized to one function, and the time and loci of dopaminergic firing during pain is best carefully examined.

The mesolimbic reward circuitry, including dopaminergic projections from VTA to the NAc, is involved in the modulation of pain. In rodents, prolonged pain triggers dopamine release in the NAc (Schmidt et al., 2002). Painful events can rapidly excite the dopaminergic neurons in the VTA. Not only do dopaminergic neurons process both acute and chronic pain, it also modulates pain relief (Wood, 2006).

Previous studies identified phasic firing activity to painful stimuli in areas of the brain thought to receive dopamine input (Chudler, 1993). Recent studies confirm this by showing that the DA neurons in the VTA are phasically activated by noxious footshocks in rodents (Brischoux et al., 2009). Brischoux et. al (2009) showed that DA neurons in the ventral VTA clearly responded in the first 500 ms upon a painful stimulus. The dopamine neurons in the dorsal VTA that were unresponsive to the footshock revealed excitation at the termination of an unfavorable footshock, where the neuron excitation peaked at 100 to 150 ms after the painful stimulus. The authors note that this may explain why "dopamine receptor antagonists interfere with avoidance learning, where the rewarding role of the offset of an aversive stimulus drives behavior" (Brischoux et al., 2009). However, (Moutoussis et al., 2008) suggested that avoiding pain leads to phasic firing activity of dopaminergic neurons. Thus, pain-induced dopaminergic firing is closely associated with learning pain aversive behavior.

It is important to note, however, that these studies do not equate the mechanism of learning from painful stimuli to that of reward stimuli. Although reward is clearly implicated in pain and pleasure, physiological response, learning rate, observed behavior, and our own experience draws a clear distinction between the two. The pathway through the brain by the painful signal is in itself highly distinct from that of a normal reward. Foremost, acute pain begins at nociceptors – specific receptors of somatic neurons that detect noxious stimuli apart from other stimuli. Moreover, one of the first modules of the pain pathway in the brain is the thalamus; typical reward circuitry does not involve the thalamus. Further distinct differences are shown in Figure [Figure comparing the pain pathway in brain with reward circuit]. Unfortunately, the exact neuromodulatory systems of learning occurring by avoided pain or received pain is largely unknown.

Discussion

Thus, it is clear that the reinforcement learning paradigm lacks the ability to sense pain in any capacity similar to human nociception. Furthermore, learning from an action resulting in negative reward mirrors neither observed human behavior in aversive punishment nor the neural circuitry involved in processing pain. This is in contrast to the rough parallels of reward based classical or operant conditioning to reinforcement learning via reward prediction error. Therefore, we propose several ways to incorporate pain into reinforcement learning, which we call deinforcement learning.

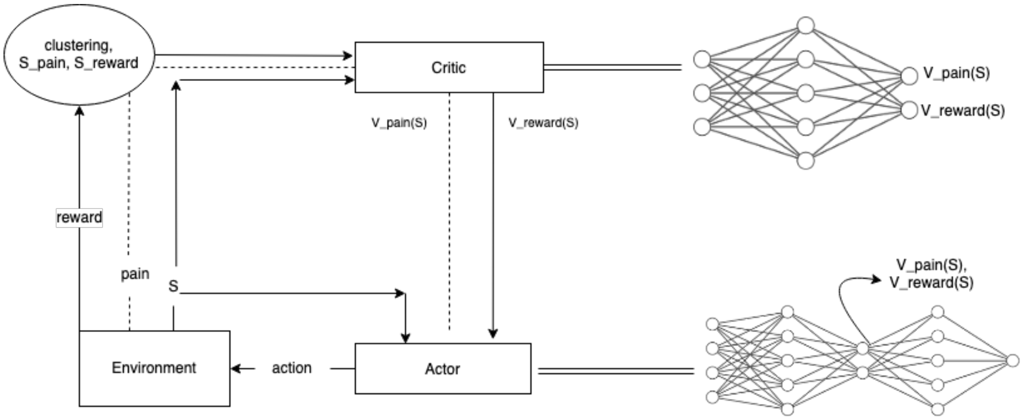

Figure 2: Possible Implementations of Pain in A2C
Dashed lines represent additional channels of information passage between networks in contrast to traditional A2C. Double solid lines indicate a suggested architecture. From possible incorporation of pain as an environmental signal to interpretation at the critic or actor level, implementation options are varied.

There appears to be two possible ways to represent pain in the context of reinforcement learning. One must distinguish between a painful state and a not painful state. This can either be the job of the environment, or the agent. In the former proposition, painful states may be a completely separate input, labeled as painful or not by the environment itself. For example, in the context of the game chess, the environment could send painful signals when pieces are lost, and reward when pieces are captured. This type of approach is seen in the current practice of reward shaping, but lacking a concept of pain (though the reward could have a negative sign). But this is not how humans perceive pain. The universe does not define pain for us. Fundamentally, it is the latter approach, the one that passes a raw state to the agent and allows the agent to interpret what is reward and what is pain, that is biologically inspired.

The path of pain through the body begins with nociception: the distinction between cells that can receive painful input and those that do not. When you touch something, a signal is propagated along mechanoreceptors. If you touch something hard enough, pointy enough, or hot enough, the signal propagates along pain specific fibers (i.e. A-fibers, A-delta, C-fibers) to signal an acute pain to the brain. As previously discussed, this pathway is disparate from that of an unpainful signal. The mechanisms that perceive pain and other stimuli fundamentally represent the state differently, before interpretation in the brain. How would this look in reinforcement learning? This may take the form of a clustering algorithm whose clusters represent painful vs rewarding stimuli and various interpolations of those classes. After clustering into discrete signals, the pain and reward signals could be processed and interpreted with different learning mechanisms as they are in the brain. It may take the form of a classification neural network (or support vector machines among other algorithms), whose logit probabilities can be interpreted as dimensions along various sensory stimuli such as touch, pain, temperature etc. This leaves the state open to interpretation by multiple perception pathways; you can not only feel pain when pricked by a needle, you can also feel pressure. There are many other possibilities for representing this distinction between pain and other sensory information at the initial reception at the sensory level.

If there is now a scalar describing the painfulness of a state, we need to augment the state to contain this new knowledge. This will allow us to describe to another network the painfulness of a state. One possibility is to use the positional encoding technique used by (Vaswani et. al) in Attention is All You Need.

Next, how does one interpret pain within the RL equivalent of a brain – in our explained example, the A2C method (see Reinforcement Learning section) uses the critic to evaluate how valued a state is with respect only to estimated future reward. Because in human behavior reward and pain pathways trigger learning at different rates, it is necessary to have different representations of V(s) with respect to pain, and with respect to reward such that learning can be modulated according to the painfulness of experiences. One way to achieve this is to first modify the state with information from the pain classification processing mentioned above. Then pass this information to the critic, whose weights should implicitly learn a representation of V(s) which estimates and takes into account both the estimation of future pain and reward, then outputs V(s) with respect to both of those estimations (Figure 1). This approach may be less efficient than explicitly modeling the estimation of V(s)pain and V(s)reward, but allows the critic network to learn how to weight the two values and combine them into an overall V(s).

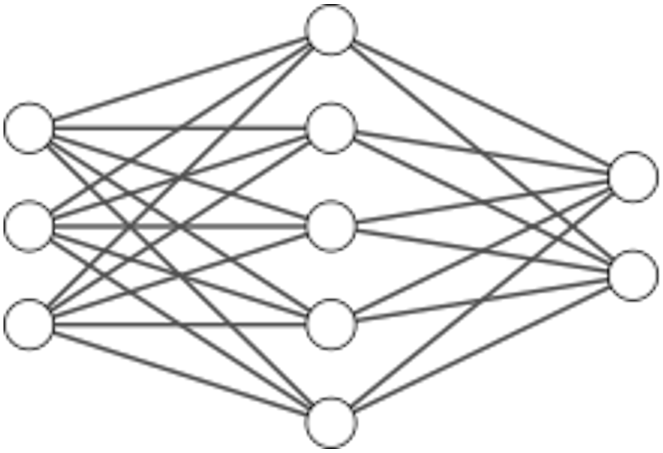

Figure 3: Simplified Critic with $V_pain(S)$ and V_reward Outputs

Alternatively, the modified state information is passed to the critic, which must learn to accurately represent the state as estimations of V(s)pain and V(s)reward and pass these values to the actor. The actor can then learn weights to combine pain and reward into a customary single V(s) rather than the two layer output in Figure 1. This shifts the burden of estimating $V_pain(S)$ and $V_reward(S)$ to the actor, as in Figure 2.

Another possible implementation is to train separate critic algorithms after binary classification of pain or non-painful stimuli. This can be likened to ensemble approaches. This approach is the least biologically faithful, since states are not best represented by such a binary classification, but it may prove more efficient with less data.

As we have seen, reward and motivation to escape is caused by painful stimuli. To mirror this type of aversion, one could use concepts like memory buffers, and algorithms may learn to associate the end of painful states with higher values, increasing the expected reward term as in TD learning.

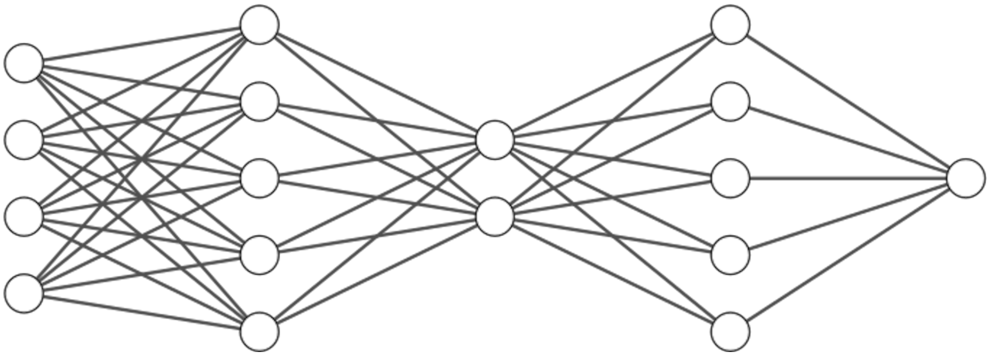

Figure 4: Actor with $V_pain(S)$ and $V_reward(S)$ in Third Layer

Finally, within this framework there is the question of what is painful. Secondarily, how do we construct a state provided by the environment that might allow us to learn what is painful? This may be the most difficult component of incorporating pain in reinforcement learning. For humans, this begins as well defined by genetics. We learn what is painful through evolutionary genetic iterations. Each iteration further defines reward as that which approaches propagating our genes, and pain as that which must be avoided to continue propagating these genes. Along with genetics, what is painful is learned through experiences as well as internally modulated through complex top-down modulatory pathways, beyond the scope of the neuroscience described in this paper.

The most literal machine learning analog might be an actor which can reproduce or spawn new networks with its learned weights. The probability of reproduction in this context would need to be correlated with the problem that the network is trying to solve. The network would also need a reward heuristic, such as time alive or reproductive success. It would learn as described above, making estimations of how painful or rewarding the environment might be. If the pain estimate is not correct, this negatively affects the network's predictions, and it would be less likely to reproduce. If an inaccurate pain estimation led to a very low performing model, the agent may be deactivated, or effectively killed. Thus, a successful, fit network should learn to define pain in a similar manner to humans, as that which should be avoided for the sake of reproduction. Reproductive odds defined within the genetic algorithm provides a separate signal to learn from besides the immediate-term reward signal and potential pain signal. For example, if the network should learn to make a stick figure walk like in OpenAI's MuJoCo framework, then reproductive success may be set as a function of time spent walking versus energy expended. Then networks that avoid fatal falls or expend less energy in the movement should have a higher probability of reproducing or replicating their weights in new networks. Here, the initial sensory layer would group together similar states and outcomes based on their features, modulate the given state of the environment to hold this painful information, then finally pass it to a critic. The critic estimates the value of the state with respect to potential pain and reward. If the agent were near a box it might trip over, the current sensed pain may be 0, but the critic may weigh future states as very painful and $V_pain(s)$ very low. Its estimation $V_pain(s)$ would be tuned as it attempts to walk, and further tuned as it replicates its weights in other agents based on its walking success. To completely mimic evolution, the "problem" would be replication itself, and the goal would be for the actor to take actions that replicate itself, which may be writing or executing code.

One may not find a need for literal analogs such as genetic algorithms. In the above case, one implicitly defines pain merely by defining what success is. In the MuJoCo case we did this – pain was implicitly that which must be avoided to achieve success of walking, for example, tripping. However, a model can receive a single signal (like reward in the current paradigm) and learn pain aversive behavior. Crucially this must split the signal with some algorithm into pain and pleasure, which together update the model's policy in dissimilar ways, just as the proposed algorithms in Figure X describe.

In environments where reward is uniformly distributed across all possible outcomes, such as a binary right/wrong object classification, there is no purpose to learning a function to approximate pain. The purpose of pain is to learn to avoid certain states much more vehemently and faster than learning from a lack of a reward in the same situation. If a misclassification is always weighted the same as any other misclassification, there is no distinction between pain and a lack of a reward.

Those familiar with RL may now be wondering: an initial clustering algorithm, and two node output of the critic instead of one node – is this really a significant change from standard RL practices? The key difference is not only in these simple, fundamental algorithmic changes but in coupling them with environmental changes conducive to pain aversive learning. As stated above, without an environment where correctness or incorrectness is non-uniformly distributed, pain is not a useful concept. This means that to test the incorporation of pain processing with something as simple as an MNIST classifier, one would need to quantify how close each classification was to the correct classification, and build that into the reward or ground truth. Ultimately, there are many possible scenarios where pain would increase convergence. Anywhere where particular states must be avoided more than a typical 'failure' is a good application of pain.

Conclusion

We have seen that reinforcement learning and neuroscience are complexly intertwined, beginning with their overlapping uses of reward prediction error and TD learning. Continuing to draw inspiration from the brain and body to enhance modern RL algorithms is, we believe, a fruitful frontier. Many areas of how humans learn from pain are yet to be investigated; the role of emotional pain and trauma was not examined in this paper. However, the growing neuroscientific body of knowledge on pain allows us to examine the phenomena as a model for RL algorithms. Thus, we conclude new RL models which learn pain aversive behavior are necessary to propel the field towards more realistic, efficient learning paradigms.

Acknowledgments

This work was created by the Interactive Intelligence team, a research group at the University of Washington fusing neuroscience and computer science to make machines learn like humans. We would like to thank the I2 advisor, Dr. Eric Chudler, for his support and guidance.

Thank you to everyone on the team, and everyone on and off it who helped make this paper and team possible.

References

Arias-Carrión, O., Caraza-Santiago, X., Salgado-Licona, S. et al. Orquestic regulation of neurotransmitters on reward-seeking behavior. *Int Arch Med* 7, 29 (2014). https://doi.org/10.1186/1755-7682-7-29

Brischoux, F., Chakraborty, S., Brierley, D. I., & Ungless, M. A. (2009). Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (12), 4894–4899. https://doi.org/10.1073/pnas.0811507106

DosSantos MF, Moura BS and DaSilva AF (2017) Reward Circuitry Plasticity in Pain Perception and Modulation. Front. Pharmacol. http://dx.doi.org/10.3389/fphar.2017.00790

Fu, B., Wen, S. N., Wang, B., Wang, K., Zhang, J. Y., Weng, X. C., & Liu, S. J. (2018). Gabapentin regulates dopaminergic neuron firing and theta oscillation in the ventral tegmental area to reverse depression-like behavior in chronic neuropathic pain state. *Journal of pain research*, 11, 2247–2256. https://doi.org/10.2147/JPR.S170167

Geron, A. (2019). Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

Ghiassian, S., Patterson, A., Garg, S., Gupta, D., White, A., & White, M. (2020). Gradient Temporal-Difference Learning with Regularized Corrections. ICML.

Horvitz J. C. (2002). Dopamine gating of glutamatergic sensorimotor and incentive motivational input signals to the striatum. *Behavioural brain research*, 137(1-2), 65–74. https://doi.org/10.1016/s0166-4328(02)00285-1

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. &amp; Kavukcuoglu, K.. (2016). Asynchronous Methods for Deep Reinforcement Learning. *Proceedings of The 33rd International Conference on Machine Learning, in Proceedings of Machine Learning Research* 48:1928-1937 Available from https://proceedings.mlr.press/v48/mniha16.html.

Perelman School of Medicine at the University of Pennsylvania. "Human learning altered by electrical stimulation of dopamine neurons." ScienceDaily. ScienceDaily, (2014) www.sciencedaily.com/releases/2014/05/140513175006.htm>.

Schultz W. (2006). Behavioral theories and the neurophysiology of reward. *Annual review of psychology*, 57, 87–115. https://doi.org/10.1146/annurev.psych.56.091103.070229

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. de L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., & Riedmiller, M. (2018). DeepMind Control Suite (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1801.00690

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (p./pp. 5998–6008), .

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural Circuitry of Reward Prediction Error. *Annual review of neuroscience*, 40, 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109

Wood P. B. (2006). Mesolimbic dopaminergic mechanisms and pain control. *Pain*, 120(3), 230–234. https://doi.org/10.1016/j.pain.2005.12.014