Simulating Addiction with Deep Q Learning by modifying Experience Replay Capacity

Dataset: Synthetic data generated from a simulated lever pulling reward environment.
People Associated: Dr. Luis Miralles
Pending Modules : Data Visualisation (Exam), Geospatial Information Systems (Exam), Deep Learning
Proposed Start Date: 5th Feb 2023
Submission Attempt: 1st

1 Background, Context and Scope

This research falls within the domain of reinforcement learning (RL). Reward seeking and past experiences will be modelled in a human behaviour psychology simulation to observe unhealthy decision making in an RL model.

The model's output is a decision not a prediction or clustering. There are no ground-truth decisions from a supervisor with a labelled dataset and although it can employ Supervised/Unsupervised learning to make decisions, it must learn from trial and error. Decisions are structured as a Markov Decision Process (MDP), based on the bellman equation where an agent performs an action in an environment of states (collections of features) to receive a new reward and a new set of states (Bellman, 1957; van Otterlo & Wiering, 2012).

Policy $\pi(s, a)$ maps actions to states. RL can use On-Policy or Off-Policy methods to learn while doing or learn, plan, then following that plan. Its goal is to maximise reward by updating a value function, which summarises the long-term effect of taking actions (Sutton & Barto, 2018).

RL can look for either the optimal state-action pair (Policy Based) or actions that give maximum reward (Value Based). It can also be Model-Based where a structured problem is given or Model-Free where the agent has no prior experience of the environment.

The model uses Temporal Difference to learn from each environment episode with no prior knowledge (Sutton, 1988).

Deep Q-Learning is a popular off-policy, value-based, model-free, RL method, where environment states combined with actions are given an estimated quality value or Q-value by a neural network. Actions with highest Q-values are selected as per equation below: (Schaul, Quan, Antonoglou, & Silver, 2015; Tokic, 2010).

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha(R(s,a)\gamma Max_a Q(s',a') - Q_{t-1}(S,A)) \quad (1)$$

RL is used in autonomous driving, recommendation systems, robotics, energy grid optimisation, fraud detection, pricing, and healthcare.(Li, 2019).

2 Problem Description

In healthcare, substance use disorders (addictions) are chronic, relapsing conditions that lead to clinically significant impairment or distress. Addictions take up 35.3% of a person's lifetime, costing $700 billion in USA alone. Symptoms, two of which are needed to classify addiction, include impaired control, physiological alterations and cravings. Many mathematical and brain based models of drug use and addition exist. However most models focus on the effects of drug use not addiction, have not been tested or supported by human data and few models capture multiple stages and addiction symptoms. To improve patients' lives, educate the public develop more efficient treatments and reduce the cost of addiction on the healthcare system, a better understanding of addiction neural circuitry is required (Mollick & Kober, 2020).

Brain behaviour to rewards (Serenko & Turel, 2020) can be modelled using Deep Q-Learning (DQL) in a lever pulling environment, grounded in extensive animal cocaine testing experiments (Keramati,

Durand, Girardeau, Gutkin, & Ahmed, 2017) and based on MDP (White, 1993), where hormones affect memory which can affect level of addiction (Popescu, Marian, Miruna, & Costea, 2021) and where stress can influence relapse in dopamine system's response to drugs (Dayan, 2009).

Traditional epidemiological methods for data collection such as follow-up surveys used to identify onset of substances abuse, are limited by resources, non-responses and social desires (Mak, Lee, & Park, 2019). EEG data is also too expensive to capture. Exposing real test subjects to hazardous substances to compare findings is not practical for safety reasons. Instead simulated reward score data will be used based on observations of real world lever-pulling experiments conducted with pigeons and rats. Actions (levers) in this simulation will be limited to between six and ten.

2.1 Approaches to solve the problem

(Dayan, 2009) identified that some people suffer more from addiction than others and both vulnerability to addiction and relapse were not well understood. Lab rats with low levels of dopamine showed more sensitivity to compulsive cocaine abuse, abstinence then relapse. Relapse occurred when re-exposed to cocaine, this was attributed to increased stress during the relapse period. (Dayan, 2009) used reinforcement learning to recreate this behaviour to better understand how dopamine influenced drug initiation and compulsion. However due to the complex unlearning process happening during abstinence, RL did not demonstrate relapse well.

(Popescu et al., 2021) stated that addiction was classified as a disease since the late 1800s due to its negative impact on individuals and society. Since then, Neuroscientists have used genetics and reward neurology to explain addiction, while Behavioural Scientists have used behavioural models but both have not agreed on its cause. However, both agreed hormones affected memory which affected the level of addiction. (Popescu et al., 2021) further noted that several models were rendered obsolete when the American Psychiatric Association (APA) which created the Diagnostic and Statistical Manual of Mental Disorders (DSM) updated Gambling from an impulse control disorder to an addictive disorder. 'Substance-Related Disorders' was changed to 'Substance-Related and Addictive Disorders'. More types of substance addictions were added notably cannabis, opioids and stimulants with the International Statistical Classification of Diseases and Related Health Problems (ICD) mirroring these changes.

(Mak et al., 2019) surveyed addiction literature noting two recent studies that used Deep Q-Learning to model addiction. The first assessed the relationships between cigarette smoking and a reward signal among twenty-five university students who were moderate smokers. Softmax action selection was used to control the relative levels of exploration and exploitation. They correlated positive reward rate differences between cigarette consumption with carbon monoxide. Negative reward was significantly enhanced for smoking abstinence when compared with cigarette consumption. These suggested that smoking states (abstinence and cigarette consumption) were related to positive probabilistic selection task reward signals, RL, and decision making. The second conducted a randomised cross-over study that included twenty-two USA adult cocaine addicts who were non-treatment seekers. Cocaine levels were assessed by urine test. Participants performed probabilistic loss-learning tasks during MRI scans. Results suggested that cocaine dependent participants showed higher positive probabilistic reward rates during deprivation, which may attribute to their inability to control the dopamine hormone.

(Keramati et al., 2017) built a homeostatic regulation simulator based on homeostatic reinforcement learning, a current addiction theory which frames addiction as a homeostatic disorder where chronic drug use induces long-lasting maladaptation in the brain to maintain a homeostatic setpoint. In this view drug seeking is not habitual, but goal-directed aimed at fulfilling the intensely escalated need for the drug. They trained a Deep Q-Learning agent inside the simulated environment which resembled observational data from real rat cocaine experiments. Their study suggested that drug addiction hijacks the brain's goal directed associative learning system, which defends the physiological stability of an organism. Their model also implied that in experimental animals and people, a prior deviation exists. They wondered what factors caused the majority of experimental animals to have pre-existing homeostatic deviations that rendered them sensitive to the rewarding effects of cocaine. They suggested it was poor environmental stimulation of the laboratory animal's brain reward system. They also suggested that at least in the majority of individual, increasing and/or diversifying access to non-drug options should reduce cocaine use and risk of escalation.

Regarding how options are selected from memory in Deep Q-Learning (Hayes et al., 2021) stated, a

method used to sample and store data from the environment called Experience Replay was inspired by observing biological neural networks during sleep, primarily in the hippocampus. Experience Replay in humans is now thought to play a critical role in memory formation, retrieval, and consolidation. For example, Prioritisation in Experience Replay was shown to be motivated by fear due to old experiences overlapping with new memories, as they were most in danger of being damaged by new learning and were preferentially replayed.

(Zhang & Sutton, 2017) stated Experience Replay was first introduced by Lin in 1992 to train an RL agent with the transitions sampled from a buffer of previously experienced transitions. Now widely used in Deep Q-Learning it is still poorly understood and the buffer size hyperparameter is underestimated in the community. Most set it to the default one million transition value that (Mnih et al., 2016) used. (Zhang & Sutton, 2017) experimented with changing the hyperparameter and discovered that both small and large replay buffer sizes can heavily hurt the learning process. They claimed the effect of important transitions are delayed in Experience Replay but increasing the buffer size partially controlled this negative effect. (Fedus et al., 2020) held other components of the Deep Q-Learning architecture fixed and studied the effects of modifying Experience Replay. They discovered that the value approximator (neural network) improves with large replay capacity, so increasing the buffer size with a fixed replay ratio has varying improvements.

2.2 Gaps in Research

The cause of addiction is not well understood (Popescu et al., 2021) nor is the pre-existing homeostatic deviation in animals that makes them sensitive to cocaine. Although Deep Q-Learning can model homeostatic regulation (Keramati et al., 2017), no current model can explain all stages and symptoms of addiction (Mollick & Kober, 2020), such as relapse (Dayan, 2009). Although hormones such as dopamine and stress can affect memory (Mak et al., 2019; Dayan, 2009; Popescu et al., 2021; Hayes et al., 2021), it is not well understood what happens to memories after consolidation from the hippocampus to the neocortex (Hayes et al., 2021). Reinforcement Learning can help narrow these gaps if the next direction of research continues from (Zhang & Sutton, 2017; Fedus et al., 2020) and varies the buffer size hyperparameter, to see what happens in Experience Replay when used by a neural network of a Deep Q-Learning architecture to update its Q-value weights to maintain a homeostatic setpoint.

3 Research Question

To what extent does changing the size hyperparameter for Experience Replay in a deep Q-Learning architecture influence compulsive selection, despite negative reward scores received from the action, and the extent to which compulsive action selection introduces prediction errors in the neural network model responsible for long term reward score estimation, any of which lead to overvaluation of bad actions even with long term abstinence from bad action selection?

4 Hypothesis

Alternative Hypothesis (H1): If the buffer size hyperparameter is reduced for 'experience replay' of a Deep Q Learning Architecture, trained on the reward scores for pulling levers inside a simulated environment then when compared to the actual reward scores returned by the simulator for pulling a lever, a statistical significant difference (α: 0.05) is expected between the distribution of predicted reward scores for the reduced buffer size model, to one where the buffer size is held at a default 1 million transitions. Null Hypothesis (H0): there is no significance in reward scores.

5 Design and Implementation

5.1 Objective 1: Build a Simulated Environment to conduct experiments; 8 weeks

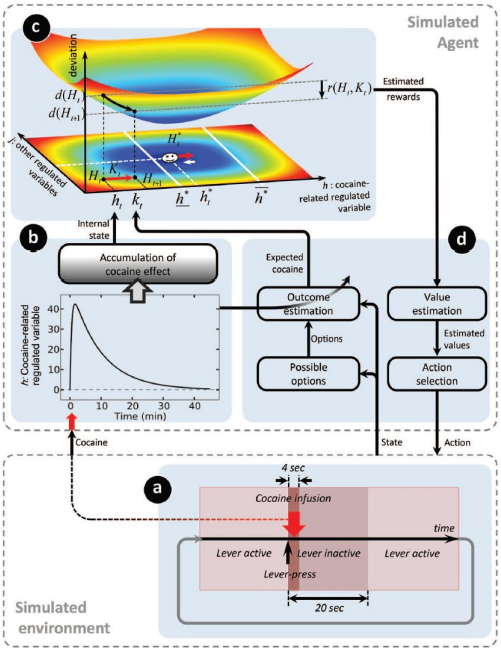
1. Create a Deep Q-Learning (DQL) agent; An ensemble comprising of a PyTorch neural network model and an array for the event replay memory model that are both used by the agent to sample data from the environment, perform the Bellman Equation then stochastic gradient descent with the Adam's optimiser to update the neural network via backpropagation for value approximation then perform action selection based on Softmax for exploration/exploitation behaviour. The DQL value estimator predicts dopamine reward score for performing an action. h is a brain internal variable (dopamine) that increases during drug use then decreases. Memory of actions that increase h accumulate over time in

the memory buffer and are sampled to train the model.

2. Create a lever pulling environment; comprising of a custom Openai-gym environment based on the CartPole environment (Kumar, 2020) to represent a homeostatic regulation mechanism. Each Lever-press initiates cocaine use for 4 sec. dopamine (h) is set high (e.g 45 units). After 4 seconds the lever is inactive for 20 sec. h reduces until zero after 45 secs unless another lever is pressed. $H*$ is a homeostatic setpoint and H is an internal state (an array that contains dopamine along with other hormone variables). The amount of reward (r) given for action (k) will be equal to the reduced distance from the internal state H to the homeostatic setpoint $H*$. Each bad action sets $H*$ to a new setpoint $H*_u$ unless no bad action is taken then $H*$ resets automatically back to its original setpoint ($H*_l$). In each state information about H and $H*$ are given to the agent to estimates how much r is needed to reach $H*$ given current H and selects an action that brings it closer to $H*$.

3. Create a simple user interface; use either a bash terminal, command prompt or Jupyter notebook to allow users to set parameters and observe the agent selecting actions in the environment.

Figure 1: *Homeostatic Regulation Simulator (Keramati et al., 2017)*



5.2 Objective 2: Conduct Experiments on Experience Replay hyperparameter to understand reward variance and test hypothesis; 4 weeks

1. Experiment 1: Set Experience Replay to default 1 million transitions and run simulation. Let agent learn, first by sampling randomly from environment to explore then exploiting actions to gain the most dopamine reward that maintains its homeostatic setpoint.

2. Experiment 2: Repeat Experiment 1 with, first 50% reduction in memory.

3. Experiment 3: Repeat Experiment 1 but dynamically change memory (50% smaller if further away from homeostatic setpoint and 50% larger if closer)

4. Evaluate DQL Agent; compare DQL agent actions and reward given memory size with an Analysis of Variance (ANOVA) to see which are statistically significant then post hoc test (e.g Tukey's method) to identify the top group.

5.3 Objective 3: Build a Data Visualisation & Explainability Portal for Educating the Public and Independent Assessment of hypothesis; 3 weeks

1. Host the environment and DRL agent on the web; Implementing a Django REST API with a PostgreSQL database and connect bash terminal or command prompt to it so others can recreate experiments to independently test hypothesis or test their own hypotheses.

2. Create a Web Portal for user access; Implement a Vue or React Javascript front end for users, for mobile and laptop devices to interact with the agent and environment, visualise the DQL agent in the environment both via live and historical logs. Finally allow datasets from logs to be downloaded for further analysis or comparison to real world data.

5.4 Objective 4: Research Writing to communicate results and future work; 5 weeks

1. Write three drafts and review; Introduction, Literature Review, design & methodology, results and discussion, conclusion and future work

2. Issue final; allow contingency for any requested revisions or repeated experiments.

6 Performance Metrics

Each experiment conducted will allow the Deep Q-Learning agent's reward scores to be divided into three different groups of behaviour, where each receive an initial memory buffer size for the experiment period. The control buffer size will be default 1 million transitions. The reward scores will be a list set containing state presented, reward predicted, action selected and actual reward received. At the end of the experiment, reward scores are measured. Then for each group, the mean reward scores are calculated. Analysis of variance (ANOVA) will be used to compare these group means to find out if they are statistically different or if they are similar. For performance the lowest memory size should be statistically significant, so a Post Hoc Test (e.g Tukey's method) will be performed after ANOVA to confirm this.

References

Bellman, R. (1957). *Dynamic Programming*. Dover Publications.

Dayan, P. (2009, 5). Dopamine, reinforcement learning, and addiction. *Pharmacopsychiatry*, *42 Suppl 1*, S56–65.

Fedus, W., Ramachandran, P., Agarwal, R., Bengio, Y., Larochelle, H., Rowland, M., & Dabney, W. (2020). Revisiting fundamentals of experience replay. JMLR.org.

Hayes, T. L., Krishnan, G. P., Bazhenov, M., Siegelmann, H. T., Sejnowski, T. J., & Kanan, C. (2021, 10). Replay in deep learning: Current approaches and missing biological elements. *Neural Computation*, *33*, 2908-2950. Retrieved from https://doi.org/10.1162/neco$_{a0}$1433 doi: 10.1162/neco$_{a0}$1433

Keramati, M., Durand, A., Girardeau, P., Gutkin, B., & Ahmed, S. H. (2017, 3). Cocaine addiction as a homeostatic reinforcement learning disorder. *Psychol. Rev.*, *124*, 130-153.

Kumar, S. (2020). Balancing a cartpole system with reinforcement learning - a tutorial. *ArXiv*, *abs/2006.04938*.

Li, Y. (2019). Reinforcement learning applications. *CoRR*, *abs/1908.06973*. Retrieved from http://arxiv.org/abs/1908.06973

Mak, K. K., Lee, K., & Park, C. (2019, 5). Applications of machine learning in addiction studies: A systematic review. *Psychiatry Res.*, *275*, 53-60.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016, 11). Asynchronous methods for deep reinforcement learning. In M. F. Balcan & K. Q. Weinberger (Eds.), (Vol. 48, p. 1928-1937). PMLR. Retrieved from https://proceedings.mlr.press/v48/mniha16.html

Mollick, J. A., & Kober, H. (2020, 8). Computational models of drug use and addiction: A review. *J. Abnorm. Psychol.*, *129*, 544-555.

Popescu, A., Marian, M., Miruna, R.-V. D. A., & Costea. (2021, 5). Understanding the genetics and neurobiological pathways behind addiction (review). *Exp. Ther. Med.*, *21*, 544.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). *Prioritized experience replay.* Retrieved from http://arxiv.org/abs/1511.05952 (cite arxiv:1511.05952Comment: Published at ICLR 2016)

Serenko, A., & Turel, O. (2020, 7). Directing technology addiction research in information systems: Part i. understanding behavioral addictions. *SIGMIS Database*, *51*, 81-96. Retrieved from https://doi.org/10.1145/3410977.3410982 doi: 10.1145/3410977.3410982

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9-44.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second ed.). The MIT Press. Retrieved from http://incompleteideas.net/book/the-book-2nd.ht

Tokic, M. (2010). Adaptive ϵ-greedy exploration in reinforcement learning based on value differences..

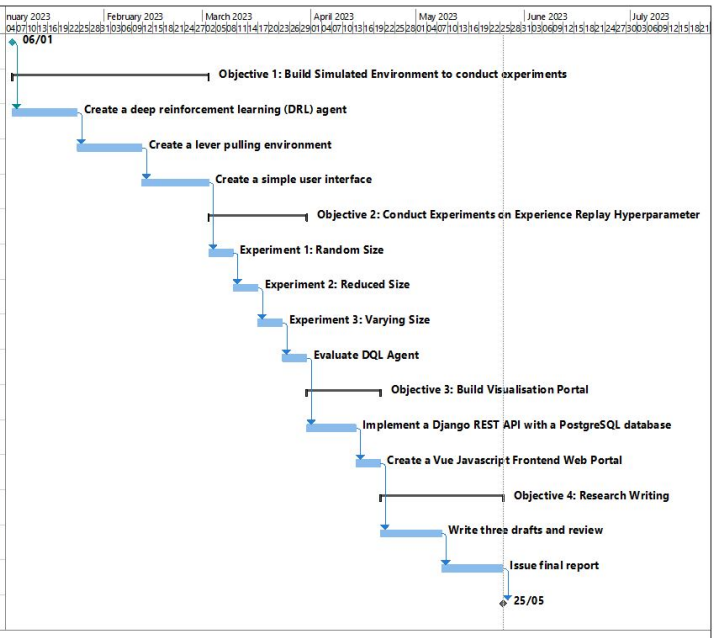
van Otterlo, M., & Wiering, M. (2012). Reinforcement learning and markov decision processes. In M. Wiering & M. van Otterlo (Eds.), *Reinforcement learning: State-of-the-art* (pp. 3–42). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-27645-3$_1$ doi: 10.1007/978-3-642-27645-3$_1$

White, D. J. (1993). A survey of applications of markov decision processes. *Journal of the Operational Research Society*, *44*, 1073-1096. Retrieved from

https://doi.org/10.1057/jors.1993.181
doi: 10.1057/jors.1993.181
Zhang, S., & Sutton, R. (2017, 12). A deeper look at experience replay.

7 Activities

ID	Task Name	Duration	Start	Finish
1	Project Start	0 wks	Fri 06/01/23	Fri 06/01/23
2	**Objective 1: Build Simulated Environment to conduct experiments**	**8 wks**	**Fri 06/01/23**	**Thu 02/03/23**
3	Create a deep reinforcement learning (DRL) agent	2.6 ewks	Fri 06/01/23	Tue 24/01/23
4	Create a lever pulling environment	2.67 ewks	Tue 24/01/23	Sun 12/02/23
5	Create a simple user interface	2.67 ewks	Sun 12/02/23	Thu 02/03/23
6	**Objective 2: Conduct Experiments on Experience Replay Hyperparameter**	**4 wks**	**Thu 02/03/23**	**Thu 30/03/23**
7	Experiment 1: Random Size	1 ewk	Thu 02/03/23	Thu 09/03/23
8	Experiment 2: Reduced Size	1 ewk	Thu 09/03/23	Thu 16/03/23
9	Experiment 3: Varying Size	1 ewk	Thu 16/03/23	Thu 23/03/23
10	Evaluate DQL Agent	1 ewk	Thu 23/03/23	Thu 30/03/23
11	**Objective 3: Build Visualisation Portal**	**3 wks**	**Thu 30/03/23**	**Thu 20/04/23**
12	Implement a Django REST API with a PostgreSQL database	2 ewks	Thu 30/03/23	Thu 13/04/23
13	Create a Vue Javascript Frontend Web Portal	1 ewk	Thu 13/04/23	Thu 20/04/23
14	**Objective 4: Research Writing**	**5 wks**	**Thu 20/04/23**	**Thu 25/05/23**
15	Write three drafts and review	2.5 ewks	Thu 20/04/23	Mon 08/05/23
16	Issue final report	2.5 ewks	Mon 08/05/23	Thu 25/05/23
17	Project Finish	0 wks	Thu 25/05/23	Thu 25/05/23



Project: Simulating Addiction
Date: Sun 18/12/22

Task	Project Summary	Manual Task	Start-only	Deadline	
Split	Inactive Task	Duration-only	Finish-only	Progress	
Milestone	Inactive Milestone	Manual Summary Rollup	External Tasks	Manual Progress	
Summary	Inactive Summary	Manual Summary	External Milestone		

Page 1