INTERNATIONAL BRAIN IBRO RESEARCH ORGANIZATION

IBRO/IBE-UNESCO Science of Learning Briefings

Learning and Reinforcement

Jean-Claude Dreher

Title	Learning and reinforcement
Series	IBRO/IBE-UNESCO Science of Learning Briefings
IBE Director	Dr. Mmantsetsa Marope
Author of brief	Dr Jean-Claude Dreher
Author affiliation	CNRS, Institute of Cognitive Science, Lyon, France
Date	22th October 2021

Dr Jean-Claude Dreher is research director of the lab 'Neuroeconomics, Reward and Decision Making' (https://dreherteam.wixsite.com/neuroeconomics) at the Institut des Sciences Cognitives Marc Jeannerod (UMR 5229, Lyon, France). He investigates the neural mechanisms underlying decision making, motivation and reward processing in humans, using concepts from cognitive neuroscience, psychology and behavioral economics. He uses experimental tools such as model-based functional Magnetic Resonance Imaging to understand the computational processes involved when making a choice.

Executive Summary

- In Neurosciences and Psychology, learning, reinforcement and rewards are closely related concepts but refer to distinct notions.
- Rewards and punishments are any external objects (or stimuli) that an organism would make an effort to obtain or avoid, respectively.
- Reinforcement Learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment to maximize cumulative rewards in the future.
- There are multiple brain systems to detect one's errors, to correct them and for checking.
- The brain constantly makes predictions regarding the outcome of its choices.
- These brain systems are key for learning by reinforcement and for learning by trial and error.
- Dopaminergic neurons located in the midbrain send a learning signal to many parts of the brain, which plays a pivotal role in encoding prediction error signals (i.e. differences between expected reward and the reward effectively delivered).
- Brain systems for learning via rewards and learning from punishments partially overlap but are distinct.
- A variety of reinforcement schedules can be used fruitfully in education and serious games, such as variable reinforcement schedules.
- These principles are useful for education in general and, not just at school.

Introduction

A key discovery was made in 1953 by Olds and Milner, called the Brain stimulation Reward (BSR), defined as a pleasurable phenomenon elicited by direct stimulation of specific brain regions. BSR allows animals to learn to execute novel behaviors, such as lever pressing in order to receive short electrical stimulation in specific brain areas (**Figure 1**). This discovery was a turning point in the study of the brain's reward system because BSR is able to act as a reward in the absence of any peripheral sensory stimulation and any physiological need. The existence of such a "pure reward" signal provided a proof that brain regions are specialized in hedonic functions as well as in motivational aspects of the reward (Sugrue et al., 2005; Wise, 2002). Stimulation activates the reward system circuitry and establishes response habits similar to those established by natural rewards, such as food and sex. Thus, electrical brain stimulation, or intracranial drug injections, can produce robust reward sensations as well as motivation to come back for more due to direct activation of the reward circuit. This discovery illustrates that rewards have several basic functions: (1) they induce subjective feelings of pleasure and contribute to positive emotion; (2) they can act as positive reinforcers by increasing the frequency and intensity of behavior that leads to the acquisition of goal objects, as described in classical and instrumental conditioning procedures; (3) they can also maintain learned behaviors by preventing their extinction (Schultz, 2000).

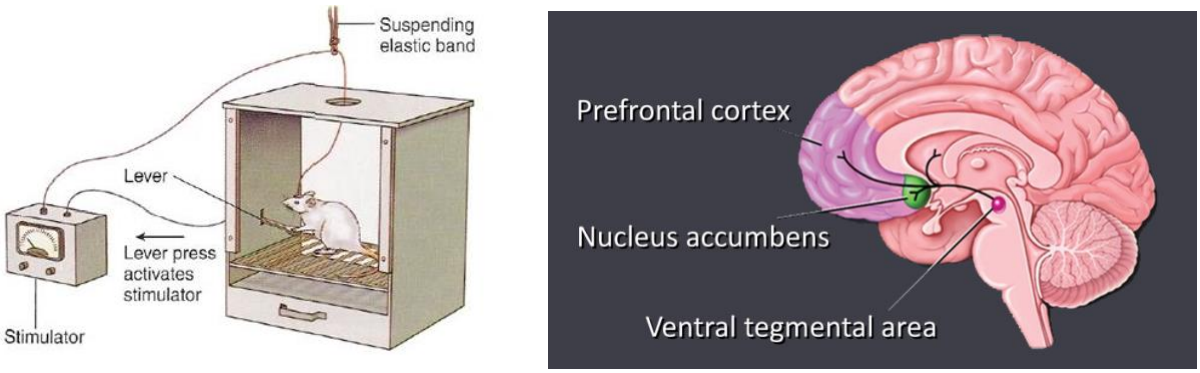

Figure 1 . **Left**. Positive reinforcement produced by self-electrical stimulation of the septal area and other regions of rat brain. **Right**. A learning signal is sent by dopaminergic neurons located in the midbrain to the striatum and prefrontal cortex to influence reward-dependent learning (Bayer et al., 2007; Schultz, 2000; Schultz & Dickinson, 2000).

Learning signal as the difference between predicted outcome and outcome effectively delivered

Rewards induce changes in observable behavior and serve as positive reinforcers by increasing the frequency of the behavior that results in reward. In Pavlovian, or classical, conditioning, the outcome follows the conditioned stimulus (CS) irrespective of any behavioral reaction, and repeated pairing of stimuli with outcomes leads to a representation of the outcome that is evoked by the stimulus and elicits the behavioral reaction. By contrast, instrumental conditioning requires the subject to execute a behavioral response. Instrumental conditioning increases the frequency of those behaviors that are followed by reward by reinforcing stimulus-response associations.

Neuroscientists have identified neurons, called dopaminergic neurons, which fire at the specific time of behavior when learning associations between stimuli and rewards during Pavlovian and instrumental conditioning paradigms. Dopaminergic cell bodies are located in small nuclei called the ventral tegmental area (VTA) and substantia nigra in the midbrain. These neurons send projections to the basal ganglia, (in particular the ventral striatum), which are engaged in motivation and learning, and, to the prefrontal cortex, engaged in cognitive control and executive functions (**Figure 1**). Dopamine is released at these projection sites when dopaminergic neurons fire, and this release modulates a large number of cognitive, motivational and learning functions.

One important function is to learn to associate a cue (stimulus) with a reward. At the time of expected reward delivery, dopaminergic neurons send a learning signal to efferent brain structures which reflects a difference between what is expected and what is effectively delivered (actual outcome – predicted reward). Prediction errors thus measure deviations from previous reward expectations. A prediction error can either be positive (when the reward delivered is better than expected) or negative (less or no reward delivered at the expected time) (Schultz et al., 1997; Sutton & Barto, 2018). Prediction errors are used to learn the value of states of the world and are critical for learning how to make better choices in the future. Electrophysiological studies in monkeys indicate that dopaminergic neurons code this prediction error signal in a transient fashion at the time of the outcome.

Interestingly, in classical conditioning experiments, where an association has to be learnt between a conditioned stimulus and a rewarding outcome (unconditioned stimulus), dopaminergic neurons fire not only at the time of the outcome delivery but also at the time of the conditioned stimulus (**Figure 2**). The signal coded at the time of the conditioned stimulus reflects the

subjective value associated to the reward. Different factors such as reward magnitude, probability, timing uncertainty and delay, influence this subjective value signal as well as the prediction error signal. In particular, the phasic response of dopaminergic neurons to the conditioned stimuli monotonically increases with probability and magnitude, but decreases with increasing reward probability at the time of the outcome, as predicted from a prediction error signal (Fiorillo et al., 2003; Kobayashi & Schultz, 2008). Taken together, these results indicate that dopamine responses reflect at least 2 important signals: the subjective value of the cue which predicts the future rewarded outcome, and the prediction error signal at the time of outcome (**Figure 2**). Finally, in addition to the roles of dopamine in encoding these two signals, electrophysiological studies also indicate that dopaminergic neurons may code a sustained signal between the cue and the outcome, that reflects reward uncertainty (maximal when reward probability=0.5) (Fiorillo et al., 2003). This signal may be functionally important for risk seeking behavior and/or exploratory behavior. This signal may explain why gambling, with its intrinsic reward uncertainty characteristics, has reinforcing properties that share common mechanisms with addictive drugs (Fiorillo et al., 2003).

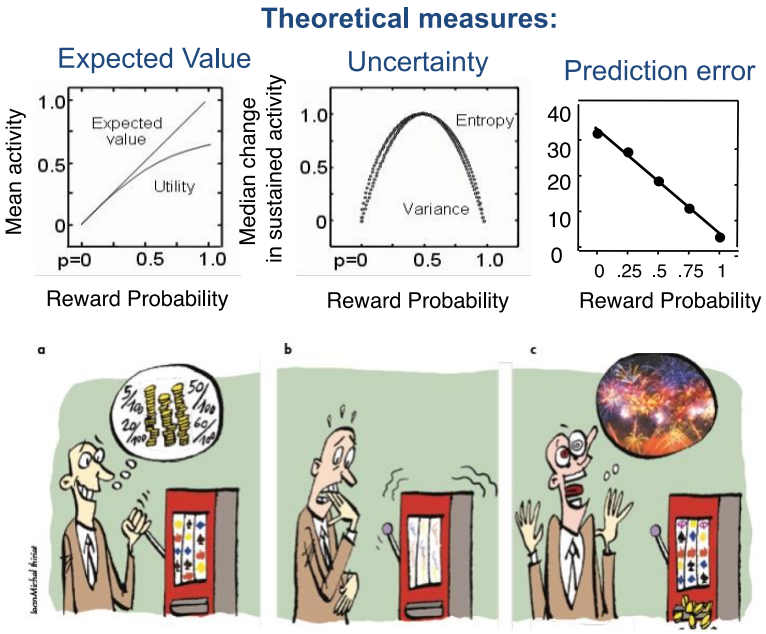

Figure 2. Dopaminergic neurons have been proposed to code 3 types of theoretical measures. The signal coded at the time of the conditioned stimulus reflects the subjective value associated to the reward. The phasic response of dopaminergic neurons to the conditioned stimuli monotonically increases with probability. During the delay between the cue and the outcome, dopaminergic neurons also encode a sustained signal that reflects reward uncertainty (maximal when reward probability=0.5). Finally, at the time of reward outcome, dopaminergic neurons encode a prediction error signal reflecting the difference between expected reward and reward effectively delivered. This signal decreases with increasing reward probability at the time of the outcome.

Human neuroimaging studies on reinforcement learning

One key property of dopaminergic neurons is that the Prediction Error signal can be formally described by an algorithm developed in the field of reinforcement learning (Schultz et al., 1997). In the Reinforcement learning framework, an agent learns a policy that maximizes the "reward function" that accumulates from the immediate rewards (**Figure 3**). A basic reinforcement learning agent interacts with its environment in discrete time steps. At each time t, the agent receives the current state s(t), and reward r(t). It then chooses an action a(t) from the set of

available actions, which is subsequently sent to the environment. The environment moves to a new state s(t+1) and the reward r(t+1) associated with the transition between s(t), a(t), s(t+1) is determined. The goal of a reinforcement learning agent is to learn a policy which maximizes the expected cumulative reward.

The beauty of this framework is that cognitive neuroscience has now combined RL models with functional neuroimaging in the so-called model-based fMRI approach. In this approach the signal observed in fMRI experiments is regressed with the computational signals coming from the RL framework, such as the signals encoding the value of each possible actions (called action value) and the prediction error signal. Model-based fMRI has been used to investigate the neural correlates of the prediction error signal in the whole brain. A number of studies suggest that activity in the ventral striatum and ventromedial prefrontal cortex correlates with the reward prediction error (Abler et al., 2006; Berns et al., 2001; Bray & O'Doherty, 2007; Dreher et al., 2006; Fletcher et al., 2001; McClure et al., 2003; O'Doherty et al., 2003). Moreover, the magnitude of the prediction error signal in the striatum has been found to correlate with behavioral performance (Pessiglione et al., 2006; Schönberg et al., 2007).

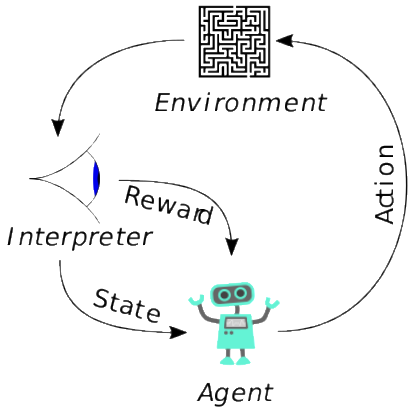

Figure 3. In the **r**einforcement learning (RL) framework, an agent learns a policy which maximizes the expected cumulative reward. At each time step, the agent is in a current state with a reward value function and it chooses an action a from a set of available actions, which is subsequently sent to the environment.

Multiple brain systems for learning by trial and error

Learning by trial and error has important implications for education and learning in general, outside of the classroom. This learning process engages different brain systems in addition to midbrain dopaminergic neurons that encode the difference between what is predicted and what is effectively obtained after one's actions. First, when we make an error in a cognitive task, a signal that reflects internal errors, (called Error Related Negativity, ERN), is generated rapidly, (in less than 80 ms), by a region called the supplementary motor area (**Figure 4**). This results in an automatic slowing down of our motor action, (post-error slowing), at the behavioral level. However, the ERN signal does not indicate what correction needs to be made, and does not in itself provide the solution. Another cognitive control mechanism subsequently engages a third brain region, (the anterior cingulate cortex), which encodes a signal known as the feedback-related negativity (FRN), an Event Related Potential (ERP) component reflecting error monitoring after the feedback has been obtained from the world. This signal is generated around 250 ms after we receive external information, (e.g. a teacher's comment), which has a discrepancy with our expectations. The anterior cingulate cortex is connected to other brain areas engaged in decision making, and can thus set in motion cognitive flexibility and strategic thinking that takes into account the mistake. To date, it seems that the signal encoded in midbrain dopaminergic neurons (prediction error signal) is sent to the anterior cingulate cortex, in which populations of neurons detect feedback and propose explorative strategies for adaptive behavior. Thus, errors are key information for the brain to learn and correct previous predictions to allow adaptive learning.

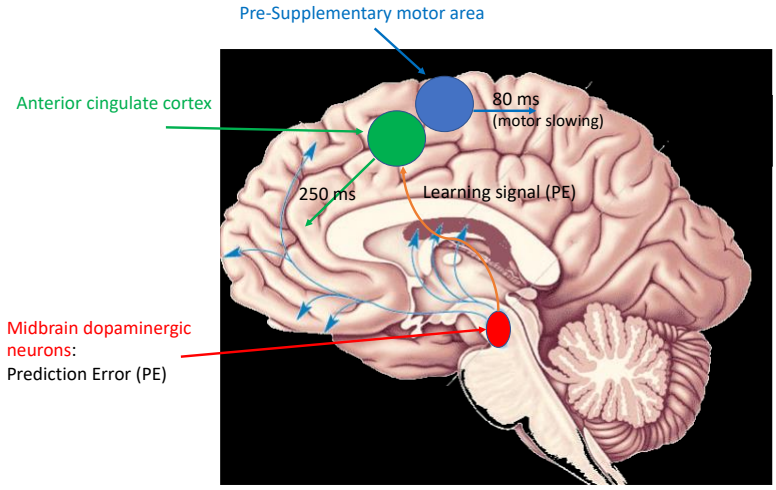

Figure 4. Different brain regions and systems are engaged when learning by trial and error. A learning signal (Prediction Error: PE) is sent by midbrain dopaminergic neurons to the frontal cortex and basal ganglia. This signal encodes a difference between what is predicted and what is effectively obtained after one's action. After error, a signal detecting internal errors (ERN) is sent by the pre-Supplementary Motor Area in less than 80 ms to motor regions to delay subsequent motor responses. Another cognitive control mechanism subsequently engages the anterior cingulate cortex (ACC), which encodes the feedback-related negativity (FRN), about 250 ms after reception of the feedback.

By "error" we often mean failure or a negative result, which can be subjectively perceived as a punishment. However, what is covered in this section is that the brain systematically recognizes whether it is wrong, whether for bad or for good: what the brain detects is deviations from its expectations. These are prediction errors, and they can be either better or worse than predicted. Concretely, how does this fundamental knowledge help to understand what happens when a student learns from trial and error? The computation of the difference between the reward

expected by the student (e.g. expecting a correct answer to a question in the classroom) and the one actually received, (the answer was wrong), allows her to constantly adjust her representations of the world according to this signal. Such trial and error learning involves feedback from the teacher or environment with respect to the student's performance. This feedback is not necessarily a sanction from the teacher, but may simply consist of the correct answer. If it is done regularly, targeted and finely adjusted, the student will rapidly learn from her mistakes. In addition to feedback from the teacher, there are also many softwares and electronic games which today allow students to monitor their errors and to self-regulate. These educational games can be adapted to the rhythm of each pupil in a class.

Decision bias must be overcome for effective learning

Why do we persist in making errors, even after receiving appropriate feedback? One reason is that we have a natural tendency to overestimate the value of an option previously selected. This decision bias is known as “choice-induced preference change” (Brehm, 1956), which inhibits learning from one’s mistakes. This phenomenon is based on the fact that one’s choices influence one’s values, such that actions or items seem to acquire value simply because one has chosen them. If a child chooses the wrong item between two items having equal value, they may persist in that choice. This type of persistent error may be difficult to overcome because decision biases are deeply rooted in evolution. One interesting proposal to overcome such errors is to favor teaching methods based on the observation of mistakes made by others (Monfardini et al., 2017). Indeed, when people are free from the preferences created by their own choices, they may learn very efficiently from the mistakes of others. A number of fMRI studies in humans, as well as direct single-cell recordings in non-human primates, indicate that learning by observation of others’ action or outcomes may engage partially different brain systems than learning from one’s own mistakes (Joiner et al., 2017). These basic principles can be translated inside the classroom into steps taken to promote correction procedures that focus on the correction of others’ mistakes. In this way, distinct brain systems can be recruited for learning by observation of others. These may be less prone to the decision bias when making choices for oneself.

Learning by carrots or sticks: appetitive versus aversive systems

Neuroimaging studies in humans and electrophysiological investigations in animals have not only focused on instrumental learning paradigms that require learning by trial and error to select the most rewarding option. A number of studies have also investigated the other side of the coin: that is learning to avoid options that are more punishing than the others. In this latter case, punishment has often been operationalized as a monetary loss or physical harm (pain such as heat to the skin inside the scanner). These two types of instrumental learning paradigms have helped to identify that in addition to the appetitive system described above, an aversive system also functions as an opponent system. While learning from stimuli-reward associations preferentially engages the striatum, the aversive system often engages the bilateral insula. In addition, some regions, such as the ventromedial prefrontal cortex (vmPFC) and amygdala show some degree of overlap, and are engaged regardless of valence type. In particular, the vmPFC may integrate information about both appetitive and aversive values across different reinforcer modalities (from rewards such as food to more abstract ones such as social reinforcers), so as to compute a net value that might guide decision making in a cost-benefit type of calculation.

fMRI studies also provide examples of overlapping appetitive and aversive processes in the human brain. As noted above, the striatum is not only engaged in appetitive learning and reward processes but is also engaged in aversive learning (Delgado et al., 2011), perhaps reflecting a role in avoidance learning (Delgado et al., 2009; Palminteri et al., 2012). There may also be some functional segregation within the striatum, with more anterior regions showing relative selectivity for rewards and more posterior regions for losses (Seymour et al., 2007).

Consistent with a role of the striatum in reinforcement learning, we have observed both appetitive and aversive prediction error signals in this region, consistent with a salience prediction error. In this fMRI study, we used RL modelling during a classical conditioning learning paradigm to investigate the prediction error related to different types of reinforcement (juice *versus* image), and also compared prediction error for rewards and punishments (Metereau & Dreher, 2013) (apple juice, salty water, money and aversive picture). Trials consisted of two phases: an anticipatory period followed by the outcome presentation. The results showed that the putamen, the insula and the anterior cingulate cortex code the taste prediction error, regardless of valence, i.e. for both the appetitive and the aversive liquids (juice and salty water). A different pattern of activation was observed in the amygdala, which coded a prediction error only for the primary/immediate reinforcers (apple juice, salty water and aversive pictures). These results demonstrate the different contributions made by distinct brain regions to compute prediction error depending upon the type and valence of the reinforcement **(Figure 5)**.

Taken together, neuroimaging evidence suggests that appetitive and aversive processing overlap in regions such as the striatum, but also show some functional segregation depending on contexts. These contexts are important because engagement of specific brain systems in appetitive/aversive conditioning depends on the specific contingencies of the associations to be learned. For example, in a simple Pavlovian conditioning paradigms in which one cue predicts either a reward or no reward delivery, and another cue predicts a punishment or no punishment delivery, the same outcome, i.e. absence of reward may be perceived as a punishment in the first case whereas the absence of punishment may be perceived as a reward in the second case.

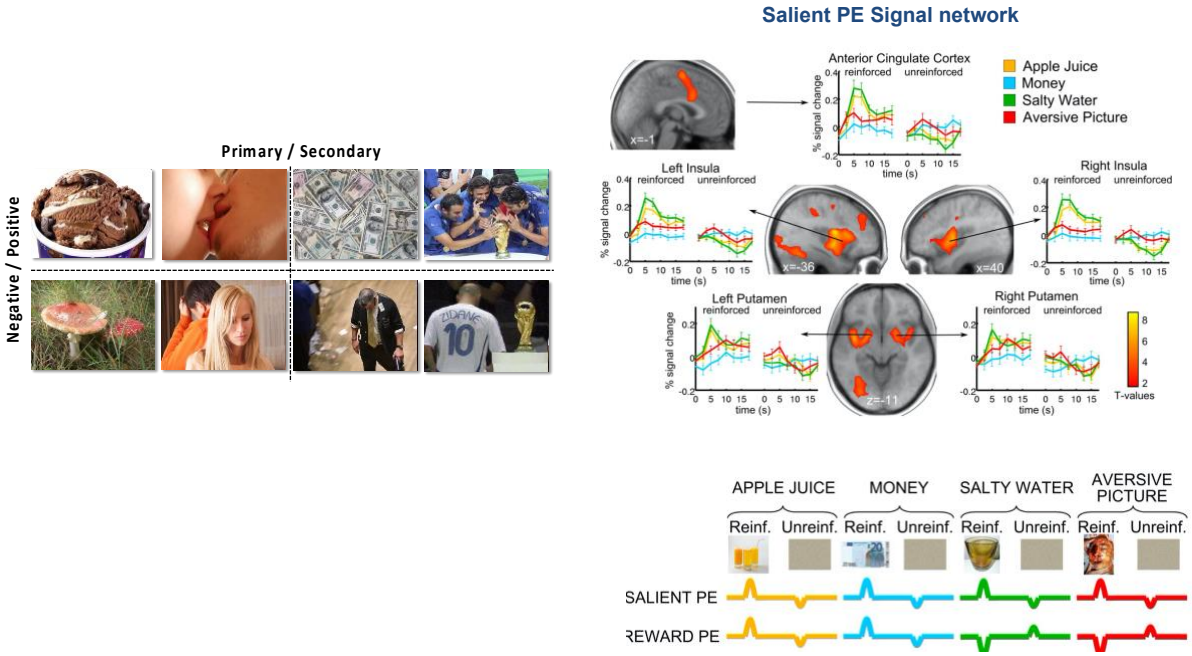

Figure 5. Left. Reinforcers can be positive (rewards) or negative (punishments) and can be classified as primary (innate value) vs secondary (learned from experience). Right. Salient Prediction Error (SPE) signal. Statistical parametric maps showing that activity in the anterior cingulate cortex, bilateral putamen, and bilateral insula correlates with the SPE in the gustatory conditions.. Reinforced and unreinforced trials are plotted separately. Bottom right. Illustration of computational signals expected for the salient PE and the Reward PE. The SPE signal responds to reward and punishment in the same way, as motivationally salient events, generating positive PE for reinforced trials and negative PE for unreinforced trials. The RPE signal responds to rewards and punishments in opposite ways, generating a positive PE when an unexpected reward is delivered or when an expected punition is missed and generating a negative PE when an

unexpected punishment is delivered or an expected reward is missed (Unreinf., Unreinforced; Reinf., Reinforced).

Variable vs fixed Reinforcement schedules: how can they help in learning and education?

Schedules of reinforcement are the rules that control the timing and frequency of reinforcer delivery to increase the likelihood a target behavior will happen again, strengthen or continue. In a schedule of reinforcement, the reinforcers are only applied when the target behavior has occurred, and therefore, the reinforcement is contingent on the desired behavior. There are two main categories of schedules: intermittent and non-intermittent. Non-intermittent schedules apply reinforcement or no reinforcement at all, after each correct response. Intermittent schedules apply reinforcers after some, but not all, correct responses. Among intermittent schedules, different schedules can be distinguished and one is of particular interest: the variable interval schedule. Variable interval schedules deliver the reinforcer after a variable time interval has passed since the previous reinforcement. This schedule usually generates a steady rate of performance due to the uncertainty of the time of the next reward, and is thought to be habit-forming.

One application for education is the schedule of exams.Variable interval schedules are more effective than fixed interval schedules of reinforcement in teaching and reinforcing behavior that needs to be performed at a steady rate. Students whose grades depend on the performance of unpredictable exams throughout the semester study more regularly as they cannot know in advance when exams will occur. In contrast, fixed interval schedules deliver a reward when a set amount of time has elapsed. This is the case when the date of the final exam is fixed. Many students whose grades depend entirely on a final exam do not study much at the beginning of the semester, but increase their work only when the exam date approaches.

Another concrete example for education is the used of multiple schedules of reinforcement as a behavioral intervention strategy to allow teachers to signal to students specific contexts under which behaviors will be reinforced or not. To signal when a teacher's attention is available (i.e., reinforcement) versus when it is not (i.e., extinction), different types of cues (e.g. color cards) can be used (Vargo, 2020). This has proven useful to reduce disruptive behaviors such as interruptions by students, which often disrupt the academic environment (Akers et al., 2019; Vladescu & Kodak, 2016). These papers offer important examples for teachers to know how to design and implement effective behavior management strategies to decrease problem behaviors and simultaneously increase academic-related behaviors.

A final example comes from the field of serious games which frequently use reinforcement schedules (Mayer, 2019; Nagle et al., 2014). Some studies have directly compared distinct methods of scheduling rewards in games, such as fixed ratio schedules, in which rewards are given after a fixed number of correct responses, and variable ratio schedules, in which they are given after an unpredictable number of correct responses. The results of one study indicate that giving rewards according to a player-centered variable-ratio schedule has the potential to make serious games more effective. The results also showed that enjoyment, performance, duration of gameplay and likelihood to play again were significantly higher when variable-ratio schedules were used.

Together, these behavioral findings demonstrate that variable reinforcement schedules, which are more likely to generate Prediction Error signals from dopaminergic neurons, are usually more effective in promoting study and learning.

Conclusions

Fundamental neuroscience knowledge derived from Reinforcement Learning and characterization of dopaminergic signals, (e.g. Prediction error) has important practical implications for learning and education in general. First, learning by trial and error requires learners to confront their mistakes. As one often tends to experience mistakes, learners need to be reassured. Students need to understand that mistakes should not be perceived as punishments and that they actually learn from trying something and receiving negative feedback (i.e. that they can be incorrect). Making a mistake allows us the opportunity to improve our skills over time. Another aspect concerns the rich field of reinforcement learning schedules, which has direct practical implications, such as the fact that frequent exams/tests will reduce the stress that results from the exam itself. Second, over the course of repeated tests, via trial and error and informative feedback, students will consider the test itself and their mistakes as a less stressful, and more importantly, as a necessity for successful learning. The benefits of discovery by oneself through the trial and error process reinforces the relationships between the correct actions and outcomes, thereby serving the learning purpose. Further aspects include personality traits such as confidence and metacognitive abilities to judge the relationship between perception of oneself and one's abilities.

References

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage*, *31*, 790–795.

Akers, J. S., Retzlaff, B. J., Fisher, W. W., Greer, B. D., Kaminski, A. J., & DeSouza, A. A. (2019). An Evaluation of Conditional Manding Using a Four-Component Multiple Schedule. *The Analysis of Verbal Behavior*, *35*(1), 94–102. https://doi.org/10.1007/s40616-018-0099-9

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *J Neurophysiol*, *98*, 1428–1439.

Berns, G. S., Mc Clure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *J Neurosci*, *21*, 2793–2798.

Bray, S., & O'Doherty, J. (2007). Neural coding of reward-prediction error signals during classical conditioning with attractive faces. *J Neurophysiol*, *97*, 3036–3045.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal Psychology*, *52*(3), 384–389. https://doi.org/10.1037/h0041006

Delgado, M., Jou, R., LeDoux, J., & Phelps, L. (2009). Avoiding negative outcomes: Tracking the mechanisms of avoidance learning in humans during fear conditioning. *Frontiers in Behavioral Neuroscience*, *3*, 33. https://doi.org/10.3389/neuro.08.033.2009

Delgado, M., Jou, R., & Phelps, E. (2011). Neural Systems Underlying Aversive Conditioning in Humans with Primary and Secondary Reinforcers. *Frontiers in Neuroscience*, *5*, 71. https://doi.org/10.3389/fnins.2011.00071

Dreher, J. C., Kohn, P., & Berman, K. F. (2006). Neural coding of distinct statistical properties of reward information in humans. *Cereb Cortex*, *16*, 561–573.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898–1902.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., Papadakis, N., & Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nat Neurosci*, *4*, 1043–1048.

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 8. https://doi.org/10.1038/s41539-017-0009-2

Kobayashi, S., & Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *J Neurosci*, *28*, 7837–7846.

Mayer, R. E. (2019). Computer Games in Education. *Annual Review of Psychology*, *70*(1), 531–549. https://doi.org/10.1146/annurev-psych-010418-102744

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*, 339–346.

Metereau, E., & Dreher, J.-C. (2013). Cerebral correlates of salient prediction error for different rewards and punishments. *Cerebral Cortex (New York, N.Y.: 1991)*, *23*(2), 477–487. https://doi.org/10.1093/cercor/bhs037

Monfardini, E., Reynaud, A. J., Prado, J., & Meunier, M. (2017). Social modulation of cognition: Lessons from rhesus macaques relevant to education. *Neuroscience & Biobehavioral Reviews*, *82*, 45–57. https://doi.org/10.1016/j.neubiorev.2016.12.002

Nagle, A., Wolf, P., Riener, R., & Novak, D. (2014). The Use of Player-centered Positive Reinforcement to Schedule In-game Rewards Increases Enjoyment and Performance in a Serious Game. *International Journal of Serious Games*, *1*(4), Article 4. https://doi.org/10.17083/ijsg.v1i4.47

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*, 329–337.

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., & Pessiglione, M. (2012). Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. *Neuron*, *76*(5), 998–1009. https://doi.org/10.1016/j.neuron.2012.10.017

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making. *Journal of Neuroscience*, *27*(47), 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007

Schultz, W. (2000). Multiple reward signals in the brain. *Nat Rev Neurosci*, *1*, 199–207.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu Rev Neurosci*, *23*, 473–500.

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nat Rev Neurosci*, *8*, 300–311.

Skvortsova, V., Palminteri, S., & Pessiglione, M. (2014). Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates. *Journal of Neuroscience*, *34*(47), 15621–15630. https://doi.org/10.1523/JNEUROSCI.1350-14.2014

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nat Rev Neurosci*, *6*, 363–375.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*.

Vargo, K. K. (2020). A Teacher's Guide to Using a Multiple Schedule of Reinforcement in Educational Settings. *Intervention in School and Clinic*, *56*(1), 36–42. https://doi.org/10.1177/1053451220910745

Vladescu, J. C., & Kodak, T. (2016). The Effect of a Multiple-Schedule Arrangement on Mands of a Child with Autism. *Behavioral Interventions*, *31*(1), 3–11. https://doi.org/10.1002/bin.1422

Wise, R. A. (2002). Brain Reward Circuitry: Insights from Unsensed Incentives. *Neuron*, *36*(2), 229–240. https://doi.org/10.1016/S0896-6273(02)00965-0