Leszek Rutkowski · Rafał Scherer ·
Marcin Korytkowski · Witold Pedrycz ·
Ryszard Tadeusiewicz · Jacek M. Zurada (Eds.)

LNAI 14125

Artificial Intelligence and Soft Computing

22nd International Conference, ICAISC 2023
Zakopane, Poland, June 18–22, 2023
Proceedings, Part I

Springer

Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence 14125

Founding Editor

Jörg Siekmann

Series Editors

Reinforcement Learning with Brain-Inspired Modulation Improves Adaptation to Environmental Changes

Eric Chalmers[1(✉)] and Artur Luczak[2]

[1] Mount Royal University, Calgary, AB T3E 6K6, Canada
echalmers@mtroyal.ca
[2] University of Lethbridge, Lethbridge, AB T1K 3M4, Canada
luczak@uleth.ca

Abstract. Developments in reinforcement learning (RL) have allowed algorithms to achieve impressive performance in complex, but largely static problems. In contrast, biological learning seems to value efficient adaptation to a constantly changing world. Here we build on a recently proposed model of neuronal learning that suggests neurons predict their own future activity to optimize their energy balance. That work proposed a neuronal learning rule that uses presynaptic input to modulate prediction error. Here we argue that an analogous RL rule would use action probability to modulate reward prediction error. We show that this modulation makes the agent more sensitive to negative experiences, and more careful in forming preferences: features that facilitate adaptation to change. We embed the proposed rule in both tabular and deep-Q-network RL algorithms, and find that it outperforms conventional algorithms in simple but highly-dynamic tasks. It also exhibits a "paradox of choice" effect that has been observed in humans. The new rule may encapsulate a core principle of biological intelligence; an important component of human-like learning and adaptation - with both its benefits and trade-offs.

Keywords: Reinforcement Learning · adaptation · lifelong learning · brain-inspired computing

1 Introduction

"Most work in biological systems has focused on simple learning problems. . . where flexibility and ongoing learning are important, similar to real-world learning problems. In contrast, most work in artificial agents has focused on learning a single complex problem in a static environment." (Neftci and Averbeck) [22]

Real-world environments are constantly changing, and the ability to flexibly adapt to these changes is imperative. But current A.I. does not always demonstrate this ability to the same degree as animals. Here, building on a recent model

L. Rutkowski et al. (Eds.): ICAISC 2023, LNAI 14125, pp. 23–34, 2023.
https://doi.org/10.1007/978-3-031-42505-9_3

of neuronal learning [20], we propose a reinforcement learning rule that demonstrates more realistic flexibility - including both its benefits and its trade-offs. We test the new reinforcement learning rule in multi-armed bandit tasks and a task inspired by the Wisconsin Card Sorting Test - a psychological test used to assess patients' ability to adapt to changing reward structures. We demonstrate that the new rule improves performance in dynamic decision-making tasks with few to moderate numbers of choices (probably like the routine decision-making faced by animals day-to-day), and that this comes at the expense of performance when selecting between many choices - a paradox-of-choice effect that has been observed in humans. We also discuss some connections between the new rule and several other paradigms from across machine learning.

2 A New Reinforcement Learning Rule

2.1 Basic Building Blocks of Reinforcement Learning

A reinforcement learning agent must be able to estimate the value V of executing action a while in state s - though during the early stages of learning its estimates may not be very good. The agent must learn from each new experience in the environment; improving the efficacy of its value estimates for the future. Suppose at time t the agent is in state s_t, executes action a_t, and then finds itself in the new state s_{t+1} with reward r. The actual, experienced value of this event can be formulated as reward r_t plus the predicted value of being in the new state s_{t+1}:

$$V\left(s_t, a_t\right)_{actual} = r_t + \gamma V\left(s_{t+1}\right) \tag{1}$$

Here γ is a discount factor applied to expected future rewards ($\gamma\epsilon$ [0, 1]). The "temporal difference error" δ expresses the difference between actual and predicted values:

$$\delta_t = V_{actual} - V = r_t + \gamma V\left(s_{t+1}\right) - V\left(s_t, a_t\right) \tag{2}$$

The temporal difference error is a measure of the agent's surprise at the recent experience, and is a useful mechanism for learning. In the canonical Q-learning algorithm, for example, the agent maintains a table of value estimates that are updated proportional to δ, and according to a learning rate parameter α:

$$V\left(s_t, a_t\right) \leftarrow V\left(s_t, a_t\right) + \alpha\delta_t \tag{3}$$

The agent selects actions for execution according to a policy π. For the purpose of this paper, let us assume π is a softmax function that calculates the probability of selecting action a out of the set of actions A, based on current value estimates, and according to a temperature parameter τ:

$$\pi\left(s, a\right) = P\left(a_t = a | s_t = s\right) = \frac{e^{V(s,a)/\tau}}{\sum_{b\epsilon A} e^{V(s,b)/\tau}} \tag{4}$$

Thus the learning process consists of iteratively using value estimates to select actions, and using the observed results to improve the value estimates.

2.2 The New Rule

Building on the Contrastive Hebbian Learning rule [1,3] Scellier and Bengio proposed "Equilibrium Propagation" (EP) as a new, more biologically plausible model for learning in artificial neural networks [24]. EP envisions the network as a dynamical system that learns in two phases. First is the "free phase", in which an input is applied, and the network is allowed to equilibrate. In the second or "weakly clamped" phase, output neurons are soft-clamped or nudged toward a target value. Weights are then updated according to the rule:

$$\Delta W_{ij} \propto \left[u_i^c u_j^c - u_i^f u_j^f \right] \tag{5}$$

where i and j are the indices of neurons on either side of the weight/synapse (note that EP assumes symmetric connections between neurons), u^c is the neuron's squashed clamped-phase activation, and u^f is the neuron's squashed free-phase activation. Luczak, et al. [20] showed that free-phase activity can be well predicted based on past activity, and proposed the following alternative rule:

$$\Delta W_{ij} \propto u_i^c (u_j^c - \tilde{u_j^f}) \tag{6}$$

where the tilde indicates the neuron's prediction of its own free-phase equilibrium given the input. They showed that this rule can explain learning without requiring two distinct phases, as free-phase activity can be predicted in advance.

Importantly, the rule arises naturally as a result of a neuron acting to optimize its own energy balance, and hints at an explanation for consciousness [19], suggesting that it may encapsulate some principle of general intelligence. This motivates our current exploration of an analogous reinforcement learning rule. Examining this new rule, we see the update consists of the prediction error term $(u_j^c - \widetilde{u_j^f})$, modulated by the presynaptic activation u_i^c. Here we abstract the basic form of this rule to produce a rule applicable to reinforcement learning. The prediction error term is easy to place in a reinforcement learning context: it is analogous to the temporal difference error δ. But if we want to formulate a reinforcement learning rule corresponding to the neuronal one, we need a scaling or modulating factor analogous to the presynaptic activation. Since the presynaptic activation is the input to the neuron and the cause of its resulting activity, a natural analog could be $\pi(s_t, a_t)$; the input to the agent's environment and the cause of the resulting experience. We can then formulate a reinforcement learning rule as a modulation of δ by $\pi(s_t, a_t)$:

$$\Delta V(s_t, a_t) \propto \pi_t(s_t, a_t)\delta_t = \pi_t(s_t, a_t)\left[r + \gamma V(s_{t+1}) - V(s_t, a_t)\right] \tag{7}$$

The analogy between the neuronal learning rule and the new RL rule is illustrated in Fig. 1.

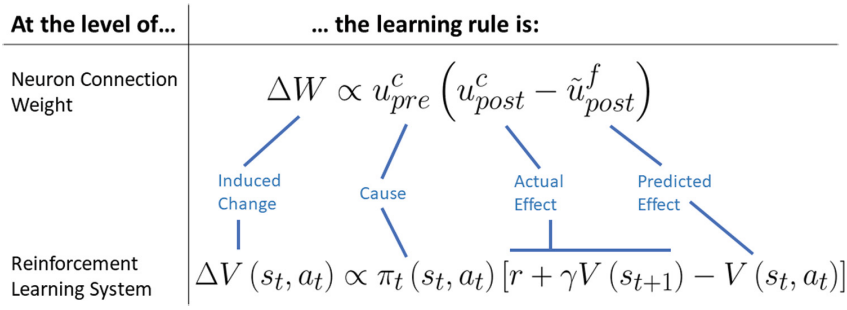

Fig. 1. We abstract the biologically-plausible neuronal learning rule of Luczak et al. to create an analogous rule for reinforcement learning. To calculate a weight update, the prediction error is modulated by the presynaptic input that caused the neuron's activity. In other words, a cause is used to modulate the error in predicting the effect. For an analogous reinforcement learning rule, updates are calculated by using action probability (the cause of the agent's experience) to modulate reward prediction error (the error in predicting the effect of that action)

2.3 Effect of the New Rule

Scaling the temporal difference error by $\pi(s_t, a_t)$ has two effects:

1. **It magnifies the agent's reactions to negative experiences.** If an action that was thought to be valuable (i.e. π is large) brings a negative out-come, the scaled (negative) reward-prediction error will be large. This will depress the perceived value of that action, creating an immediate aversion to it.
2. **It slows down the development of an action preference** - making the agent somewhat more careful in selecting actions. If the agent is unlikely to take an action, the scaled reward prediction error will be small - even if the experience was rewarding. Thus, the agent needs a lot of "convincing" that an un-likely action is desirable.

Thus, modulating the temporal difference error by the action probability $\pi(s_t, a_t)$ in this way biases the agent's learning somewhat toward negative experiences. We hypothesize this will allow the agent to adapt to environmental changes: when a previously rewarding action is no longer rewarded, the agent will quickly suppress its perceived value and carefully search for a new preference.

3 Experiments

3.1 Experiment 1 - Changing Multi-armed Bandit

A changing, n-armed bandit experiment was designed to test the new rule's ability to adapt to changes. Multi-armed bandits are a simple experiment often used to illustrate learning algorithms' performance [30]. The bandit was given one high-reward arm with $p_{reward} = 0.9$ and one no-reward arm with $p_{reward} = 0$. The rest of the arms had random reward probabilities $p_{reward} \sim U(0.25, 0.75)$. The agent receives a $+1$ reward when the arm it samples is rewarded, and a -1 reward otherwise. The reward probabilities are periodically rotated in such a way that all reward probabilities change, and the arm that was previously high-reward becomes no-reward.

The new rule was implemented in a tabular reinforcement learning agent by modifying the classical Q-learning algorithm to use Eq. 7 as its update rule. This algorithm maintains a table of the perceived values of each action and updates the relevant value after each experience. We also implemented a variety of standard bandit-solving algorithms for comparison: a conventional Q-learning algorithm [27] and an Upper Confidence Bound (UCB) algorithm [2]. All these algorithms are memoryless and so cannot learn the pattern to the reward probabilities' rotation: they perceive each change as a random, unexpected, and complete change to the reward landscape. For reference, we also included a UCB algorithm which has the advantage of being automatically reset each time the rewards change - note the other algorithms are not informed of changes this way; they must figure it out themselves. Thus, this "perfectly-informed" UCB algorithm represents a performance cap that the other algorithms are not expected to reach.

In our experiments each algorithm was allowed to select the best values for learning rate $\alpha \epsilon [0, 2]$ and softmax temperature $\tau \epsilon \{0.5, 1, 2\}$. Note that α is conventionally set to be (much) less than 1, but a large value of α can also produce a quick response to environmental changes, so here we allow each algorithm to select α as high as 2. The parameter searches and the experiments themselves were performed on different bandit instances. The cumulative reward for a 7-armed bandit with changes every 100 steps is shown in Fig. 2.

We note that the conventional Q learning algorithm achieved quick reaction to reward changes by self-selecting a large learning rate α (usually somewhere between 0.7 and 1.5). But this large α also causes the algorithm to switch to a new arm very quickly when it finds a chance reward at that arm - sometimes it switches too quickly and selects a sub-optimal arm, and cumulative reward suffers as a result. The new algorithm, on the other hand, scales reward-prediction-error down when the probability of selecting that arm is low, and so spends more time convincing itself that a new arm is desirable. This longer time spent identifying the new high-reward arm yields more reward overall, as shown in Fig. 2.

To quantify this extra time taken to develop a new arm preference, we first ran a 10-period moving average on the probabilities of selecting each arm. When

the maximum probability of any arm except the previously high-reward arm exceeded 50%, we considered the agent to have developed a new preference for that arm. Time-to-preference as the number of bandit arms increases is shown in Fig. 3.

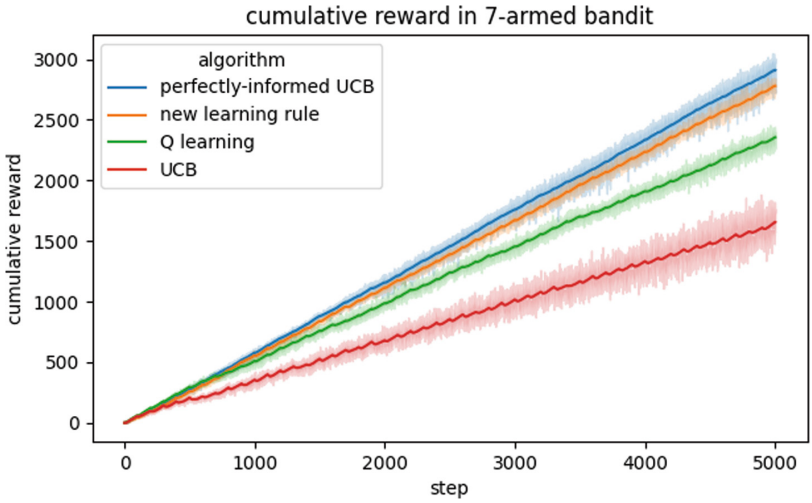

Fig. 2. Cumulative reward obtained on a 7-armed bandit with reward probabilities that change every 100 steps. The "perfectly-informed UCB" algorithm is reset (informed) when the reward probabilities change, and so represents a cap on possible performance. The new learning rule is *not* given this information, yet it performs almost as well. The shaded area is the 95% confidence interval of the mean over 10 repetitions.

3.2 Experiment 2 - Task Inspired by the Wisconsin Card Sorting Test

The Wisconsin Card Sorting Test is a neuropsychological test used to assess patients' ability to adapt to a changing set of rules [4], and has historically been used to identify brain injury and neurodegenerative disease [21]. The test presents patients with cards that can be matched based on several features, such as color, shape, number, etc. The patients are not told the correct matching criteria, but are rewarded when they make a match correctly. The rewarded matching criteria changes periodically throughout the test: healthy patients can generally adapt quickly when the rule changes, while patients with prefrontal cortex damage cannot.

Here we simulate a similar test using a multiclass classification task. Normally distributed clusters of points are created in n-dimensional space and assigned to

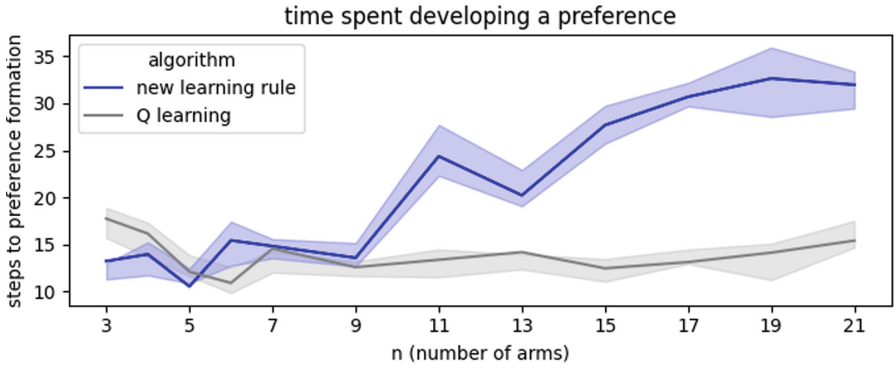

Fig. 3. Time (steps) taken to develop a preference for a new arm after each change, for varying numbers of bandit arms. See text for description of how "preferences" were detected. The new rule is more careful in evaluating options, and this helps it to identify the optimal arm when the number of choices is small. The shaded area represents the 95% confidence interval of the mean over 20 repetitions.

each of k classes. The agent is rewarded when it correctly matches a randomly-drawn point to its current class, but the classes are periodically scrambled (such that all the points previously assigned to class "0" now belong to class "2", for example).

For this test we use a deep Q network based on the new rule in Eq. 7. The network is a perceptron with one hidden layer of 20 neurons, and tanh squashing functions. We use separate policy and value networks that synchronize every 5 trials, and a replay buffer of the last 10 trials. The same network is also instantiated with a conventional update rule for comparison. Figure 4 shows the new rule allowing the network to adapt to each change, while the conventional deep Q network adapts less effectively. Here the classification rule is changed every 100 steps.

3.3 The Paradox of Choice

As humans we often take for granted our ability to change: to update our beliefs in response to new information, or to change a strategy when necessary. But our gift for quick adaptation in everyday situations comes with a trade-off: less-than-optimal performance in situations with many choices. Psychologist Barry Schwartz calls this "the paradox of choice" [25]: As the number of choices increases, our ability to select a satisfying option decreases and our preferences become weaker [9].

Our experiments show a similar paradox-of-choice effect, illustrated in Fig. 5. The new rule creates a bias toward negative experiences that - when the reward landscape changes - quickly depresses perceived value of the previously high-reward arm, and also makes the agent more careful in choosing a new preferred arm. This is an advantage when the number of arms is small, but can become

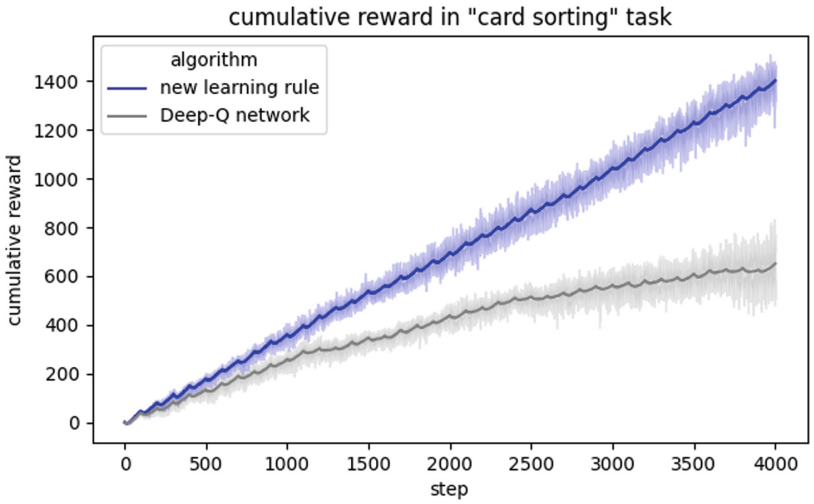

Fig. 4. Cumulative reward obtained in the 4-class version of the task inspired by the Wisconsin Card-Sorting Test, with classes being shuffled every 100 steps. The shaded area is the 95% confidence interval of the mean over 10 repetitions.

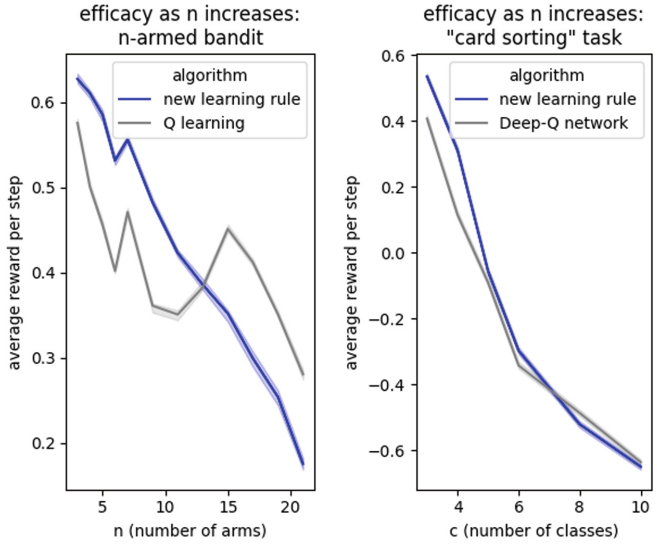

Fig. 5. Average reward-per-step obtained in the n-armed bandit task (left) and "card sorting" task (right). The new rule provides an advantage over conventional learning when the number of choices is small, with the trade-off of a disadvantage when the number of choices is large. This trade-off is likely favorable for many real-world situations, and similar to the "paradox of choice" effect observed in humans.

a disadvantage if there are many arms. In large-n cases, the agent takes too long evaluating new arms and sometimes fails to select one in time. Thus the same effects of the new rule that allow effective adaptation to environmental changes involving few choices, hinder it when the number of choices becomes large. This trade-off likely works heavily in favor of biological agents in the natural world, who rarely have to select between many attractive options, but must continuously adapt to simple though potentially dramatic environmental changes (e.g. the animal discovers that its favorite watering hole now has an alligator in it, and so reverts to a different water source).

4 Discussion

The new rule demonstrates some features of human-like learning. Humans are known to increase decision-making time as the number of options increases, in a relationship known as Hick's law [15]. Our new rule exhibits similar increasing decision time in Fig. 3, while the conventional learning algorithm does not. A paradox-of-choice effect is also observed in Fig. 5, where the new rule outperforms conventional learning until the number of choices becomes large. Humans exhibit this trade-off as well, where "selections made from large assortments can lead to weaker preferences" [9] though it should be noted that the relationship between number of choices and the choice overload effect in humans is complex [10]. The new rule is derived from a recently proposed neuronal predictive learning rule, and thus may encapsulate some basic principles of learning and intelligence that exist at both the neuronal and system levels. We hope this paper will add to the important conversation around A.I. that can adapt to the constant environmental changes of the real world.

The topics of adaptability and continuous learning represent a growing research field [6,14,17,18], and paradigms for detecting and responding to environmental change do exist in the machine learning literature. For example, model-based reinforcement-learning approaches maintain an internal model of the world, with which new experiences can be compared to detect environmental changes. Previous work has stored world models and switched between them when recent experiences indicated an environmental change [7,8], adapted time series change-point algorithms to detect environmental changes [23], and used consciousness-inspired approaches to improve the generalization of a model to a new task [32]. However, these approaches require maintenance of a world model, which can be costly. Ultimately the quick, model-free effect of our rule could work well in conjunction with the more complex goal-oriented-planning effect of a model-based approach: the brain employs both model-free and model-based mechanisms [26], and the combination likely holds promise for A.I. as well.

Another related approach is transfer-learning or meta-reinforcement learning, which aims to accelerate learning in new tasks from a previously experienced family. One meta-RL approach [5] uses a particular recurrent (memory-equipped) network architecture that learns general features of a task family through back-propagation, allowing the recurrent dynamics to quickly tune into details of a

new task from the family, in what is thought to be a brain-like mechanism [29]. Meta-RL is currently an active research field [11,13,28]. This general approach could be seen as adaptation through knowledge transfer, though unfortunately the network must be informed (reset) each time the task changes. Again, the quick memory-free effect provided by our rule could work well in conjunction with such transfer-learning methods, resulting in more human-like learning.

The idea of modulating a prediction error appears elsewhere in machine learning literature, and modulating the error in different ways or by different signals produces different effects. Here we have shown that modulating reward prediction error by action probability creates a human-like adaptation-to-change effect, including improved performance in simple but dynamic tasks, as well as a paradox-of-choice effect. Conversely, the Inverse Propensity Score Estimation (IPSE) approach used in counterfactual learning uses the inverse of the probability as a modulating factor [12,16]. This can have the effect of de-biasing learning from data collected in a population that differs from a target population. However, during online learning of dynamic tasks it would result in slower adaptation; opposite to our rule. We could also consider REINFORCE-style reinforcement learning algorithms, which modulate a prediction error by a "characteristic eligibility" term that expresses the gradient of the action probability with respect to the parameter being updated [31]. This quickly makes rewarding actions more likely - in static environments where the gradient has consistent meaning. Our rule, on the other hand, demonstrates a similar learning effect in dynamic tasks. Making predictions is a central operation of the brain, and it is likely that neural circuits modulate prediction errors in many ways to get the right effect at the right time, creating what we know as human-like learning.

Among the various effects that can be obtained by modulating prediction errors in different ways, we believe the one proposed here deserves special future study for two reasons. First, the ability to cope gracefully in dynamic situations is still relatively understudied (high-profile successes of machine learning are typically in static environments like games). Second, since this new RL rule is derived from a biologically plausible neuronal learning rule, it creates a link between neuron learning and system-level learning which could shed light on universal principles of learning and intelligence.

5 Code

See https://github.com/echalmers/modulated_td_error for code accompanying this paper. Experiments described in this paper used this code and were executed on commodity hardware without a GPU.

Acknowledgements. This work was supported by Compute Canada, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Canadian Institutes of Health Research (CIHR) grants to Artur Luczak.

References

1. Almeida, L.B.: A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In: Artificial Neural Networks: Concept Learning, pp. 102–111. IEEE Press, January 1990
2. Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. J. Mach. Learn. Res. **3**(Nov), 397–422 (2002)
3. Baldi, P., Pineda, F.: Contrastive learning and neural oscillations. Neural Comput. **3**(4), 526–545 (1991). https://doi.org/10.1162/neco.1991.3.4.526
4. Berg, E.A.: A simple objective technique for measuring flexibility in thinking. J. Gen. Psychol. **39**(1), 15–22 (1948). https://doi.org/10.1080/00221309.1948.9918159
5. Botvinick, M., Ritter, S., Wang, J.X., Kurth-Nelson, Z., Blundell, C., Hassabis, D.: Reinforcement learning, fast and slow. Trends Cogn. Sci. **23**(5), 408–422 (2019). https://doi.org/10.1016/j.tics.2019.02.006
6. Caccia, M., et al.: Online Fast Adaptation and Knowledge Accumulation (OSAKA): a new approach to continual learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 16532–16545. Curran Associates, Inc. (2020)
7. Chalmers, E., Contreras, E.B., Robertson, B., Luczak, A., Gruber, A.: Context-switching and adaptation: brain-inspired mechanisms for handling environmental changes. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 3522–3529, July 2016
8. Chalmers, E., Luczak, A., Gruber, A.J.: Computational properties of the hippocampus increase the efficiency of goal-directed foraging through hierarchical reinforcement learning. Front. Comput. Neurosci. **10**, 128 (2016)
9. Chernev, A.: When more is less and less is more: the role of ideal point availability and assortment in consumer choice. J. Consum. Res. **30**(2), 170–183 (2003). https://doi.org/10.1086/376808
10. Chernev, A., Böckenholt, U., Goodman, J.: Choice overload: a conceptual review and meta-analysis. J. Consum. Psychol. **25**(2), 333–358 (2015). https://doi.org/10.1016/j.jcps.2014.08.002
11. Dorfman, R., Shenfeld, I., Tamar, A.: Offline meta reinforcement learning – identifiability challenges and effective data collection strategies. In: Advances in Neural Information Processing Systems, vol. 34, pp. 4607–4618. Curran Associates, Inc. (2021)
12. Dudik, M., Langford, J., Li, L.: Doubly robust policy evaluation and learning, May 2011. https://doi.org/10.48550/arXiv.1103.4601
13. Fallah, A., Georgiev, K., Mokhtari, A., Ozdaglar, A.: On the convergence theory of debiased model-agnostic meta-reinforcement learning. In: Advances in Neural Information Processing Systems, vol. 34, pp. 3096–3107. Curran Associates, Inc. (2021)
14. Harrison, J., Sharma, A., Finn, C., Pavone, M.: Continuous meta-learning without tasks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 17571–17581. Curran Associates, Inc. (2020)
15. Hick, W.E.: On the rate of gain of information. Q. J. Exp. Psychol. **4**(1), 11–26 (1952). https://doi.org/10.1080/17470215208416600
16. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**(260), 663–685 (1952). https://doi.org/10.1080/01621459.1952.10483446

17. Kwon, J., Efroni, Y., Caramanis, C., Mannor, S.: Reinforcement learning in reward-mixing MDPs. In: Advances in Neural Information Processing Systems, vol. 34, pp. 2253–2264. Curran Associates, Inc. (2021)
18. Liu, H., Long, M., Wang, J., Wang, Y.: Learning to adapt to evolving domains. In: Advances in Neural Information Processing Systems, vol. 33, pp. 22338–22348. Curran Associates, Inc. (2020)
19. Luczak, A., Kubo, Y.: Predictive neuronal adaptation as a basis for consciousness. Front. Syst. Neurosci. **15**, 767461 (2021). https://doi.org/10.3389/fnsys.2021.767461
20. Luczak, A., McNaughton, B.L., Kubo, Y.: Neurons learn by predicting future activity. Nat. Mach. Intell. **4**(1), 62–72 (2022). https://doi.org/10.1038/s42256-021-00430-y
21. Milner, B.: Effects of different brain lesions on card sorting: the role of the frontal lobes. Arch. Neurol. **9**(1), 90–100 (1963). https://doi.org/10.1001/archneur.1963.00460070100010
22. Neftci, E.O., Averbeck, B.B.: Reinforcement learning in artificial and biological systems. Nat. Mach. Intell. **1**(3), 133–143 (2019). https://doi.org/10.1038/s42256-019-0025-4
23. Padakandla, S., Prabuchandran, K.J., Bhatnagar, S.: Reinforcement learning algorithm for non-stationary environments. Appl. Intell. **50**(11), 3590–3606 (2020). https://doi.org/10.1007/s10489-020-01758-5
24. Scellier, B., Bengio, Y.: Equilibrium propagation: bridging the gap between energy-based models and backpropagation. Front. Comput. Neurosci. **11**, 24 (2017)
25. Schwartz, B., Kliban, K.: The Paradox of Choice: Why More Is Less. Brilliance Audio, Grand Rapids, Mich., unabridged edition, April 2014
26. Steinke, A., Lange, F., Kopp, B.: Parallel model-based and model-free reinforcement learning for card sorting performance. Sci. Rep. **10**(1), 15464 (2020). https://doi.org/10.1038/s41598-020-72407-7
27. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, 2nd edn. A Bradford Book, Cambridge (1998)
28. Tang, Y., Kozuno, T., Rowland, M., Munos, R., Valko, M.: Unifying gradient estimators for meta-reinforcement learning via off-policy evaluation. In: Advances in Neural Information Processing Systems, vol. 34, pp. 5303–5315. Curran Associates, Inc. (2021)
29. Wang, J.X.: Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. **21**(6), 860–868 (2018). https://doi.org/10.1038/s41593-018-0147-8
30. Wang, J.X., et al.: Learning to reinforcement learn, January 2017
31. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**(3), 229–256 (1992). https://doi.org/10.1007/BF00992696
32. Zhao, M., Liu, Z., Luan, S., Zhang, S., Precup, D., Bengio, Y.: A consciousness-inspired planning agent for model-based reinforcement learning. In: Advances in Neural Information Processing Systems, vol. 34, pp. 1569–1581. Curran Associates, Inc. (2021)