# A role for phasic serotonergic signaling in regulating augmentations during open-ended learning

**Caroline Haimerl, Daniel McNamee**
Neuroscience Programme
Champalimaud Centre for the Unknown
Lisbon, Portugal
caroline.haimerl@research.fchampalimaud.org

## Abstract

Humans and animals need to rapidly adapt to dynamically changing environments given only few experiences while high-performance artificial systems require large datasets. In order to bridge this gap, we consider data augmentations, which have been shown to substantially improve the data-efficiency and generalization capabilities of many machine learning models including reinforcement learning agents. However, how augmentation should be coordinated online during open-ended interactions with the world is unclear. We take inspiration from the brain in addressing this issue. Encountering unexpected environment states (signaled by state prediction errors, SPEs) has been associated with the phasic release of serotonin, a neurotransmitter known to mediate cognitive flexibility in humans. We hypothesize that serotonin triggers augmentations and that this facilitates adaptation to novel environments. In our agent-based simulations, learning from augmentations improves state-prediction in unfamiliar contexts within a minimal circular environment and in gridworlds. Furthermore, we find that augmentations timed to high SPEs are particularly effective. These preliminary results are consistent with a functional role for serotonergic neuromodulation in open-ended adaptation of natural and artificial systems based on regulating the augmentation of experience.

## 1 Introduction

Data augmentation is a particularly useful machine learning strategy for the regime of sparse data and large systems such as brains, where generalization is a critical property [1]. It has emerged as a instrumental method for training large models especially in the self-supervised setting [2, 3, 4, 5]. We investigate whether an analogous computational process may be implemented in the brain and propose a mechanistic implementation. Specifically, we examine a novel hypothesis involving an ancient, and widely conserved, neuromodulator, serotonin (5-HT) [6]. We postulate that the functional role of serotonin in the brain is to cause the generation of neural augmentations of experience. That is, long-range projections of serotonergic activity throughout the brain trigger stochastic augmentations of experiences in order to enhance generalization and aid learning from sparse environmental interactions in biological systems.

We develop our theory by exploring its implications via the simulation of reinforcement learning (RL) agents engaged in open-ended learning of novel environments [7]. Data augmentation has recently been a focus of study in the RL context whereby an agent's experience of state-action trajectories is augmented [8, 9, 3, 10]. Indeed, novel augmentation methods, such as random amplitude scaling, have been proposed specifically for continuous state-based control [10]. A critical distinction can be drawn between data augmentation in the offline [10] versus online [3] scenarios. In the former case, previous experiences in the agent's buffer are randomly sampled and augmented during periods of quiescence.

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

In the latter case, augmentations may be generated from a recent experience of the agent while environmental engagement is ongoing. Particularly in this online case, it is unclear *when* precisely a continuous stream of open-ended experiences should be interrupted to generate an augmentation. In this study, we propose to treat online augmentation as a meta-task whereby the agent can optimize *when* to generate augmentations. Specifically, given a limited budget of augmentations, when should they be applied to the agent's experience? Inspired by neuroscientific studies regarding the role of serotonergic modulation in the brain, we propose that a state novelty or uncertainty signal may serve the purpose of determining when augmentations are generated. Specifically, we hypothesize that the phasic activity of serotonin-producing neurons triggers experiential augmentations when an animal or human encounters a novel environment state. In order to examine whether this could contribute to the performance of open-ended learning mechanisms, we investigated a key prediction in our simulations. We tested whether an RL agent, endowed with the capacity to regulate the generation of augmentations of recent experiences according to an environment state novelty signal, learns and generalizes better than an agent which randomly augments regardless of the interactive context of the agent.

## 2 Methods

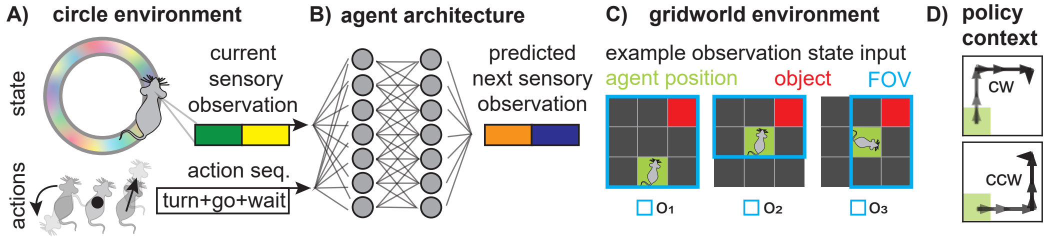

Figure 1: **A.** We used a minimal circle environment with random two-dimensional real-valued observations to illustrate our theory. The agent (a virtual rodent) can traverse the circle via egocentric actions "turn", "go", "wait". The current sensory observation (green and yellow rectangle) and the egocentric action sequence are inputted to a neural network. **B.** A neural network was trained to predict the next sensory observation (orange and red rectangle) from the previous sensory observation and action sequence. **C.** We trained an analogous network in an open gridworld with one object (red). The agent observes one of 28 possible egocentric observations (field of view, FOV, in blue). The input to the network is the scalar corresponding to the observation state (1-28). **D.** We examined the generalization properties of our augmented learning algorithm between two possible behavioral policy contexts; namely, whether the agent was biased to traverse the space in a cw (top) or ccw (bottom) fashion.

We study the problem of state prediction in Markov decision processes. Agents receive continuous or discrete state observations $\mathbf{o}_t$ together with a sequence of $k$ discrete actions $\mathbf{a}_{t:t+k}$ and aim to predict the next state observations $\mathbf{o}_{t+k}$. The agent consists of a two-layer feedforward neural network with 100 neurons per layer, and an output layer that maps to predicted state observation space (Fig. 1B). Weights were optimized using backpropagation with Adam [11], either mean squared error (MSE) or cross-entropy loss and a learning rate of 0.001 for the circle and 0.0001 for the gridworld environments (see below for a detailed description of the environments).

**Augmentations:** The agent is able to produce and learn from augmentations of its experienced trajectories. Augmentations are random combinations of experienced environmental states and actions that the agent "imagines" traversing. Concretely, during augmentations the neural network is trained on such a set of randomly generated state-action-state sequences. *Triggered augmentations* occur after a high state prediction error (SPE) while *random augmentations* can occur at any point with their probability matched to the rate of triggered augmentations. We test the state-prediction performance of the dynamic triggered-augmentation agent against random-augmentation and no-augmentation baselines in two environments.

**Circle environment:** The first environment (Fig. 1A) is a circular track which an agent traverses in one of two contexts; clockwise (cw) or counterclockwise (ccw). The circle has 36 discrete states each with a unique continuous sensory 2-dimensional observation vector that depends on the angular position and the direction that the agent faces (i.e. at one angle the agent might get a sensory observation of $\mathbf{o} = [1, 2]$ if facing cw, or $\mathbf{o} = [2, 1]$ if facing ccw). The agent can take one of three actions: a forward step to reach the next angular state according to its current direction, a turn that changes its direction and flips its sensory observation but does not change its angular state, or a wait action. The agent predicts the next sensory observation $\mathbf{o}_{t+k}$ given the current state observation $\mathbf{o}_t$ and a sequence of either wait or forward actions $\mathbf{a}_{t:t+k}$ (for instance $\mathbf{a}_t = [1, 1, 1, 1, 0, 0, 0, 0]$). We train the agent in an unidirectional context (cw or ccw) on the MSE loss of its predictions and true observations. We then test the agent's ability to predict states of cw or ccw trajectories, one of which is unfamiliar to the agent.

**Gridworld environment:** We further test our algorithm in a minigrid environment [12] with a 3 × 3 grid and 3 possible actions, turn left, turn right, step forward (Fig. 1C). The egocentric field of view depends on the agent's orientation in the environment leading to 28 possible 7 × 7 pixel observation states [12], which we simplify into an observational state category label, that is provided to the network as a scalar input. We train the network on this state input and an action sequence to predict the next state category using a cross-entropy loss. The agent moves through the gridworld following one of two possible policies. In the first, clockwise, policy context the agent takes mostly actions that result in cw trajectories. With a probability of $10\%$ the agent still chooses random actions. The second context follows the same rules, but ccw (Fig. 1D).

## 3 Results

State prediction poses a key problem in model-based RL, especially if environmental structure changes or multiple contexts are encountered by an agent. Here we propose a new approach which regulates automatic augmentations and enhances zero-shot performance in the online learning setting. This method can be used widely and flexibly combined with different policy learning algorithms.

**Learning protocol and zero-shot performance:** First, we train an agent in the circular environment given a unidirectional policy (cw context, $500$ training batches of $64$ trials, maximum of $k = 8$ forward actions, test MSE in familiar context $< 0.01$). We then test the agent's ability to predict the next state observation in a new, unfamiliar context, ccw trajectories. We find that state prediction errors increase significantly in this environmental condition (t-test $p = 0.004$, Fig. 2A). Next we test whether augmentation can improve the performance of the trained agent in the unfamiliar context and find significantly decreased MSE for the same trained model after additional training on augmented random trajectory data (t-test $p = 0.03$, Fig. 2A). Similarly we train an agent in the gridworld environment on a directional policy, where the agent moves in the environment with a strong cw or ccw bias (see Methods). An agent trained on one context (e.g. cw) has significantly decreased performance in a new context (e.g. ccw), measured as $\%$ correctly predicted next state categories (t-test $p < 1e - 10$ Fig. 2B). This zero-shot context-switch performance is significantly increased through augmentations (t-test between performance of trained versus trained + augmented agents in

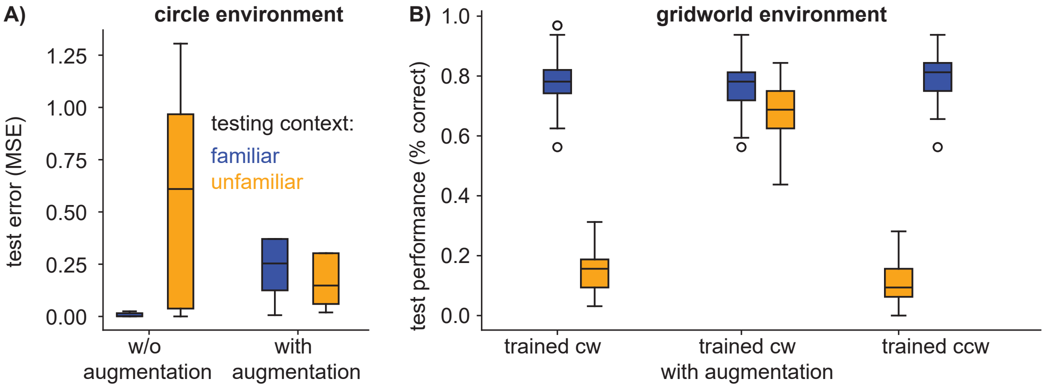

Figure 2: Augmentations improve zero-shot state prediction in both the circle **A** and gridworld B environment. Boxplot indicates inter-quartile range (IQR) and median, whiskers indicate $1.5\times$ IQR.

new context $p < 1e-10$), without significant decreases in performance in the familiar cw context (t-test between trained and trained+augmented agents for performance in familiar context $p = 0.12$).

**Online augmentations:** We hypothesize a role for serotonin in the brain for triggering stochastic augmentations of recent experiences to accelerate adaptation to novel environmental contexts. Specifically, increased SPE, associated with unfamiliar sensory experiences and signaled by serotonin, activates augmented neural representations from which the brain learns more generalized knowledge. We tested whether such timed augmentations following high SPEs in an online learning setting improve performance over randomly interspersed augmentations.

First, we took an agent trained on one context of the circular environment and test it on multiple blocks of trials from either the familiar or unfamiliar context. SPEs is substantially higher in the unfamiliar context blocks (Fig. 3A). We then rerun this block-design but allow the agent to trigger augmentations whenever the SPE surpasses a threshold $\theta$ (i.e. MSE $> \theta$). In this work, the threshold is arbitrarily set to $\theta = 0.05$ however a future avenue of investigation would be to meta-optimize this hyperparameter. We find a significant difference between the average online SPE without and with augmentations (average MSE without augmentations $0.11$, average MSE with augmentations $0.04$, t-test $p < 1e-14$, Fig. 3B&C).

In order to test, whether the timing of the augmentations is crucial, we compare to simulations where augmentations happen randomly, independent of SPE, but keep the total number of augmentations the same as for the triggered augmentations. We find that while random augmentations are also able to improve the overall performance of the agent, SPE-triggered augmentations are significantly better than random augmentations (ave. MSE with random augmentations $0.05$, t-test $p < 1e-4$, Fig. 3C). We confirm the findings of SPE-triggered augmentations in the gridworld task. We find that both random and SPE-triggered augmentations improve the trained model's performance (% correct across context-blocks) significantly, and that SPE-triggered augmentations outperform

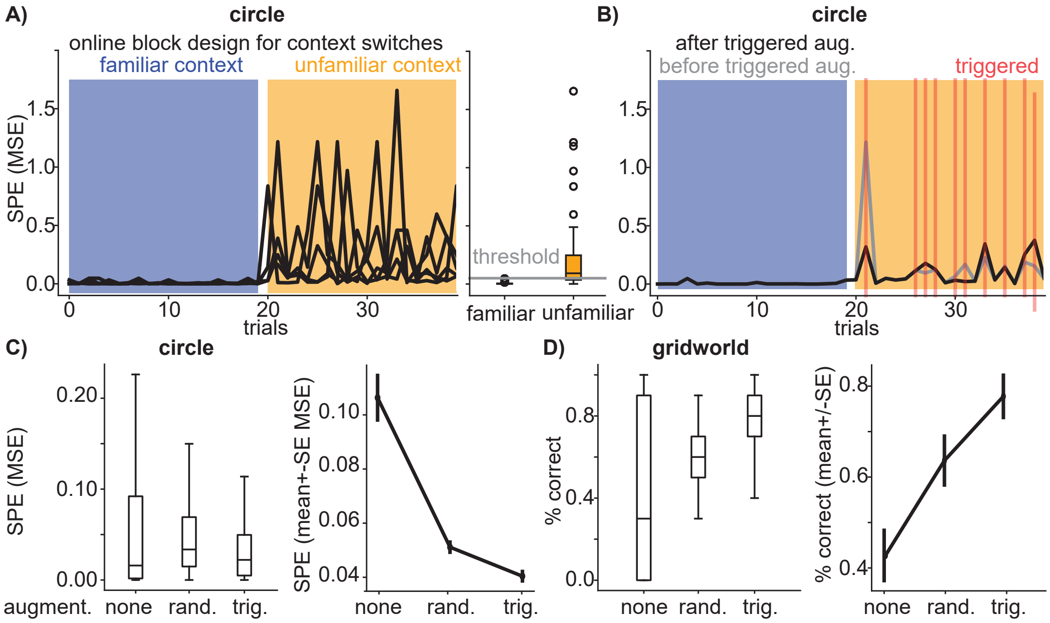

Figure 3: SPE-triggered augmentations lead to better performance than random augmentations in the circle environment (A-C) and gridworld (D). A) Left: State predictions across familiar/unfamiliar context blocks in circular environment (two blocks shown). Black lines indicates SPE of individual simulations. Right: Distribution of SPE by context. Grey line indicates threshold of $0.05$. Boxplot statistics as in Fig. 2. B) Augmentations are triggered when SPE surpasses threshold (red lines). Grey line indicates the MSE before the augmentation and black line indicates the MSE after the augmentation. C) SPE across contexts in circle environment for no augmentations, random augmentations and triggered augmentations. Left plot shows the distribution, median and IQR, right plot shows mean and standard error. D) As C but for the gridworld environment.

random augmentations (trained model without augmentation ave. $43\%$ correct, with SPE-triggered augmentations $78\%$, with random augmentations $64\%$, t-test between trained and SPE-triggered augmentations $p < 1e - 13$, t-test between random and SPE-triggered augmentations $p < 1e - 5$, Fig. 3D). These results illustrate that augmentations of recent experiences not only help an agent to constantly adapt its internal models of state transitions to novel environmental context, but that timing these augmentations to follow high SPEs significantly increases their positive impact.

## 4 Discussion

With respect to previous work on the role of serotonin in the brain, our theory provides a new perspective which integrates models relating 5-HT to state prediction error signals indicating contextual novelty, with evidence suggesting a contribution of serotonin towards cognitive flexibility [13, 14, 15]. We explicitly model serotonin signals as state prediction errors; these drive augmentations, which in turn facilitate generalization and therefore increase cognitive flexibility as measured by zero-shot inference in this work. In support of our hypothesis, we have performed preliminary simulations investigating a role for state novelty signals in regulating experience augmentations during open-ended learning. Such signals have long been explored as intrinsically motivating (IM) objective functions for driving exploration in RL agents seeking to construct useful state and action representations - a key goal of open-ended learning [16]. Here, we show that such IM signals can also be productively applied to automatically trigger and modulate augmentations of the agent's experience albeit in relatively simple environments. However, analogous IM measures have recently been scaled to high-dimensional continuous problems using random network distillation [8] which we plan to leverage for future work in applying our framework to more complex environments.

Though our simulations have focused on serotonergically triggered augmentations in the online awake state amid bouts of agent-environment engagement, we envision this to be mutually consonant with the proposed role of dreaming as an augmentation process during offline sleep states [17]. Indeed, serotonin is known to affect both the structure of sleep and the content of dreams [18]. This also aligns with a previous proposal conceptualizing diffusive hippocampal replay, which specifically occurs during the sleep state [19], as augmentations optimized for spatial learning [20]. This observation opens up another potential avenue for future investigation regarding possible mechanistic implementations of augmentations in the brain. From a computational perspective, an intriguing possibility is that qualitatively distinct forms of augmentations may be symbiotically generated in the brain during the online and offline phases with different generalization properties that reciprocally benefit.

In future work, we expect this complex interplay to provide further theoretic guidance regarding our hypothesis relating serotonin and experience augmentations across distinct states of consciousness throughout open-ended world interactions. While previous work has focused on the question of *what* augmentations to generate leading to automated algorithms for augmentation selection [21], our theory introduces the problem of *when* to trigger augmentations. In general, best practices for automating the initiation and parametrization of augmentations remain poorly understood [5]. Too little augmentation results in limited generalization and inflexibility while too much augmentation leads to inaccurate hallucinatory inferences and bias. Potentially, the brain may provide insights into effective regulation strategies regarding when and how augmentations should be generated.

## References

[1] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 2019.

[2] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

[3] Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In *Advances in Neural Information Processing Systems*, volume 34, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[5] Chi-Heng Lin, Chiraag Kaushik, Eva L. Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

[6] Romain Ligneul and Zachary F. Mainen. Serotonin. *Current Biology*, 33(23):1216–1221, 2023.

[7] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2 edition, 2018.

[8] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. *ICLR*, 2019.

[9] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. A simple randomization technique for generalization in deep reinforcement learning. *ICLR*, 2020.

[10] Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 5402–5415, 2021.

[11] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *ICLR*, pages 1–15, 2015.

[12] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

[13] H. F. Clarke, J. W. Dalley, H. S. Crofts, T. W. Robbins, and A. C. Roberts. Cognitive inflexibility after prefrontal serotonin depletion. *Science*, 304(5672):878–880, 2004.

[14] Sara Matias, Eran Lottem, Guillaume P Dugué, and Zachary F Mainen. Activity patterns of serotonin neurons underlying cognitive flexibility. *eLife*, 2017, 2017.

[15] Cooper D. Grossman, Bilal A. Bari, and Jeremiah Y. Cohen. Serotonin neurons modulate learning rate through uncertainty. *Current Biology*, 32(3):586–599.e7, 2022.

[16] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019. arXiv: 1908.06976v2 [cs.LG].

[17] Erik Hoel. The overfitted brain: Dreams evolved to assist generalization. *Patterns*, 2(5):100244, 2021.

[18] Barry L. Jacobs and Michael E. Trulson. Dreams, hallucinations, and psychosis - the serotonin connection. *Trends in Neurosciences*, 2:276–280, Jan 1979.

[19] F. Stella, P. Baracskay, J. O'Neill, and J. Csicsvari. Hippocampal reactivation of random trajectories resembling brownian diffusion. *Neuron*, 102(2):450–461.e7, 2019.

[20] Daniel C McNamee. The generative neural microdynamics of cognitive processing. *Current Opinion in Neurobiology*, 85:102855, 2024.

[21] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2019.