Learning from the Unexpected: A Von Economo Neuron-Inspired Efference Copy Architecture for Continual Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Efficient continual learning requires computing the orthogonal subspaces in the learnt parameter space, separating parameters to be updated for the new task from the ones that should be retained. This implies a continuous one-to-one mapping between the input and generated output to distinguish which inputs are caused by the model output, which is also required for animal self-recognition in the mirror. We derived a novel reinforcement learning framework inspired by von Economo neurons (VEN), a type of neuron that is only found in the small number of species with documented mirror self-recognition. Morphology of VEN suggests that it compares efference copy motor plan to the sensory input, helping to yield truly unpredictable residual signals for learning. Unlike the reward-prediction error of standard TD learning, these residuals isolate self-caused from externally induced change. We distilled VEN architecture into an explicit efference copy circuit for reinforcement learning. In Atari-2600 games and control-suite environments, VEN-augmented DDQN learned much more efficiently than capacity-matched controls under self-position perturbation, actively probed its own position in the game, and performed continual learning across sequentially presented tasks. This architecture provides a computational account of how preserved latent state information could support continual learning.

1 Introduction

Retraining a network on a shifted distribution usually erases what it knew (McCloskey & Cohen, 1989; French, 1999). The geometry of high-dimensional representations makes this the default outcome: old and new data separate along near-orthogonal feature directions (Gorban & Tyukin, 2018; Stringer et al., 2019; Gur-Ari et al., 2018), so gradients from the new task populate directions that carry nothing of the old solution (Ramasesh et al., 2021). Methods that constrain updates to directions unimportant for old tasks (Kirkpatrick et al., 2017; Zenke et al., 2017; Farajtabar et al., 2020) presuppose that the old solution has already been delineated in parameter space. A network that learns continually instead needs a component of its representation that persists across distributions by construction.

Continual reinforcement learning gives this abstract problem a concrete structure. For an embodied agent, every task is experienced through the same sensorimotor loop: the same actions move the same limbs or paddle, and the sensory consequences of those actions obey the same dynamics despite the fact that the goal, the opponent, or the viewpoint may change. Input changes therefore decompose into a self-caused part ("what my action produced") and a world-caused part. The first part is invariant across tasks; a representation that explicitly carries it has a natural candidate for the persistent core that continual learning requires. Interestingly, such capability is also essential for animal recognition of itself in a mirror: it attributes a fully self-controlled visual stream to its own motor commands (Gallup, 1970; Gallagher, 2000).

Section 3 recalls one biological clue: von Economo neurons (VENs), found in the anterior cingulate and fronto-insular cortex of humans and of the small set of species with documented mirror self-recognition, have a bipolar morphology consistent with comparing thalamic sensory input against efferent copies of motor plans (von Economo, 1926; Nimchinsky et al., 1995; Allman et al., 2010). We translated this clue into a minimal circuit and tested it in a reinforcement learning setting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related work

Sequential gradient descent overwrites earlier learning and leads to catastrophic forgetting (McCloskey & Cohen, 1989; Ratcliff, 1990; French, 1999). Regularization (Kirkpatrick et al., 2017; Zenke et al., 2017), gradient projection (Lopez-Paz & Ranzato, 2017; Farajtabar et al., 2020), and rehearsal (Rebuffi et al., 2017; Rolnick et al., 2019) all presuppose that the old solution is already delineated in parameter space. Near-orthogonal feature directions make forgetting the default and a retained-versus-updated split hard to find (Gur-Ari et al., 2018; Stringer et al., 2019; Gorban & Tyukin, 2018). In reinforcement learning the component that persists by construction is the sensory consequence of the agent's own action, and an efference copy is the minimal circuit that isolates it. Continual-RL methods instead expand capacity (Rusu et al., 2016), consolidate parameters (Schwarz et al., 2018; Kaplanis et al., 2019), or replay experience (Rolnick et al., 2019; Khetarpal et al., 2022). VEN does none of these: it changes what the representation makes distinguishable, and its retention benefit appears with replay and vanishes without it.

Separate from these parameter-update methods, a motor-command copy that predicts self-caused sensation is the reafference principle (von Holst & Mittelstaedt, 1950; Sperry, 1950), with substrates in electric fish (Bell, 1981), crickets (Poulet & Hedwig, 2002), and primate corollary discharge (Crapse & Sommer, 2008; Blakemore et al., 2000), and with the forward internal model as its computational form (Wolpert et al., 1995; Kawato, 1999; Wolpert & Ghahramani, 2000). The free-energy principle and predictive coding cast the same idea as surprise minimization (Friston et al., 2006; Friston, 2010; Clark, 2013; Rao & Ballard, 1999; Friston, 2005; Friston et al., 2010; Adams et al., 2013), but leave the self/world split generic; deep-RL readings stay interpretive (Ueltzhöffer, 2018). Dyna, action-conditional video prediction, and modern world models learn that forward map for imagination or control (Sutton, 1990; Oh et al., 2015; Ha & Schmidhuber, 2018; Hafner et al., 2019; Schrittwieser et al., 2020; Hafner et al., 2025); curiosity uses the error as reward (Schmidhuber, 1991; Pathak et al., 2017; Burda et al., 2019; Stadie et al., 2015); joint-embedding methods predict latents under a stop-gradient (LeCun, 2022; Assran et al., 2023; Jaderberg et al., 2017). VEN follows this comparator tradition with two departures: the prediction does not cancel reafference, and the predicted latents enter the value head rather than serving as a scalar error, a reward bonus, or a purely representational loss. For a change of viewpoint rather than of task, equivariant networks and augmentation handle symmetries known in advance (van der Pol et al., 2020; Yarats et al., 2021; Tobin et al., 2017); our flip is unannounced, closest on the ML side to causal confusion (de Haan et al., 2019). Bell's negative image and the ICM forward model are the nearest ancestors; neither feeds the prediction to the value head, and neither addresses continual learning.

3 Von Economo neurons

Mirror self-recognition requires the same split: the animal attributes a fully self-controlled visual stream to its own commands (Gallup, 1970; Gallagher, 2000). Besides humans, only great apes, elephants, bottlenose dolphins, magpies, and monkeys after supervised training are reported to pass (Plotnik et al., 2006; Reiss & Marino, 2001; Prior et al., 2008; Chang et al., 2015). VENs, a specific type of large bipolar excitatory neurons (von Economo, 1926; Nimchinsky et al., 1995; 1999; Allman et al., 2005), are only found in that same set of species (Allman et al., 2010; Hakeem et al., 2009; Butti et al., 2013; Evrard et al., 2012). Furthermore, VEN is distributed in layer Vb of anterior cingulate and fronto-insular cortex, areas tied to self-related processing (Han et al., 2010; 2008; Ge et al., 2009). It has a distinct morphology with an apical dendrite reaching layer IV (thalamic sensory input) and a basal dendrite reaches layer VI (cortico-cortical motor-planning input) (Nimchinsky et al., 1995; Allman et al., 2010); the thick shafts favour near-linear transmission of small inputs (Koch, 1999). Patch clamp recording of VEN shows increased excitability and characteristic bursting (Ke et al., 2026), with hyper-activation by layer VI input (Liu et al., 2026). Their morphology and electrophysiology are therefore consistent with the function to integrate an efferent motor plan with sensory input and facilitating to learn about the unpredicted residual (Allman et al., 2010; Seeley et al., 2007; Craig, 2009). Unlike a reward-prediction error, that residual separates self-caused from world-caused change. We hypothesize that VENs implement this comparison, and that the same computation can support continual learning.

4 VEN circuit architecture

4.1 Basic structure

Throughout, s_t is the visual input, a_t is the motor output, z_t is the visual embedding, and $\hat{z}_t$ is the predicted visual embedding. In DDQN for Atari 2600 (Mnih et al., 2015; van Hasselt et al., 2016), three convolutional layers map s_t to z_t, and a head of widths $448 \to 512 \to 6$ estimates the Q-value of each action. The motor output a_t is the action selected from those Q-values. At each step the agent receives a visual input

$$s_t \in \mathbb{R}^{84\times84\times4}, \tag{1}$$

a stack of four frames, picks an action a_t from a discrete set of size 6, and receives reward r_t. An encoder f_θ (three convolutional layers and one fully connected layer) maps that stack to a 64-dimensional embedding,

$$z_t = f_\theta(s_t) \in \mathbb{R}^{64}. \tag{2}$$

A Q head g_ψ maps latent features to Q-values.

4.2 VEN module

On top of this backbone we add a small feed-forward network as the efference copy. It is a two-layer fully connected module

$$V_\phi : \mathbb{R}^{70} \to \mathbb{R}^{64} \to \mathbb{R}^{64} \tag{3}$$

that receives the current embedding and a one-hot action and predicts the next embedding,

$$\hat{z}_{t+1} = V_\phi(z_t, a_t). \tag{4}$$

The input dimension is $64 + 6 = 70$. Counting biases, the two affine maps have $(70 \times 64 + 64) + (64 \times 64 + 64) = 8704$ parameters. The prediction of the current embedding, the quantity available at decision time, is the same map shifted by one step,

$$\hat{z}_t = V_\phi(z_{t-1}, a_{t-1}). \tag{5}$$

The Q head receives the current embedding, the three preceding true embeddings, and the three preceding predictions,

$$Q(s_t, \cdot) = g_\psi\big([\, z_t,\, z_{t-1},\, z_{t-2},\, z_{t-3},\, \hat{z}_t,\, \hat{z}_{t-1},\, \hat{z}_{t-2}\,]\big), \tag{6}$$

an input of $7 \times 64 = 448$ dimensions. Ablated variants zero selected slots of this 448-dimensional input and/or disconnect the auxiliary loss. Parameter shapes are identical everywhere, so no variant can win by capacity (verified up to an explicit parameter-count-matched control; Table 1). The complete agent of Table 2 totals 302,454 parameters.

Table 1: Parameter counts by module. The flip-matrix agent (Table 2) totals 302,454 parameters, of which VEN is 8,704. The action-reversal pair of Section 5.2 widens the VEN hidden layer to 1,871 and the DDQN head to 662 so that both agents reach $\sim$546.7K trainable parameters.

Agent	Module	Parameters
flip matrix (Table 2)	encoder f_θ	60,784
	VEN V_ϕ (hidden 64)	8,704
	Q head g_ψ (hidden 512)	232,966
	total	**302,454**
action reversal (Section 5.2)	encoder f_θ	60,784
	VEN V_ϕ (hidden 1,871)	252,649
	Q head g_ψ (hidden 512)	232,966
	total	**546,399**
	matched DDQN (head hidden 662), trainable	546,698[†]

[†] The matched DDQN instantiates the 8,704-parameter VEN module but excludes it from the optimizer and never reads its output; the count is trainable parameters only.

4.3 Losses

Let $\mathcal{D}$ be the replay buffer. On a sampled transition (s, a, r, s', d), with $z = f_\theta(s)$, the temporal-difference loss is the double-DQN objective. The next action is selected by the online network and evaluated by the target network (θ^-, ψ^-) (van Hasselt et al., 2016):

$$a^* = \arg\max_{a'} Q(s', a'; \theta, \psi), \tag{7}$$

$$\mathcal{L}_{\text{TD}}(\theta, \psi) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\Big[\big(r + \gamma(1-d)\, Q(s', a^*; \theta^-, \psi^-) - Q(s, a; \theta, \psi)\big)^2\Big]. \tag{8}$$

The efference-copy loss is a prediction loss in embedding space with a stop-gradient on the target. Here s' is the next visual input, so $f_\theta(s')$ is the embedding that $\hat{z}_{t+1}$ is trained to match:

$$\mathcal{L}_{\text{VEN}}(\phi) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\Big[\big\|\text{sg}\big[f_\theta(s')\big] - V_\phi(z, a)\big\|_2^2\Big], \tag{9}$$

where sg denotes stop-gradient: the encoder receives no gradient from the prediction loss, so the predictor cannot collapse the representation (Grill et al., 2020; Chen & He, 2021). The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{TD}} + \lambda \mathcal{L}_{\text{VEN}}, \qquad \lambda = 0.1. \tag{10}$$

Only ϕ receives $\mathcal{L}_{\text{VEN}}$ gradients; θ and ψ receive $\mathcal{L}_{\text{TD}}$ gradients; the target network is a periodic copy. Training otherwise follows the standard Atari recipe; a sweep over $\lambda \in \{0.1, 0.5, 3.0\}$ in a separate gridworld testbed found 0.5 competitive with 0.1 and no clean optimum, so we fix $\lambda = 0.1$ throughout.

In the VEN-enhanced network, the motor-cortex counterpart receives both z_t and $\hat{z}_t$.

Write the true next embedding as

$$z_{t+1} = V^*(z_t, a_t) + e_{t+1}, \tag{11}$$

where V^* is the best action-conditioned predictor available in the training distribution and e_{t+1} is whatever the agent's own action does not predict: an opponent's move, a perturbation of the sensory input, or a change of task convention (such as action direction). The prediction residual $z_{t+1} - \hat{z}_{t+1}$ therefore estimates e_{t+1}. Unlike the scalar TD error, which is defined only against reward, this residual is a vector in the agent's own input stream, and it is nonzero exactly when the world deviates from what the agent's action would have caused. Feeding $\hat{z}$ to the Q head makes that deviation a feature of the value function: at every step the agent sees both what happened and what it believed its action would cause.

5 Experiments

5.1 Mirror symmetric perturbation and self-position identification

In some Atari 2600 games the visual scene is mirror-symmetric. In Pong, for example, if the frames are converted to black-and-white and cropped to the playing field, it is difficult to tell which paddle is controlled by the learning model (hereinafter the "agent") and which by the game program (hereinafter the "opponent") (Figure 1b). In the default setting the agent always plays the right-hand paddle and the opponent the left, and DDQN learns the game well as reported before (Mnih et al., 2015). If we instead flip the visual input randomly at the start of each episode, horizontally (left-to-right, "h"), vertically (top-to-bottom, "v"), or both ("hv"), then to play the game the agent needs to know which paddle is its own and which direction its actions move it. Under this perturbation shape-matched DDQN learns much more slowly and reaches a lower final performance. VEN-enhanced DDQN (hereafter "VEN"), in contrast, performs similarly whether the input is perturbed or not (Figure 1c, Table 2). VEN therefore helps the model know which part of the image is itself.

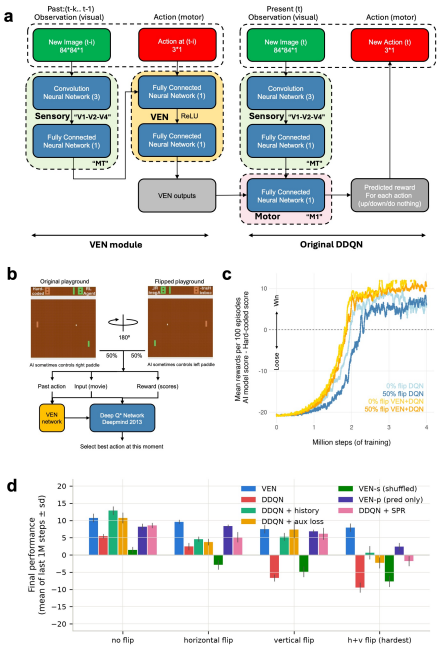

Figure 1: VEN improves learning under mirror-symmetric perturbation. **(a)** VEN architecture. **(b)** Testing game environment. The Pong image is converted to black-and-white (to a similar grayscale of brightness). In each episode, the game image is presented either as is or randomly flipped (horizontal example shown) along the x or y axis. **(c)** Mean reward during training for DDQN and VEN under 100% original (0% flip) or 50% original / 50% horizontally flipped training. **(d)** Performance of VEN ablations and DDQN additions after 4M steps of training.

Table 2: VEN and DDQN performance in binarized Pong (4M steps; reward, mean of the last 1M steps $\pm$ s.d.). Bold marks the best entry in each column.

Architecture	noflip	hflip	vflip	hvflip
VEN	$+10.8 \pm 1.2$	$+\mathbf{9.6 \pm 0.6}$	$+\mathbf{7.5 \pm 1.2}$	$+\mathbf{7.9 \pm 1.2}$
VEN-p (no z-history)	$+8.2 \pm 0.9$	$+8.4 \pm 0.5$	$+6.8 \pm 0.5$	$+2.4 \pm 1.1$
VEN-s (shuffled actions)	$+1.5 \pm 0.9$	-2.8 ± 1.4	-4.9 ± 1.6	-7.7 ± 1.6
DDQN	$+5.5 \pm 0.6$	$+2.5 \pm 1.1$	-6.7 ± 1.0	-9.5 ± 1.4
DDQN + z-history	$+\mathbf{12.9 \pm 1.2}$	$+4.6 \pm 0.7$	$+5.3 \pm 1.2$	$+0.7 \pm 1.9$
DDQN + aux. pred. loss	$+10.7 \pm 1.6$	$+3.7 \pm 0.9$	$+7.4 \pm 2.5$	-2.3 ± 1.6
DDQN + SPR multi-step aux.	$+8.6 \pm 0.8$	$+5.2 \pm 1.5$	$+6.1 \pm 1.7$	-1.8 ± 1.4

We performed systematic ablation experiments to test whether VEN is necessary for such a task. Removing the VEN output to the Q head, or shuffling the action sequence for VEN, significantly reduced task performance. On the other hand, injecting history observations into shape-matched DDQN, or using a VEN-style input only as an auxiliary loss, only improves the "h"/"v" but not the "hv" condition (Table 2). Therefore, the addition of VEN does not merely result in up-scaling of the network to increase learning capacity. In fact, a parameter-size-matched DDQN never achieves performance similar to VEN. We consider that VEN introduces a particular network architecture that computes sensorimotor integration for future prediction.

Three observations follow from Table 2. Shuffling the action history that VEN sees (VEN-s) abolishes the advantage, so the module is not extra capacity or a generic regularizer. History, an auxiliary prediction loss, and multi-step SPR help under single-axis flips but fail under the combined h+v flip ($+0.7$, -2.3, -1.8 versus $+7.9$); only the value head that reads the prediction at decision time handles both an ambiguous paddle and an ambiguous up/down mapping. With no perturbation, history

alone can improve performance (DDQN+history $+12.9$); the efference copy matters when the agent must identify its own position.

5.2 Within-episode action reversal and behavioral readouts

A harder setting swaps the meaning of the action keys at random moments within an episode, with no visual cue; only a comparison of predicted and observed consequences can reveal the swap. Against a rule-based opponent, VEN reaches $+14.0$ mean trailing-100 score (best $+15.2$) after 20M steps, while a DDQN matched to 546,698 trainable parameters stays at -17.7 (VEN has 546,399; Table 1; Figure 2a).

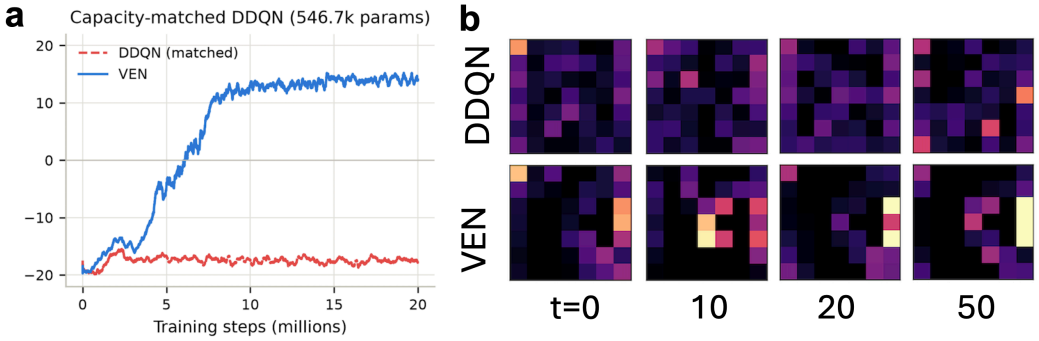

Figure 2: VEN improves learning under action perturbation by actively probing its own location. **(a)** Training performance of VEN and capacity-matched DDQN against a rule-based opponent under the action-perturbation scheme. **(b)** Visual occlusion effect on the action decision for DQN and VEN after action perturbation.

During non-threat periods (ball moving away from the agent's side; 20 evaluation episodes per model), the trained VEN dwells at a wall: 72% of its steps lie within 16 pixels of a wall, holding the paddle at a predictable position by continuously pressing a single key. Any change of key would move it away from that position, so the strategy keeps its own location exactly known and forces the opponent to serve into a geometry it controls; the parameter-matched DDQN, which never resolves the mapping, shows no such organized behavior (22%). When a reversal does arrive, VEN's decision-relevant attention, measured by visual occlusion around the event (30 evaluation seeds; 25 and 44 isolated events for VEN and DQN), is already on the far side of the court, where the consequence of its own last action returns (side-mass 0.56 far versus 0.22 own), while DQN attends to its own side instead (0.27 versus 0.38; Figure 2b). Dwelling at a known position and watching where one's own action comes back from are the two halves of self-testing at the mirror, the maneuvers that animals and young children show on first mirror exposure (Gallup, 1970; Prior et al., 2008); we draw the analogy only at the level of "act to find out what is controlled by you."

5.3 VEN enhances the predictability of future embeddings

If the circuit works as designed, VEN-trained embeddings must predict the future better when actions are known. We trained regression probes (MLPs) to predict z_{t+k} from $[z_{t-h:t}]$ with or without $[a_{t-h:t}]$, on episodes held out from checkpoint rollouts. The action-information gain, ΔR^2, is 0.316 for VEN against 0.083 for DDQN ($h = 8$, $k = 4$), a $3.8\times$ ratio, positive on 100% of held-out episodes, and largely destroyed when the action sequence is shuffled for the probe (the residual gain shrinks to a 1.2–$1.4\times$ ratio, the DDQN level) (Figure 3). Decoding the flip state from VEN encoder embeddings reaches cross-validated AUC 0.989 with four flip-sensitive units, so the representation carries an explicit, compact code of "which world am I in."

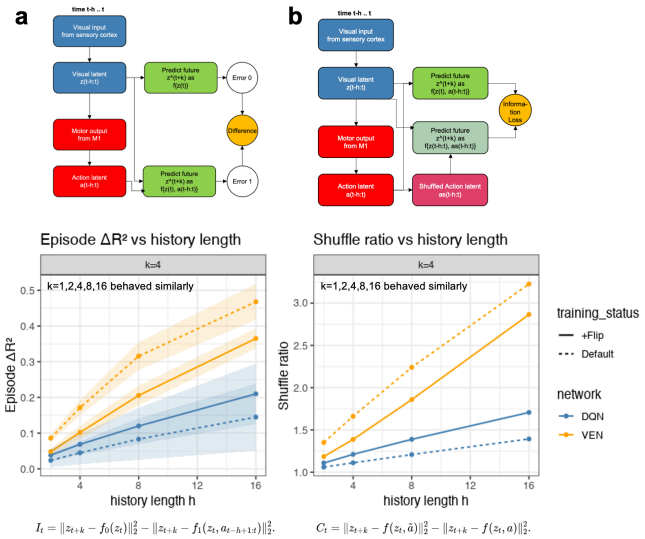

Figure 3: VEN enhances the predictability of the future embedding. **(a)** Prediction-power gain from using action history, for DDQN and VEN. **(b)** Shuffling the action history strongly affects VEN prediction performance, and less so for DDQN. The panels plot episode ΔR^2 and the shuffle ratio against history length; the figure notes that $k \in \{1, 2, 4, 8, 16\}$ behaved similarly.

5.4 Other games and dynamic control suite flip perturbations

The flip perturbation extends to any game with a mirror-symmetric scene. In seven additional Atari games (Boxing, Breakout, FishingDerby, KungFuMaster, RoadRunner, Tennis, VideoPinball; 3 seeds, 3 5M steps), VEN reached a higher best mean reward than DDQN under flip in six (from $+3.2$ in Tennis to $+44$K in VideoPinball). In five classic-control / DM-control environments with per-dimension state flips (2M steps, 4–9 seeds), VEN clearly outperformed DDQN (Table 3). With three Atari seeds the point estimates favour VEN in six of seven games, but only FishingDerby, RoadRunner, and VideoPinball are individually significant; the five DM-control comparisons are all significant ($t \geq 3.3$).

Table 3: Performance of VEN and DDQN on other games and control environments. Reward is the best mean 100-episode reward (Atari) or the best mean reward (DMC). h: horizontal; hv: horizontal+vertical. Seeds are listed as VEN/DDQN. t: Welch's t-test on per-seed (Atari) or per-run (DMC) best rewards, positive favouring VEN (df 2–13); with three Atari seeds only FishingDerby, RoadRunner, and VideoPinball reach individual significance.

Domain	Environment	Setting	VEN (mean±sd)	DDQN (mean±sd)	VEN win by	Seeds	t
Atari-2600	Boxing	h	$+\mathbf{30.5 \pm 4.4}$	$+17.9 \pm 16.8$	$+\mathbf{12.6}$	3/3	1.3
Atari-2600	Breakout	hv	$+\mathbf{38.8 \pm 1.0}$	$+37.5 \pm 0.4$	$+\mathbf{1.3}$	3/3	2.0
Atari-2600	FishingDerby	h	$+\mathbf{0.9 \pm 6.6}$	-62.4 ± 15.7	$+\mathbf{63.4}$	3/3	6.4
Atari-2600	KungFuMaster	h	$+\mathbf{6646 \pm 6491}$	$+4889 \pm 914$	$+\mathbf{1757}$	3/3	0.5
Atari-2600	RoadRunner	h	$+\mathbf{24077 \pm 2094}$	$+5193 \pm 2255$	$+\mathbf{18884}$	3/3	10.6
Atari-2600	Tennis	h	$-\mathbf{10.7 \pm 3.0}$	-13.9 ± 4.4	$+\mathbf{3.2}$	3/3	1.0
Atari-2600	VideoPinball	h	$+\mathbf{58295 \pm 5865}$	$+14108 \pm 1935$	$+\mathbf{44187}$	3/3	12.4
DMC	CartPole	hv	$+\mathbf{252.3 \pm 35.6}$	$+46.9 \pm 6.3$	$+\mathbf{205.4}$	7/6	15.0
DMC	Inv. Double Pendulum	hv	$+\mathbf{909.5 \pm 235.5}$	$+509.5 \pm 132.5$	$+\mathbf{400.0}$	7/7	3.9
DMC	Inverted Pendulum	hv	$+\mathbf{204.0 \pm 99.4}$	$+66.9 \pm 57.0$	$+\mathbf{137.1}$	9/9	3.6
DMC	MountainCar	hv	$-\mathbf{146.6 \pm 29.5}$	-188.6 ± 10.9	$+\mathbf{41.9}$	6/6	3.3
DMC	Reacher	hv	$-\mathbf{7.6 \pm 0.2}$	-10.3 ± 1.0	$+\mathbf{2.7}$	5/5	6.0

5.5 Continual learning

We trained on one Pong flip configuration, then retrained on a mixture in which 50% of episodes kept the old configuration. Both architectures saw the identical mixture. With 50% old-task replay, VEN retained the old task at $+20.9 \pm 0.3$ (920 evaluation episodes; 3 seeds, 4 transitions) while DDQN fell to -3.8 ± 12.8; the gap ranges from $+15.9$ to $+30.4$ across cells. VEN also acquired the new task: on the configurations that the phase-2 mixture actually trains, it reached $+8.0 \pm 5.6$ (best cell $+10.8 \pm 0.6$, the h-flip task after no-flip pretraining), while DDQN reached -2.1 ± 11.5.

Configurations presented in neither session score the random-play floor (-21.0) for both architectures, so the new-task scores reflect learning rather than leakage. In the binarized protocol, VEN acquired the new h-flip variant to $+18.2 \pm 3.9$ while holding the no-flip source task at $+17.9 \pm 1.6$; DDQN reached $+6.4 \pm 2.0$ on the new task (Figure 4). With 0% replay both architectures forget the old task (about -21) while still learning the new one, and VEN learns it faster and more reliably: pooled over the single-target transitions, $+18.8 \pm 3.3$ versus $+4.5 \pm 13.1$, and VEN solves all four single-target transitions ($+13.9$ to $+21.0$) where DDQN fails the no-flip→v-flip transition (-14.6). The protection of the old task therefore requires rehearsal; the acquisition of the new task does not (Table 4).

Table 4: Continual learning: old-task retention and new-task acquisition. Mean±s.d. over evaluation episodes; both architectures saw the identical mixture (3 seeds, 4 transitions).

Agent	Replay	Old task	New task	Untrained
VEN	50%	$+\mathbf{20.9 \pm 0.3}$	$+\mathbf{8.0 \pm 5.6}$	-21.0 ± 0.0
DDQN	50%	-3.8 ± 12.8	-2.1 ± 11.5	-21.0 ± 0.0
VEN	0%	-20.6 ± 1.4	$+\mathbf{18.8 \pm 3.3}$	-20.9 ± 0.3
DDQN	0%	-20.9 ± 0.2	$+4.5 \pm 13.1$	-21.0 ± 0.1

Old task: phase-1 configuration. New task at 50%: all configurations trained by the phase-2 mixtures but absent from phase 1 ($n_{episodes}$=1,848 per architecture). New task at 0%: the target configuration of the single-target transitions (no-flip→h, no-flip→v, h→v, h→hv; n=888); the mixture transitions at 0% replay learn only their designated orientation and are excluded. Untrained: configurations presented in neither session, the random-play floor. The 0% regime used an earlier retraining schedule.

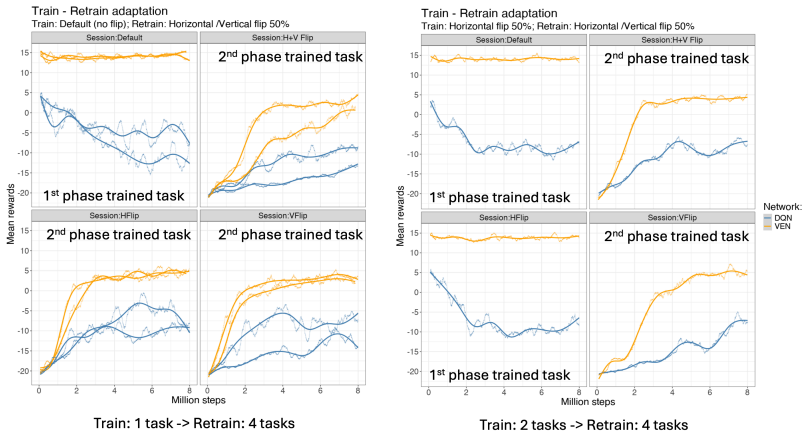

Figure 4: Performance of VEN and capacity-matched DDQN over the second training phase. Models were trained in the first phase on old tasks, and in the second phase on a mixture of old and new tasks. Curves are the mean reward over 100 episodes at the indicated step.

Interventions localize the mechanism. Through retraining, VEN's encoder keeps 3–5× more phase-1 representational overlap than DDQN's (CKA). Zeroing $\hat{z}$ in the trained VEN head changes Q by 0.6–0.8 and flips the chosen action in 50–75% of decisions. Decomposing the temporal pathway by intervention (A: no history; F: true-embedding history only; E: full pathway), the history carries essentially all of the advantage in distribution (F $+20.4$ versus E $+21.0$; the prediction's share is 0–1%), but under partial transfer to an unseen flip the prediction's share rises to 44% of recoverable reward (A -20.7, F -3.1, E $+10.5$): when history is unreliable, the head leans on the efference copy.

5.6 World model encompasses VEN function

DreamerV3 already runs an action-conditioned forward model in its RSSM (Hafner et al., 2025). Adding VEN changed nothing beyond single-run spread in 39 runs across four game families (Table 5), as expected once the backbone already predicts its own actions. On KungFuMaster the pooled mean favors noVEN by about 2×, but a single 793K run drives it: the medians are 94K (VEN) and 91K (noVEN), and the run-to-run s.d. (212K) spans the difference. On the other three families the

means overlap. On Pong, DDQN+VEN still matches the DreamerV3 band (best $+15.6$ to $+20.3$ versus 21) with a 302,454-parameter agent, of which VEN is 8,704 (Table 1).

Table 5: DreamerV3 negative control: best episode reward, mean$\pm$s.d. over runs pooled across flip variants (n runs in parentheses; one seed per cell). Three Pong runs labelled VEN were launched with VEN disabled and are counted under noVEN. VEN never exceeds noVEN beyond single-run spread.

Game family	VEN	noVEN
Pong	21.0 ± 0.0 (3)	20.8 ± 0.4 (5)
KungFuMaster	$84{,}540 \pm 20{,}485$ (5)	$165{,}376 \pm 212{,}201$ (10)
VideoPinball	$174{,}888 \pm 67{,}098$ (3)	$109{,}884 \pm 13{,}430$ (5)
Catcher (L40)	11 ± 4 (4)	14 ± 2 (4)

6 Conclusions and limitations

Self-recognition in a mirror and continual learning in a network may share a computational requirement: both need an online separation between input changes the agent caused and input changes the world imposed. Von Economo neurons, present in the mirror-recognition capable species, have a morphology that suggests exactly the comparator such a separation needs. We built the comparator explicitly as a stop-gradient action-conditioned latent forward model whose prediction enters the value head, and tested it where the difference between self and world matters: under unannounced mirror flips, under within-episode action reversal, and across sequential task changes.

The experiments support our statements. First, the effect is causal and specific: performance under perturbation tracks the presence of a valid efference copy (shuffling it abolishes the advantage) and cannot be reproduced by adding parameters, history features, or auxiliary prediction losses. Second, the circuit computes what it was designed to compute: latents become action-dependent and configuration-decodable, and the trained Q head reads the prediction, especially under distribution shift. Third, the continual-learning benefit is real but conditional: +24.7 points of old-task retention with rehearsal, but nothing on top of a backbone that already models its own actions.

Our study is not without limitations. The biological reading remains a hypothesis; we have shown that a VEN-inspired circuit is sufficient for the computational function, not that the neurons implement it. What the results do establish is a minimal, cheap (8.7K-parameter), and reproducible circuit that turns "what did my own action just cause" into a value feature, with measurable consequences for robustness and retention, and a concrete experimental handle on why self-attribution and persistent learning might share neural machinery.

Reproducibility statement

The single-file trainer behind Table 2 (all seven architectures and the input ablations), the deterministic per-flip evaluation script, and five reference checkpoints (VEN and DDQN trained with and without h-flip, plus the 10M continual-learning retrain) are available at `https://anonymous.4open.science/r/VEN_Pong_Minimal_Reproducibility-B853` (anonymized for review). Every number in the tables regenerates from the released scripts and the archived run logs.

References

Rick A. Adams, Stewart Shipp, and Karl J. Friston. Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218(3):611–643, 2013. doi: 10.1007/s00429-012-0475-5.

John M. Allman, Karli K. Watson, Nicole A. Tetreault, and Atiya Y. Hakeem. Intuition and autism: A possible role for von Economo neurons. *Trends in Cognitive Sciences*, 9(8):367–373, 2005. doi: 10.1016/j.tics.2005.06.008.

John M. Allman, Nicole A. Tetreault, Atiya Y. Hakeem, Kebreten F. Manaye, Katerina Semendeferi, Joseph M. Erwin, Soyoung Park, Virginie Goubert, and Patrick R. Hof. The von Economo neurons in frontoinsular and anterior cingulate cortex in great apes and humans. *Brain Structure and Function*, 214:495–517, 2010. doi: 10.1007/s00429-010-0254-0.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Curtis C. Bell. An efference copy which is modified by reafferent input. *Science*, 214(4520):450–453, 1981. doi: 10.1126/science.7291987.

Sarah-Jayne Blakemore, Daniel M. Wolpert, and Chris D. Frith. Why can't you tickle yourself? *NeuroReport*, 11(11):R11–R16, 2000. doi: 10.1097/00001756-200008030-00002.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1lJJnR5Ym`.

Camilla Butti, Micaela Santos, Neha Uppal, and Patrick R. Hof. Von Economo neurons: Clinical and evolutionary perspectives. *Cortex*, 49(1):312–326, 2013. doi: 10.1016/j.cortex.2011.10.004.

Liangtang Chang, Qin Fang, Shikun Zhang, Mu-ming Poo, and Neng Gong. Mirror-induced self-directed behaviors in rhesus monkeys after visual-somatosensory training. *Current Biology*, 25 (2):212–217, 2015. doi: 10.1016/j.cub.2014.11.016.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. doi: 10.1109/CVPR46437.2021.01549.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. doi: 10.1017/S0140525X12000477.

A. D. Craig. How do you feel—now? the anterior insula and human awareness. *Nature Reviews Neuroscience*, 10(1):59–70, 2009. doi: 10.1038/nrn2555.

Trinity B. Crapse and Marc A. Sommer. Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600, 2008. doi: 10.1038/nrn2457.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Henry C. Evrard, Thomas Forro, and Nikos K. Logothetis. Von Economo neurons in the anterior insula of the macaque monkey. *Neuron*, 74(3):482–493, 2012. doi: 10.1016/j.neuron.2012.03.023.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. doi: 10.1016/S1364-6613(99)01294-2.

Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360 (1456):815–836, 2005. doi: 10.1098/rstb.2005.1622.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010. doi: 10.1038/nrn2787.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1–3):70–87, 2006. doi: 10.1016/j.jphysparis.2006.10.001.

Karl J. Friston, Jean Daunizeau, James Kilner, and Stefan J. Kiebel. Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102(3):227–260, 2010. doi: 10.1007/s00422-010-0364-z.

Shaun Gallagher. Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4(1):14–21, 2000. doi: 10.1016/S1364-6613(99)01417-5.

Gordon G. Gallup. Chimpanzees: Self-recognition. *Science*, 167(3914):86–87, 1970. doi: 10.1126/science.167.3914.86.

Jianqiao Ge, Xiaosi Gu, Meng Ji, and Shihui Han. Neurocognitive processes of the religious leader in Christians. *Human Brain Mapping*, 30(12):4012–4024, 2009. doi: 10.1002/hbm.20825.

Alexander N. Gorban and Ivan Y. Tyukin. Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A*, 376(2118): 20170237, 2018. doi: 10.1098/rsta.2017.0237.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. URL `https://arxiv.org/abs/1812.04754`.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Atiya Y. Hakeem, Chet C. Sherwood, Christopher J. Bonar, Camilla Butti, Patrick R. Hof, and John M. Allman. Von Economo neurons in the elephant brain. *The Anatomical Record*, 292(2): 242–248, 2009. doi: 10.1002/ar.20829.

Shihui Han, Lihua Mao, Xiaosi Gu, Ying Zhu, Jianqiao Ge, and Yina Ma. Neural consequences of religious belief on self-referential processing. *Social Neuroscience*, 3(1):1–15, 2008. doi: 10.1080/17470910701469681.

Shihui Han, Xiaosi Gu, Lihua Mao, Jianqiao Ge, Gang Wang, and Yina Ma. Neural substrates of self-referential processing in Chinese Buddhists. *Social Cognitive and Affective Neuroscience*, 5 (2–3):332–339, 2010. doi: 10.1093/scan/nsp027.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=SJ6yPD5xg`.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Policy consolidation for continual reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3218–3228, 2019.

Mitsuo Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9(6):718–727, 1999. doi: 10.1016/S0959-4388(99)00028-8.

Wei Ke, Shuxuan Lyu, Mengmeng Jin, Shaoqing Jiao, Liang Li, Heyuan Zhang, Kexin Ling, Cuiping Tian, Yiquan Song, Liu Chen, Quansheng He, Yujie Xiao, Suixin Deng, Liu Fan, Huawei Mu, Hui Yang, Hua-Tai Xu, Hao Wang, Ai-Hui Tang, Jiajie Peng, Jie He, Hui Guo, and Yousheng Shu. Spindle neurons in human cortex possess distinctive firing properties and transcriptomic signatures. *Nature Communications*, 17(1):6330, 2026. doi: 10.1038/s41467-026-72935-2.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Machine Learning Research*, 23(290):1–44, 2022. URL `https://www.jmlr.org/papers/v23/21-1232.html`.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Christof Koch. *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, New York, 1999.

Yann LeCun. A path towards autonomous machine intelligence, 2022. URL `https://openreview.net/forum?id=BZ5a1r-kVsf`. OpenReview technical report, version 0.9.2.

Rui-Feng Liu, Mengyao Huang, Yuhui Shen, Mingting Shao, Junzhan Jing, Nana Xu, Lei Tang, Biaodi Liu, Jianming Shi, Fanrui Chen, Zhao-Zhe Hao, Xiaolong Jiang, and Sheng Liu. An atlas of primate insular cortex reveals a signal-processing strategy in von Economo neurons. *Nature Cell Biology*, 28:1742–1757, 2026. doi: 10.1038/s41556-026-02009-4.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989. doi: 10.1016/S0079-7421(08)60536-8.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. doi: 10.1038/nature14236.

Esther A. Nimchinsky, Brent A. Vogt, John H. Morrison, and Patrick R. Hof. Spindle neurons of the human anterior cingulate cortex. *Journal of Comparative Neurology*, 355(1):27–37, 1995. doi: 10.1002/cne.903550106.

Esther A. Nimchinsky, Emmanuel Gilissen, John M. Allman, Daniel P. Perl, Joseph M. Erwin, and Patrick R. Hof. A neuronal morphologic type unique to humans and great apes. *Proceedings of the National Academy of Sciences*, 96(9):5268–5273, 1999. doi: 10.1073/pnas.96.9.5268.

Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, and Satinder Singh. Action-conditional video prediction using deep networks in Atari games. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787, 2017.

Joshua M. Plotnik, Frans B. M. de Waal, and Diana Reiss. Self-recognition in an Asian elephant. *Proceedings of the National Academy of Sciences*, 103(45):17053–17057, 2006. doi: 10.1073/pnas.0608062103.

James F. A. Poulet and Berthold Hedwig. A corollary discharge maintains auditory sensitivity during sound production. *Nature*, 418:872–876, 2002. doi: 10.1038/nature00919.

Helmut Prior, Ariane Schwarz, and Onur Güntürkün. Mirror-induced behavior in the magpie (Pica pica): Evidence of self-recognition. *PLoS Biology*, 6(8):e202, 2008. doi: 10.1371/journal.pbio.0060202.

Vinay V. Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=LhY8QdUGSuw`.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.

Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308, 1990. doi: 10.1037/0033-295X.97.2.285.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017. doi: 10.1109/CVPR.2017.587.

Diana Reiss and Lori Marino. Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Proceedings of the National Academy of Sciences*, 98(10):5937–5942, 2001. doi: 10.1073/pnas.101086398.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. URL `https://arxiv.org/abs/1606.04671`.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pp. 222–227. MIT Press, 1991.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537, 2018.

William W. Seeley, Vinod Menon, Alan F. Schatzberg, Jennifer Keller, Gary H. Glover, Heather Kenna, Allan L. Reiss, and Michael D. Greicius. Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9):2349–2356, 2007. doi: 10.1523/JNEUROSCI.5587-06.2007.

Roger W. Sperry. Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43(6):482–489, 1950. doi: 10.1037/h0055479.

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015. URL `https://arxiv.org/abs/1507.00814`.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571:361–365, 2019. doi: 10.1038/s41586-019-1346-5.

Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pp. 216–224. Morgan Kaufmann, 1990.

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 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Kai Ueltzhöffer. Deep active inference. *Biological Cybernetics*, 112(6):547–573, 2018. doi: 10.1007/s00422-018-0785-7.

Elise van der Pol, Daniel E. Worrall, Herke van Hoof, Frans A. Oliehoek, and Max Welling. MDP homomorphic networks: Group symmetries in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 2094–2100, 2016. doi: 10.1609/aaai.v30i1.10295.

Constantin von Economo. Eine neue art von ganglienzellen der grosshirnrinde. *Zeitschrift für die gesamte Neurologie und Psychiatrie*, 100:438–448, 1926. doi: 10.1007/BF02871127.

Erich von Holst and Horst Mittelstaedt. Das Reafferenzprinzip. *Naturwissenschaften*, 37:464–476, 1950. doi: 10.1007/BF00622503.

Daniel M. Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature Neuroscience*, 3:1212–1217, 2000. doi: 10.1038/81497.

Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995. doi: 10.1126/science.7569931.

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=GY6-6sTvGaf`.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, 2017.