Geometry of abstract learned knowledge in deep RL agents

James Mochizuki-Freeman JMOCHIZU@IU.EDU
Md Rysul Kabir MDRKABIR@IU.EDU
Mitesh Gulecha MGULECHA@IU.EDU
Zoran Tiganj ZTIGANJ@IU.EDU
Department of Computer Science, Indiana University Bloomington

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Nina Miolane

Abstract

Data from neural recordings suggest that mammalian brains represent physical and abstract task-relevant variables through low-dimensional neural manifolds. In a recent electrophysiological study (Nieh et al., 2021), mice performed an evidence accumulation task while moving along a virtual track. Nonlinear dimensionality reduction of the population activity revealed that task-relevant variables were jointly mapped in an orderly manner in the low-dimensional space. Here we trained deep reinforcement learning (RL) agents on the same evidence accumulation task and found that their neural activity can be described with a low-dimensional manifold spanned by task-relevant variables. These results provide further insight into similarities and differences between neural dynamics in mammals and deep RL agents. Furthermore, we showed that manifold learning can be used to characterize the representational space of the RL agents with the potential to improve the interpretability of decision-making in RL.

Keywords: Evidence accumulation, Low-dimensional embedding, Manifold learning, Population analysis

1. Introduction

A number of neuroscience studies have proposed that the hippocampus encodes a cognitive map as a low-dimensional manifold spanned by task-relevant variables (Nieh et al., 2021; Keefe and Nadel, 1978; Low et al., 2018; Aronov et al., 2017; Knudsen and Wallis, 2021). At any moment, the activity of a population of neurons can be characterized as a point in a high-dimensional space where each dimension represents the firing rate of a single neuron. For low-dimensional manifold representation to emerge, the network properties need to confine the high-dimensional neural trajectories into a low-dimensional subspace.

A recent study by Nieh et al. (2021) demonstrated that neurons in the hippocampus integrate neural representations of cognitive and physical variables, forming a low-dimensional manifold. Nieh et al. (2021) trained mice on the "accumulating towers task" that combines navigation with decision-making such that mice had to accumulate evidence as they navigated through the environment. As mice moved along a virtual track, objects (referred to as "towers") appeared on both sides of the track. When they arrived at the end of the track, to earn a reward, mice had to choose the left- or right-hand side, depending on which side had more towers (see also Morcos and Harvey (2016), Pinto et al. (2018) and Engelhard et al. (2019) all of which used the same experimental paradigm). The difference in the number of towers is an abstract latent variable that corresponds to the amount of

evidence for either of the two options, connecting this task with a large body of literature in cognitive science and neuroscience that aims to advance the understanding of evidence accumulation and decision-making (see, e.g., Brody and Hanks (2016); Gold and Shadlen (2007); Dayan and Daw (2008)). Nieh et al. (2021) recorded the activity of hundreds of individual neurons from the dorsal CA1 sub-region of mice hippocampus while they performed the accumulating towers task. The results indicated the existence of cells tuned to a particular difference in the number of towers, such that a population of neurons tiled the entire *evidence* axis. Furthermore, nonlinear dimensionality reduction suggested that the activity was constrained to a task-specific low-dimensional manifold that contained a geometric representation of learned knowledge. When projected into 3D space, physical and abstract variables were jointly mapped in an orderly manner such that the position of the animal along the track and the amount of evidence appeared organized as gradients. The trajectory of the neural state typically progressed along a position direction over the course of a trial, while splitting along the evidence direction. This result provided valuable insight into how biological neural networks represent abstract and physical variables.

Deep RL agents have shown remarkable success in learning a variety of tasks at a level similar to or exceeding humans (Mnih et al., 2013, 2015; Silver et al., 2017). A number of studies have compared neural activity in artificial agents to the neural activity in mammalian brains, finding remarkable similarities (Banino et al., 2018; Lin and Richards, 2021; Deverett et al., 2019). Furthermore, a substantial effort has been invested in understanding neural representations in deep RL agents (Greydanus et al., 2018; Iyer et al., 2018; Heuillet et al., 2021; Wells and Bednarz, 2021; Mott et al., 2019). A better understanding of how these agents represent information and make decisions can help advance both AI and cognitive neuroscience (Bengio et al., 2021; Hassabis et al., 2017; Ullman, 2019).

Inspired by neuroscience research, here we investigated whether deep RL agents represent task-relevant variables in the form of low-dimensional manifolds. We trained agents on the accumulating towers task (similar to Lee et al. (2022) and using a more realistic environment and more complex agents than in our previous work (Mochizuki-Freeman et al., 2023)) and applied neuroscience analysis methods to the neural activity in artificial agents. Specifically, we examined the activity of neurons as a function of evidence, position, and luminance and applied the same dimensionality reduction technique as Nieh et al. (2021) called MIND (Low et al., 2018). MIND constructs a set of latent variables with a specific emphasis on incorporating temporal dynamics, and it has been argued to be particularly suited to finding low-dimensional representations in data with sequential activity. We used linear and non-linear decoding techniques to compare how task-relevant (evidence and position) and task-irrelevant (luminance) variables were represented in the original high-dimensional space and in the low-dimensional space learned with the MIND algorithm.

2. Methods

2.1. Accumulating towers task

We designed the accumulating towers task environment within the PyBullet physics engine (Coumans and Bai, 2021) based on the virtual reality environment used in mice recordings described in Nieh et al. (2021) and Pinto et al. (2018). The environment receives actions (*forward*, *left*, or *right*) and outputs pixel observations (Fig. 1). It has two 10.5 cm-wide

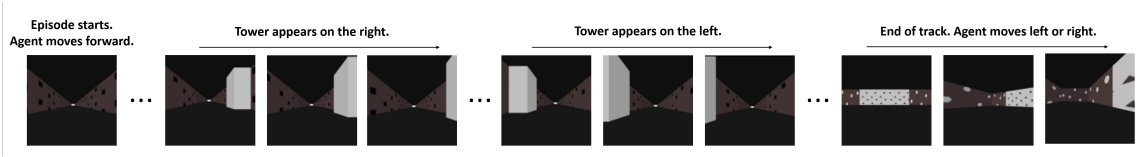

Figure 1: The accumulating towers task. In each trial, the agent moves down a narrow corridor and observes "towers" (white objects) on the left- and right-hand sides of the wall. To obtain the reward, the agent needs to turn left or right at the end of the corridor, depending on which side has more towers. Similar to the task performed by mice, the towers only become visible as the agent approaches them and then disappear shortly after. The walls of the environment have a textured pattern to provide optic flow to the agent.

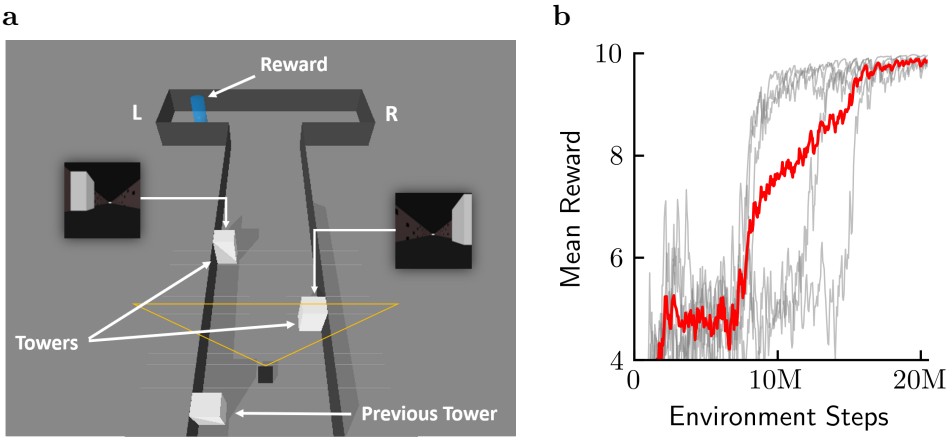

Figure 2: **a** Schematic of the accumulating towers environment showing two inserts that illustrate the agent's visual input. **b** Agent performance on the accumulating towers task. Each gray line represents the performance of a single agent over the course of training. The red line represents the mean performance of the 5 agents. The maximum mean reward was 10.

arms at the end of a 200cm long straight track (Fig. 2). The agent must go through a confined path that is bounded by walls that are 10 cm apart. At the end of the track, the agent selects one of the two arms. Different numbers of towers are placed along each of the two sides with a minimum gap of 7 cm between any two towers. At each trial, the number of towers on each side is chosen randomly, between 1 and 15. The towers only appear when they are within 5 cm of the agent and disappear as the agent passes them. The towers have a 1 cm width, while the entire track has a 6 cm height. The *forward* action moves the agent down the track at a fixed speed of 1cm/step. The *left* and *right* actions change the agent's alignment in 7.5° increments to a max of ±15° along the straight portion of the maze and without a bound in the arms. Once the agent reaches the end of the track, it receives a

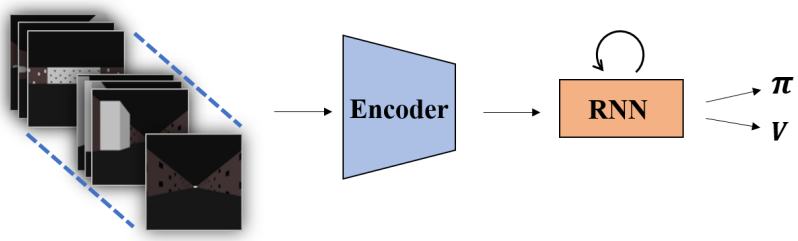

Figure 3: Agent architecture. Pixel inputs are fed into a CNN encoder followed by a GRU-based recurrent layer. The output of the recurrent layer is fed into an A2C-based RL agent.

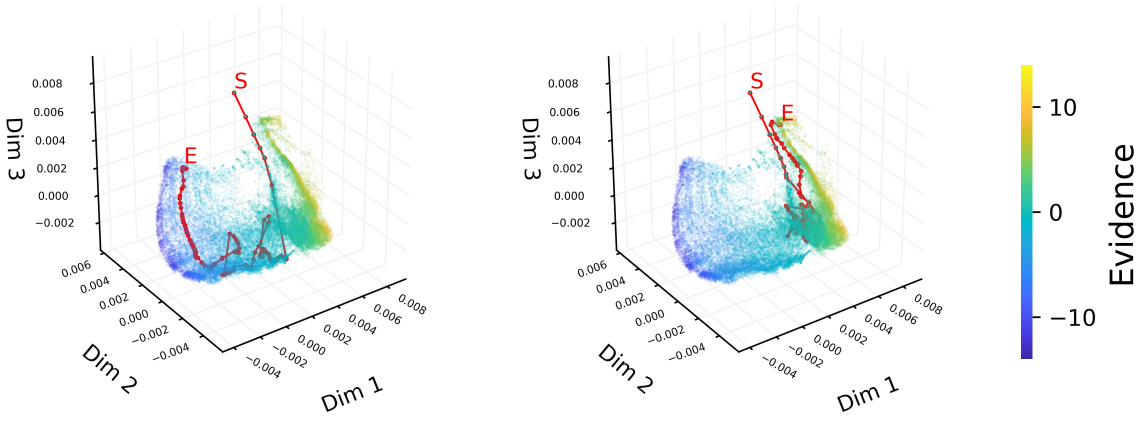

Figure 4: Trajectory of a left- and right-choice trial in MIND embedding space of the recurrent layer for one of the agents. The start and end of the trajectory are marked with S and E.

reward of either 10 if it made a correct choice or 0 otherwise. It also receives a reward of -0.1 for attempting to turn beyond 15° in either direction when in the main track or for making contact with the back wall.

2.2. Agent architecture and training

We constructed deep RL agents based on the A2C architecture (Mnih et al., 2016; Sutton and Barto, 2018) (Fig. 3). The agents consisted of three main modules: a convolutional encoder to reduce an input image to a latent vector, a recurrent network to encode a series of latent vectors over time, and actor and critic networks to generate action and value predictions.

At each time step, the agents received as input 60×60 grayscale pixel values from the environment. Those images were fed into three successive convolutional layers: 64 kernels of size 8×8 (stride 2), 32 kernels of size 2×2 (stride 1), and 64 kernels of size 3×3 (stride

2). This reduced the input to tensors of size $27 \times 27 \times 64$, then $26 \times 26 \times 32$, and then $12 \times 12 \times 64$, respectively. The resulting tensor was flattened into a 9216-element vector and passed through a fully connected layer, which reduced the 9216 intermediate units to 180 output features. All layers in the encoder had ReLU activation functions.

The convolutional encoder was followed by a GRU layer consisting of 180 hidden units. The output of the recurrent layer was passed to an actor network and a critic network, which respectively produced action logits and state value estimates. The actor and the critic network each consisted of two-layers: a fully connected layer of 180 neurons followed by a fully connected output layer of 3 (actor head) or 1 (critic head) neurons.

During training, we explored five different learning rates (0.001, 0.0005, 0.0001, 0.00005, and 0.00001) for the Adam optimizer, and three different entropy bonus coefficients (0.0, 0.001, and 0.0001) to encourage exploration by the agents. We ran each agent configuration five times for 20 million environment steps. For each training step, we fed into the agent 5120 total environment steps, collected from 20 environments running in parallel for 256 time steps each, and then performed a backpropagation pass and stepped the optimizer. Agents with a learning rate of 0.00005 and an entropy coefficient of 0.0 obtained the highest average reward, and they were used in the subsequent analyses.

2.3. Dimensionality reduction

We applied a dimensionality reduction technique called Manifold Inference from Neural Dynamics (MIND), which was introduced in Low et al. (2018). MIND takes as input a high-dimensional neural representation where each neuron corresponds to a unique coordinate axis. In our case, the input was the activity of the 180 recurrent neurons. MIND aims to describe the population activity using a small number of coordinates under the assumption that the high-dimensional space is intrinsically low-dimensional.

To perform the dimensionality reduction, MIND learns transition probabilities between states using an algorithm based on a randomized ensemble of decision trees. Each tree adaptively partitions the neural state space into local neighborhoods, with a continuous probability distribution over future states for each neighborhood. After combining estimates across trees, transition probabilities are used to construct a distance metric between states, which defines their proximity on the manifold. Distances have the property that they are smaller for pairs of states that transitioned to each other with higher probability through a smaller number of intermediate states. Intrinsic manifold coordinates representing the network trajectory are obtained by embedding states into a low-dimensional Euclidean space such that distances are approximately preserved.

3. Results

3.1. Performance of deep RL agents

We trained and evaluated deep RL agents on the accumulating towers task. After 20 million environment steps, agents were able to reach near-perfect performance (Fig. 2b). This is further illustrated using psychometric curves and confusion matrices that show performance before and after the training (Fig. S1 and Fig. S2). The y-axis on psychometric curves represents the probability of turning right. For an agent that performs perfectly, that

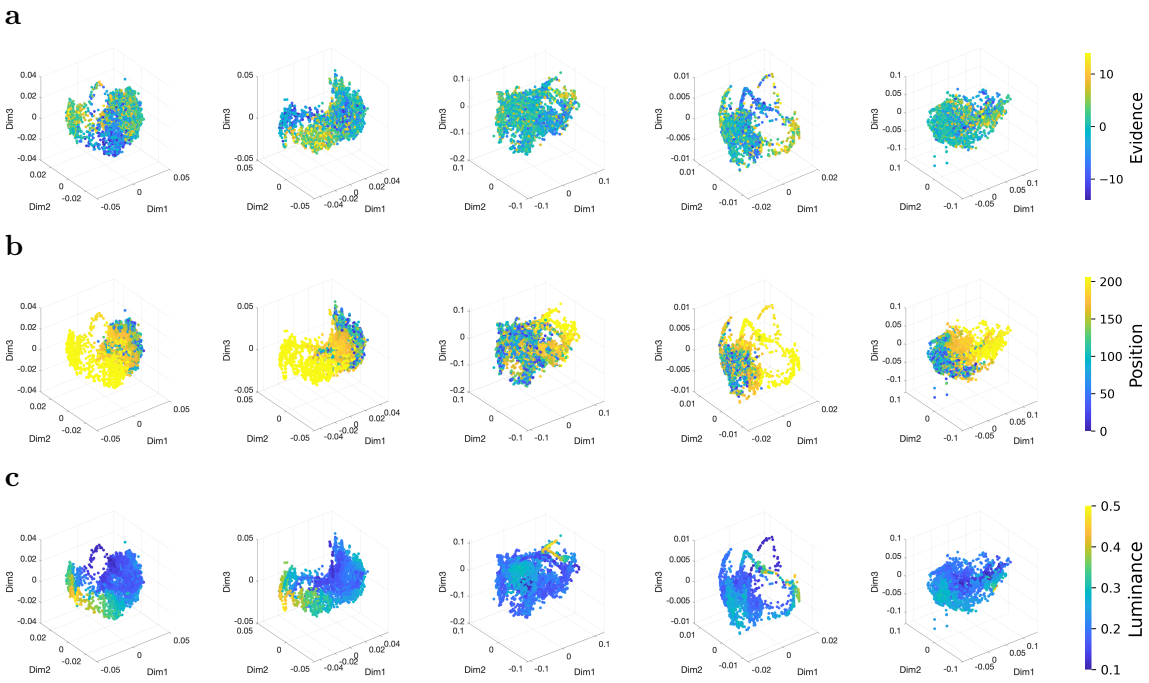

Figure 5: Three-dimensional embedding of the neural activity in the encoder layer for each of the five deep RL agents. The embedding was computed using MIND and it is color-coded by **(a)** amount of evidence, **(b)** position, and **(c)** luminance.

probability would be 0 for the trials with a negative amount of evidence (*i.e.*, more towers on the left-hand side) and 1 for trials with a positive amount of evidence (*i.e.*, more towers on the right-hand side). Similar to animal behavior, when agents made mistakes, they were made in the most difficult trials when the evidence was close to 0. Confusion matrices confirmed that the mistakes were not specific to a particular number of towers on either side.

3.2. Geometry of the neural manifold

For each of the five agents, we constructed a 7-dimensional embedding using MIND. The embedding was constructed for the 180 neurons from the encoder layer and for the 180 neurons from the recurrent layer. Fig. 5 and Fig. 6 show the first three dimensions of the 7-dimensional embedding for the encoder and recurrent layer, respectively, with the color coding indicating evidence, position, and luminance. For each agent, we used 50000 time steps, and each embedding was constructed using 37500 time steps per agent. The remaining 12500 points were held out for testing the properties of the manifold (we used a 75/25 train/test split).

To quantify how well the 7-dimensional manifold captured the variability of the 180-dimensional input, we computed a reconstruction index and reconstruction error (Fig. 7). The reconstruction index was computed as the correlation coefficient between the predicted

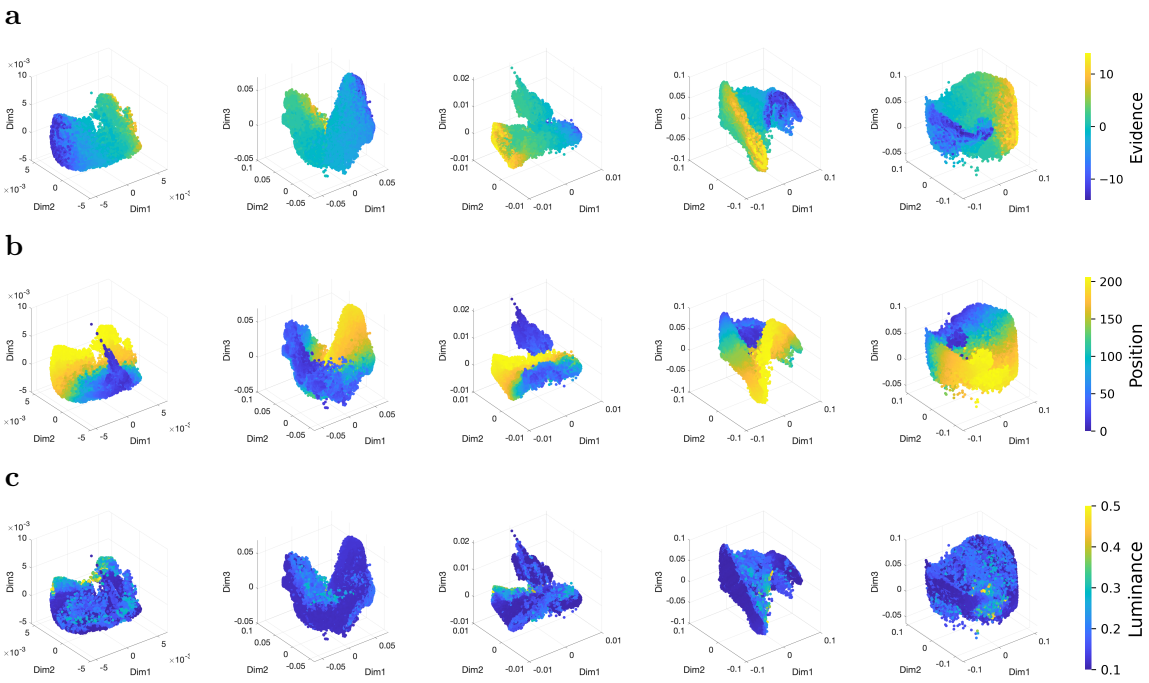

Figure 6: Three-dimensional embedding of the neural activity in the recurrent layer for each of the five deep RL agents. The embedding was computed using MIND and it is color-coded by **(a)** amount of evidence, **(b)** position, and **(c)** luminance.

and true neural activity in held-out data. The reconstruction error was the Mean Squared Error (MSE) between the predicted and true neural activity in held-out data. The results indicated that adding additional dimensions after the first 2 to 3 dimensions led to only minor improvement. This is consistent with the neuroscience data and suggests that the neural activity in this task can be well captured using a low-dimensional manifold.

Visual inspection of the 3-dimensional space constructed for the recurrent layer (Fig. 6) reveals that position and evidence gradually changed along the first two dimensions. This ordinal organization and gradients that follow evidence and position were also reported in Nieh et al. (2021). The position has a gradient from the beginning of the track to the end. When the agent is around the start of the track, the evidence is around 0 and increases, either in a positive or negative direction, as the agent moves toward the end. This is captured by dark blue and bright yellow colors ending up on the opposite parts of the manifold. Importantly, the two gradients point in directions that appear perpendicular and consistent with the biological data. The trajectory progresses along the position from low (blue) to high (yellow) and then splits up or down depending on the amount of evidence (see Fig. 4 for illustrations of the trajectories in the 3-dimensional space).

Further visual inspection of Fig. 5 and Fig. 6 suggests that evidence (task-relevant variable) was part of low-dimensional embedding in the recurrent layer but not in the encoder layer. On the other hand, luminance (not task-relevant variable) was part of low-dimensional embedding in the encoder layer but not in the recurrent layer.

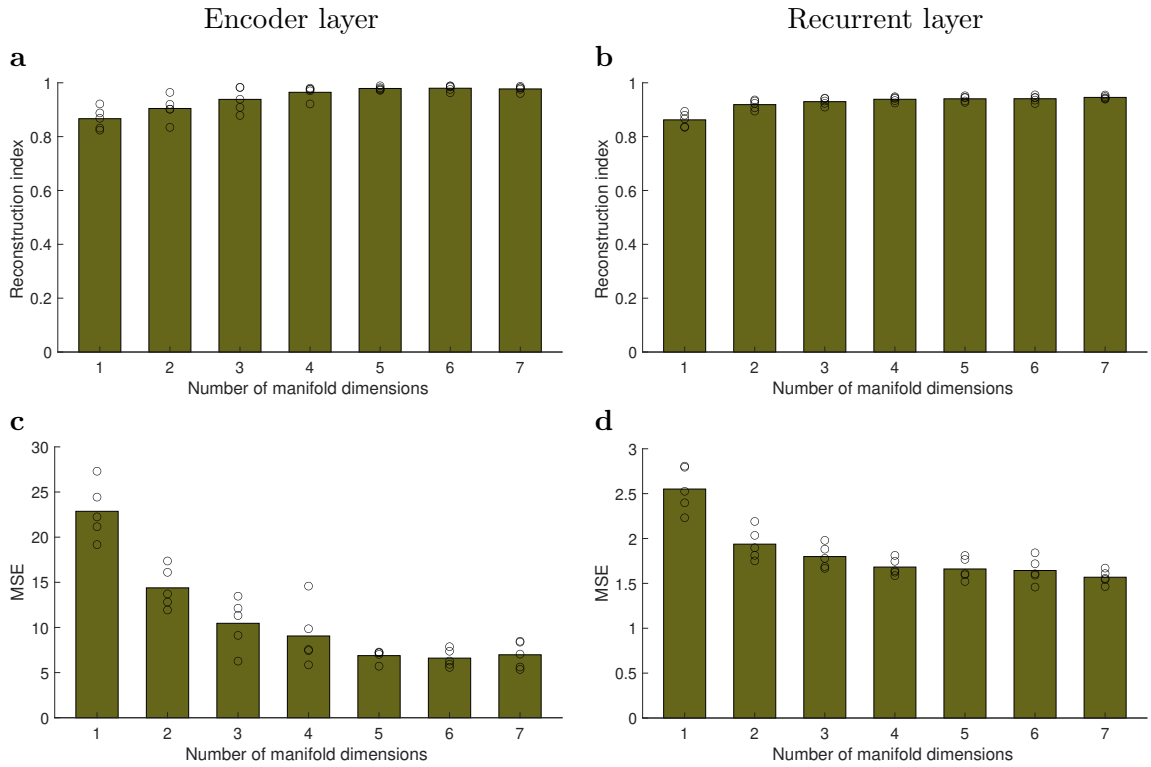

Figure 7: Evaluation of low-dimensional MIND embeddings using **(a,b)** reconstruction index, and **(c,d)** Mean Squared Error. Circles represent results from individual agents, and bars provide the mean value.

In addition to MIND, we also performed dimensionality reduction using t-SNE (Van der Maaten and Hinton, 2008) (Fig. S3 and Fig. S4), but unlike MIND, t-SNE did not reveal a structured representation of task-relevant variables.

3.3. Reconstruction from embeddings

To quantify the results of the above visual inspection, we computed the decoding index from d-dimensional embeddings of the manifold using Gaussian Process Regression (GPR) and linear regression (Fig. 8). The decoding index is the correlation coefficient between the predicted and true values of evidence, position, and luminance computed using the held-out portion of the dataset. The decoding results confirmed that the low-dimensional embedding of the encoder layer contained information primarily about luminance and position, with little information about evidence (Fig. 8a,c) – this is expected since the encoder has no memory that could accumulate evidence. In contrast, the low-dimensional embedding of the recurrent layer contained information primarily about position and evidence with much less information about luminance (Fig. 8b,d), indicating that the low-dimensional manifold was spanned by the task-relevant variables.

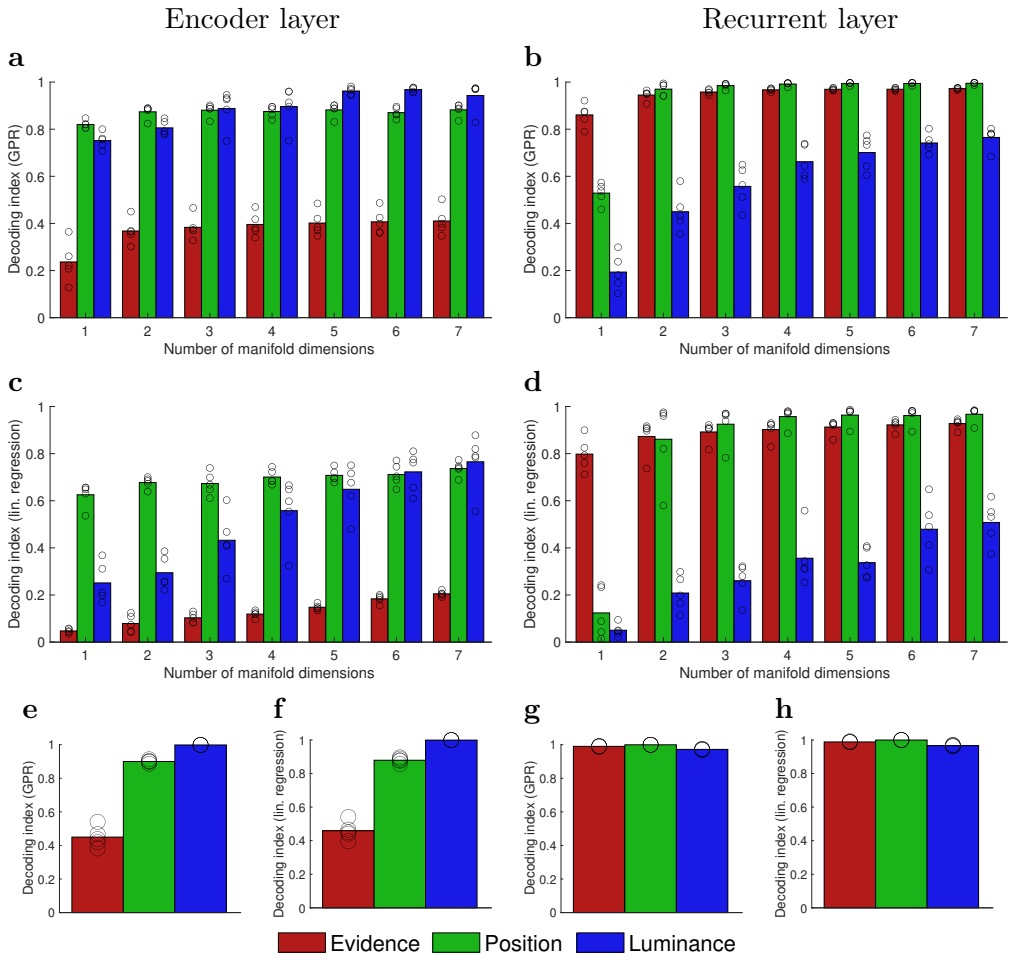

Figure 8: Decoding of evidence, position, and luminance using GPR and linear regression from low-dimensional MIND embeddings (panels **a** - **d**) and from original 180-dimensional space (panels **e** - **h**). Circles represent results from individual agents, and bars provide the mean value.

To showcase that these findings were critically dependent on the ability to learn the low-dimensional embedding of the neural representation, we performed decoding in the original 180-dimensional space (Fig. 8e-h). In the encoder layer, information decoded from the 180-dimensional space was similar to the information decoded from the low-dimensional space, with position and luminance being decoded better than evidence. However, decoding from the 180-dimensional space of the recurrent layer was able to decode evidence and position as well as luminance. Therefore, the information about the luminance remained present in the neural activity, but its contribution to the low-dimensional embedding was much smaller than the contribution from evidence and position.

3.4. Neural activity in deep RL agents

To better understand how learning shapes neural activity, we visualized the activity of neurons in the encoder (Fig. S5 and Fig. S6) and recurrent layers (Fig. S7 and Fig. S8) before and after the training as a function of evidence, position and luminance. While Nieh et al. (2021) reported that many neurons tuned to a particular magnitude of evidence and a particular position, we observed that the majority of neurons in the recurrent layer changed their activity monotonically (growing or decaying) as a function of evidence and position. In particular, we observed that negative and positive evidence were encoded by two distinct populations of neurons whose firing rate increased as a function of the amount of evidence.

4. Discussion

To solve the accumulating towers task, the agents had to keep track of the amount of evidence, which was the difference in the number of towers on the left- and the right-hand side of the environment. The agents also had to learn to identify the end of the track, where they had to turn in the direction with more evidence to receive a reward. The visualization and decoding analysis of the low-dimensional embedding of the neural activity in the recurrent layer revealed a manifold spanned by evidence and position dimensions (Fig. 6 and Fig. 8). Critically, luminance, which was not a task-relevant variable, was not well encoded in the low-dimensional space, even though information about luminance was present in the high-dimensional space. This indicates that the low-dimensional structure reveals the variables that the agent used to solve the task, providing a tool for improving the interpretability of deep RL agents.

The neural activity of the recurrent units was characterized by a monotonic increase as a function of evidence, with different populations of neurons showing selectivity for positive and negative evidence (Fig. S8a). The rates of increase also differed across neurons. Nieh et al. (2021) did not observe such monotonic changes and reported sequential neural activity, suggesting that the artificial agents might converge to a different coding scheme despite using a low-dimensional representation. Earlier computational neuroscience work on evidence accumulation has suggested a connection between neural activity that changes monotonically with a spectrum of rate constants across neurons and a sequential neural activity (Howard et al., 2018). That work suggested that sequential activity could be generated through a simple linear transformation of the monotonic activity.

Our analyses paralleled those from Nieh et al. (2021), providing an example of how the same tools can be used to examine biological and artificial agents. The results indicated that artificial agents can learn tasks that include abstract variables, and that the underlying neural representation relies on a low-dimensional manifold. Studying neural dynamics from a manifold perspective has been increasingly relevant in neuroscience (Langdon et al., 2023; Ebitz and Hayden, 2021; Gallego et al., 2017), providing an opportunity to better understand similarities and differences between biological and artificial systems. The fact that MIND was able to tease apart task-relevant and task-irrelevant variables shows potential for using manifold learning techniques to advance understanding of learning and decision-making in artificial agents.

Acknowledgments

We thank Edward Nieh, Manuel Schottdorf, Rachel Lee and Ehren Newman for helpful conversations. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Dmitriy Aronov, Rhino Nevers, and David W Tank. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647):719–722, 2017.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557 (7705):429–433, 2018.

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Communications of the ACM*, 64(7):58–65, 2021.

Carlos D Brody and Timothy D Hanks. Neural underpinnings of the evidence accumulator. *Current opinion in neurobiology*, 37:149–157, 2016.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2021.

Peter Dayan and Nathaniel D Daw. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, 2008.

Ben Deverett, Ryan Faulkner, Meire Fortunato, Greg Wayne, and Joel Z Leibo. Interval timing in deep reinforcement learning agents. *arXiv preprint arXiv:1905.13469*, 2019.

R Becket Ebitz and Benjamin Y Hayden. The population doctrine in cognitive neuroscience. *Neuron*, 109(19):3055–3068, 2021.

Ben Engelhard, Joel Finkelstein, Julia Cox, Weston Fleming, Hee Jae Jang, Sharon Ornelas, Sue Ann Koay, Stephan Y Thiberge, Nathaniel D Daw, David W Tank, et al. Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. *Nature*, 570 (7762):509–513, 2019.

Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annu. Rev. Neurosci.*, 30:535–574, 2007.

Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *International conference on machine learning*, pages 1792–1801. PMLR, 2018.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.

Marc W Howard, Andre Luzardo, and Zoran Tiganj. Evidence accumulation in a laplace domain decision space. *Computational brain & behavior*, 1(3):237–251, 2018.

Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150, 2018.

John O Keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Clarendon Press, 1978.

Eric B Knudsen and Joni D Wallis. Hippocampal neurons construct a map of an abstract value space. *Cell*, 184(18):4640–4650, 2021.

Christopher Langdon, Mikhail Genkin, and Tatiana A Engel. A unifying perspective on neural manifolds and circuits for cognition. *Nature Reviews Neuroscience*, pages 1–15, 2023.

Rachel S Lee, Ben Engelhard, Ilana B Witten, and Nathaniel D Daw. A vector reward prediction error model explains dopaminergic heterogeneity. *bioRxiv*, pages 2022–02, 2022.

Dongyan Lin and Blake A Richards. Time cell encoding in deep reinforcement learning agents depends on mnemonic demands. *bioRxiv*, 2021.

Ryan J Low, Sam Lewallen, Dmitriy Aronov, Rhino Nevers, and David W Tank. Probing variability in a cognitive map using manifold inference from neural dynamics. *BioRxiv*, page 418939, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

James Mochizuki-Freeman, Sahaj Singh Maini, and Zoran Tiganj. Characterizing neural activity in cognitively inspired rl agents during an evidence accumulation task. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 01–09. IEEE, 2023.

Ari S Morcos and Christopher D Harvey. History-dependent variability in population dynamics during evidence accumulation in cortex. *Nature neuroscience*, 19(12):1672–1681, 2016.

Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. *Advances in neural information processing systems*, 32, 2019.

Edward H Nieh, Manuel Schottdorf, Nicolas W Freeman, Ryan J Low, Sam Lewallen, Sue Ann Koay, Lucas Pinto, Jeffrey L Gauthier, Carlos D Brody, and David W Tank. Geometry of abstract learned knowledge in the hippocampus. *Nature*, pages 1–5, 2021.

Lucas Pinto, Sue A Koay, Ben Engelhard, Alice M Yoon, Ben Deverett, Stephan Y Thiberge, Ilana B Witten, David W Tank, and Carlos D Brody. An accumulation-of-evidence task using visual pulses for mice navigating in virtual reality. *Frontiers in behavioral neuroscience*, 12:36, 2018.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shimon Ullman. Using neuroscience to develop artificial intelligence. *Science*, 363(6428): 692–693, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Lindsay Wells and Tomasz Bednarz. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in artificial intelligence*, 4:550030, 2021.

Supplemental Information

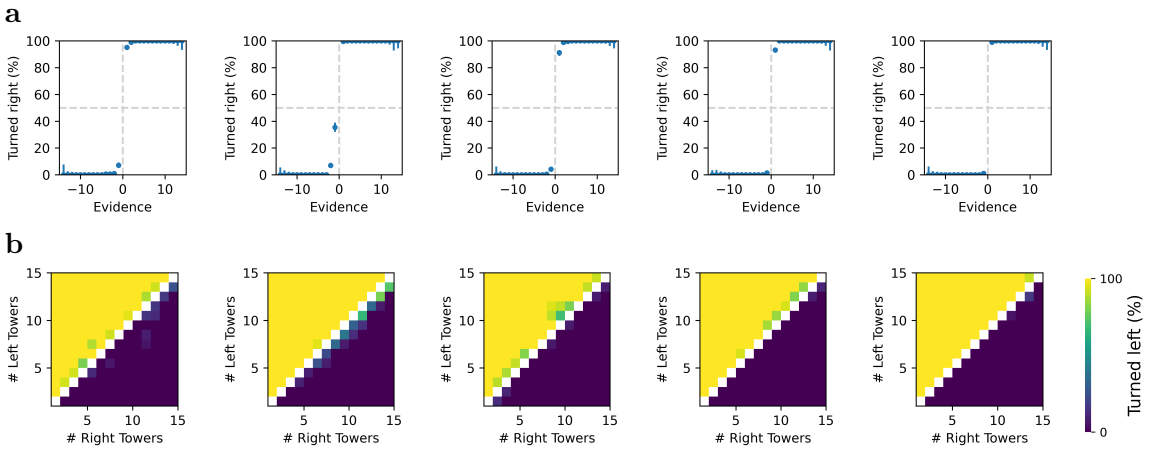

Figure S1: Performance of the agents on the accumulating towers task *after training*, captured through: **(a)** Psychometric curves. **(b)** Confusion matrix.

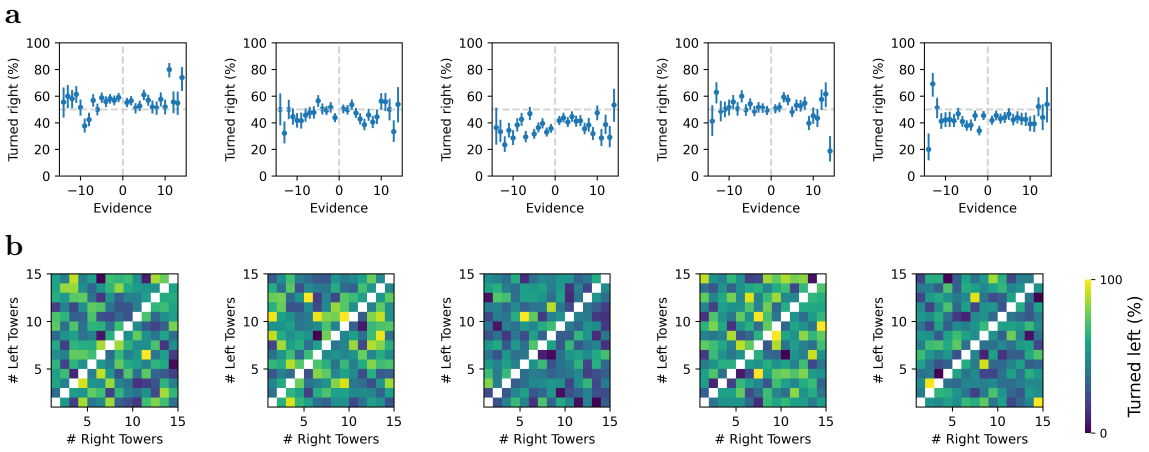

Figure S2: Performance of the agents on the accumulating towers task *before training*, captured through: **(a)** Psychometric curves. **(b)** Confusion matrix.

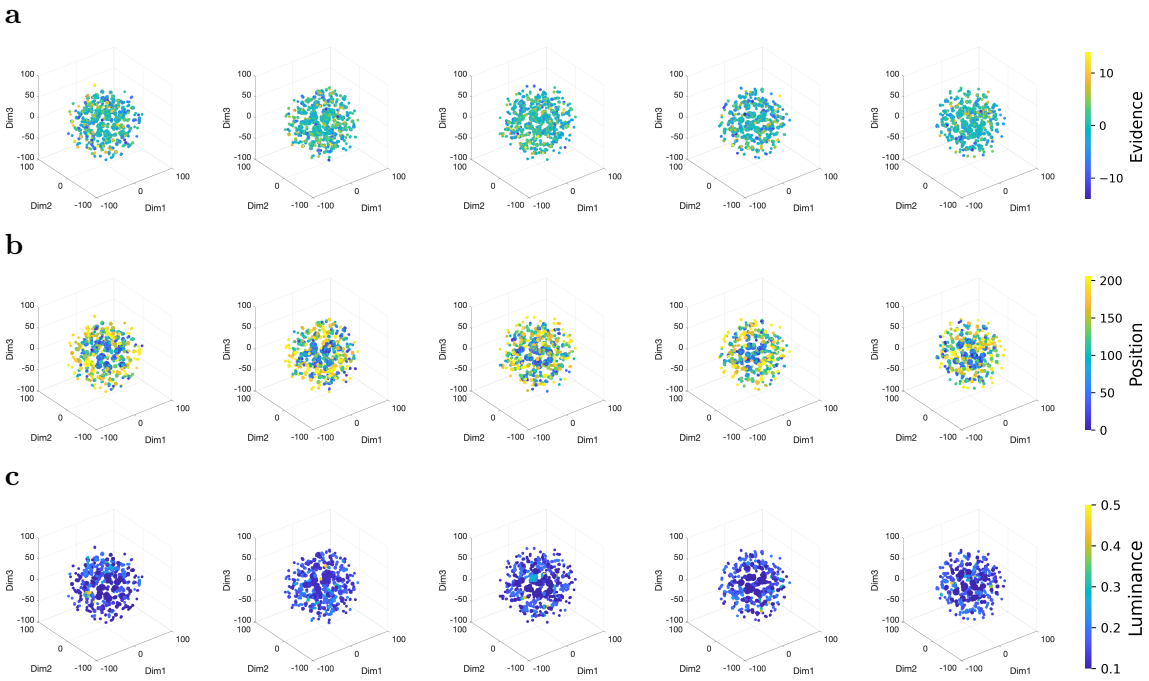

Figure S3: Three-dimensional embedding of the neural activity in the encoder layer for each of the five deep RL agents. The embedding was computed using t-SNE and it is color-coded by **(a)** amount of evidence, **(b)** position, and **(c)** luminance.

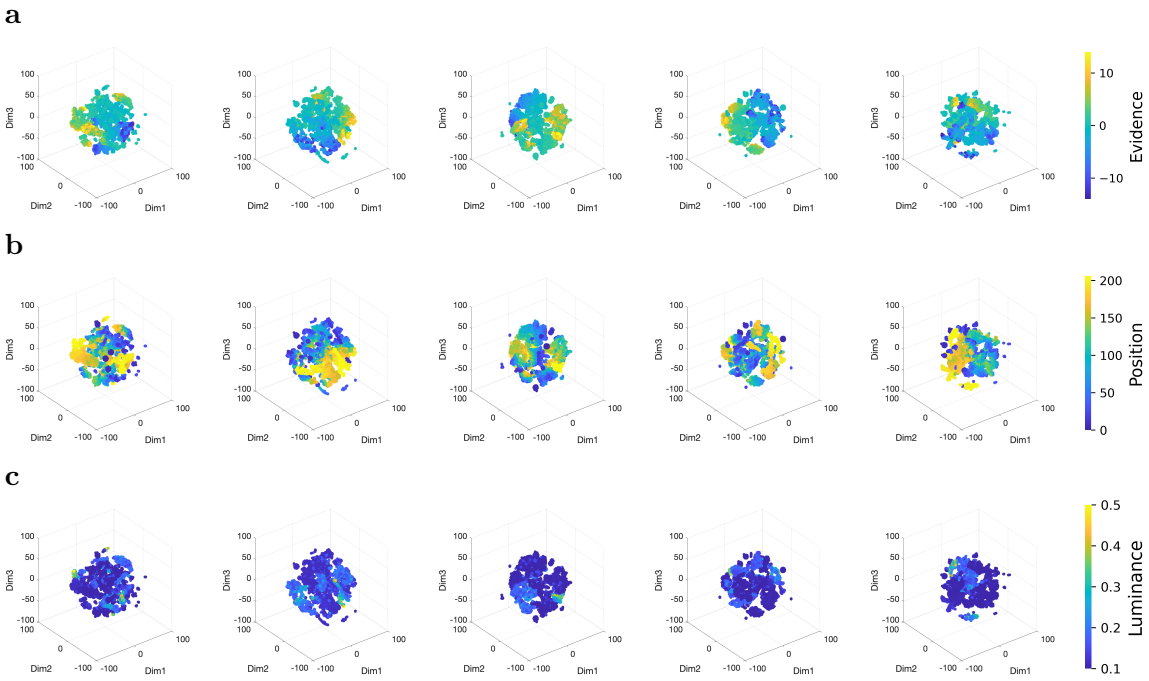

Figure S4: Three-dimensional embedding of the neural activity in the recurrent layer for each of the five deep RL agents. The embedding was computed using t-SNE and it is color-coded by **(a)** amount of evidence, **(b)** position, and **(c)** luminance.

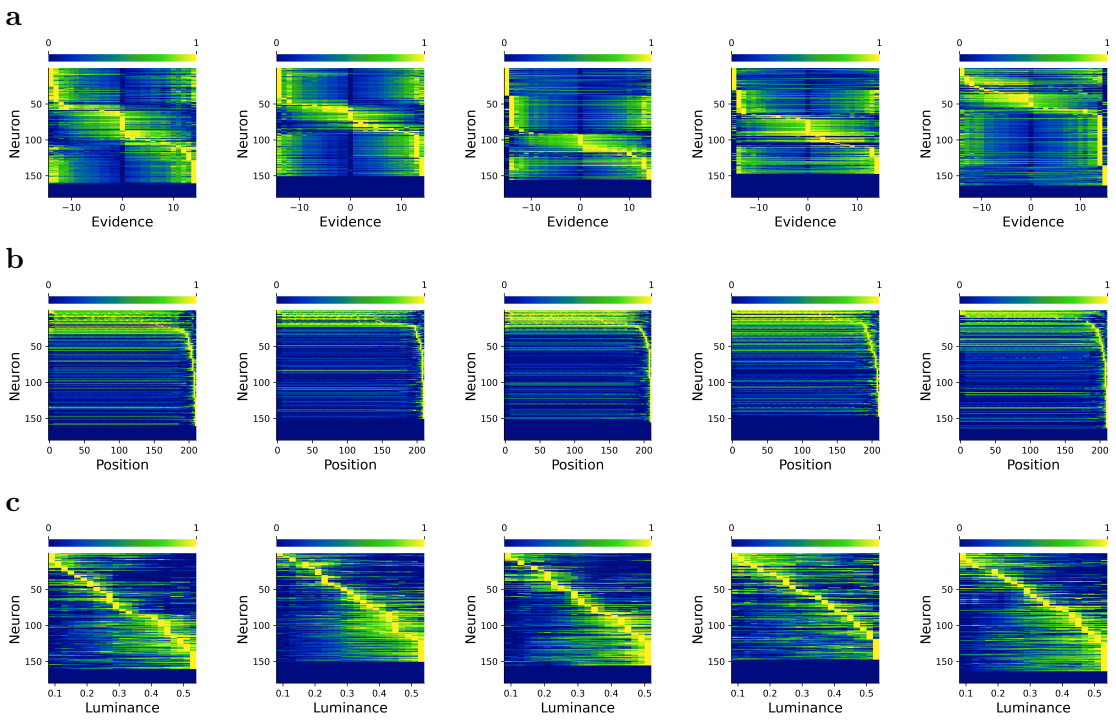

Figure S5: Activity of the neurons in the encoder layer *before training* for each of the five agents as a function of **(a)** evidence, **(b)** position, and **(c)** luminance. Neurons are sorted by their peak activity which is rescaled to a 0−1 range.

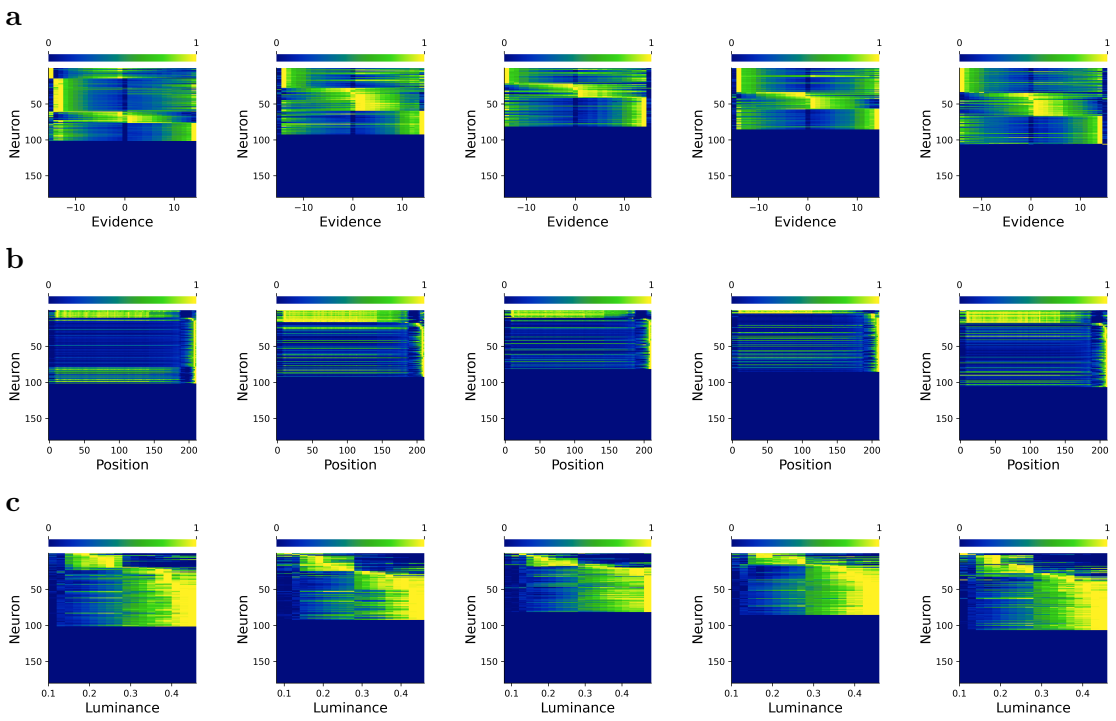

Figure S6: Activity of the neurons in the encoder layer *after training* for each of the five agents as a function of **(a)** evidence, **(b)** position, and **(c)** luminance. Neurons are sorted by their peak activity which is rescaled to a 0−1 range.

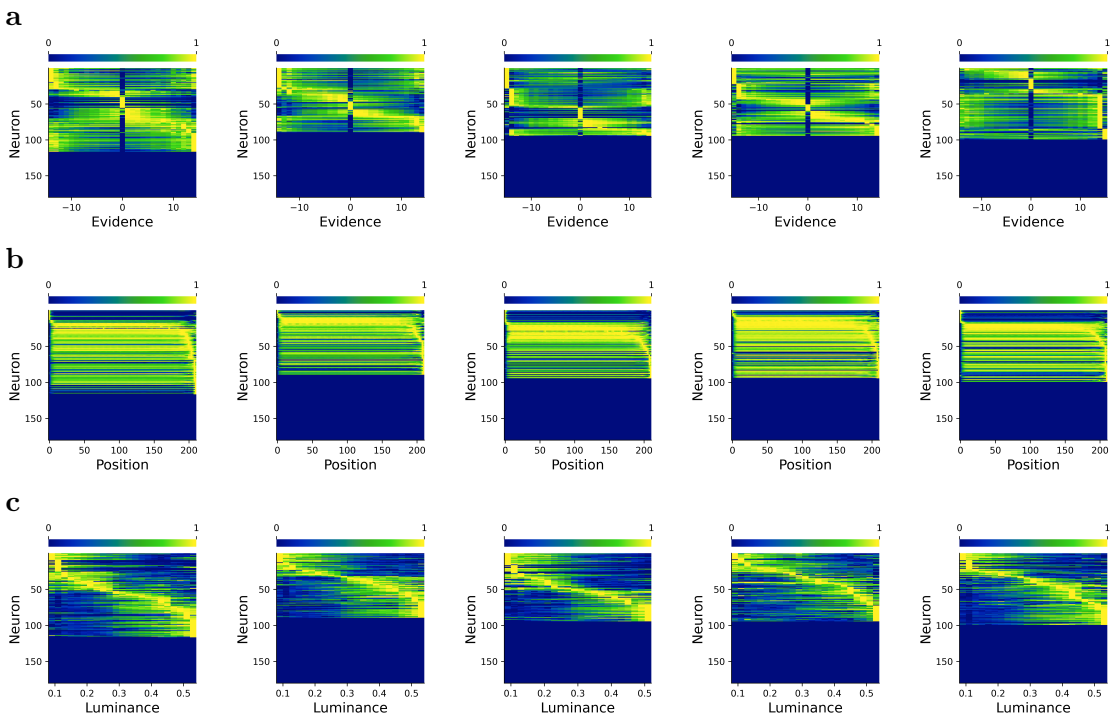

Figure S7: Activity of the neurons in the recurrent layer *before training* for each of the five agents as a function of **(a)** evidence, **(b)** position, and **(c)** luminance. Neurons are sorted by their peak activity which is rescaled to a 0−1 range..

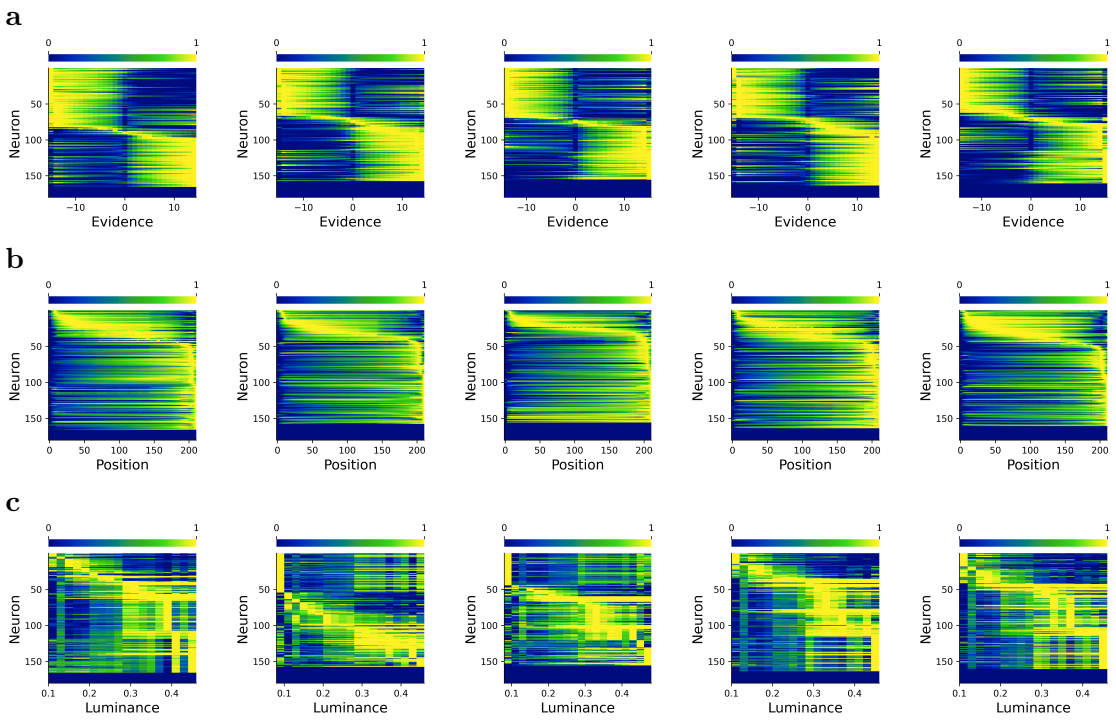

Figure S8: Activity of the neurons in the recurrent layer *after training* for each of the five agents as a function of **(a)** evidence, **(b)** position, and **(c)** luminance. Neurons are sorted by their peak activity which is rescaled to a 0−1 range.