Neurobehavior of exploring AI agents

Isaac Kauvar* Stanford University ikauvar@stanford.edu

Chris Doyle* Stanford University crd@stanford.edu

Nick Haber Stanford University nhaber@stanford.edu

Abstract

We study intrinsically motivated exploration by artificially intelligent (AI) agents in animal-inspired settings. We construct virtual environments that are 3D, vision-based, physics-simulated, and based on two established animal assays: labyrinth exploration, and novel object interaction. We assess Plan2Explore (P2E), a leading model-based, intrinsically motivated deep reinforcement learning agent, in these environments. We characterize and compare the behavior of the AI agents to animal behavior, using measures devised for animal neuroethology. P2E exhibits some similarities to animal behavior, but is dramatically less efficient than mice at labyrinth exploration. We further characterize the neural dynamics associated with world modeling in the novel-object assay. We identify latent neural population activity axes linearly associated with representing object proximity. These results identify areas of improvement for existing AI agents, and make strides toward understanding the learned neural dynamics that guide their behavior.

1 Introduction

To survive, animals have evolved effective strategies for exploration. Animals are intrinsically motivated and guided by curiosity to investigate novel stimuli [Glickman and Sroges, 1966, Ahmadlou et al., 2021, Ogasawara et al., 2022], seek information [Lydon-Staley et al., 2021, Bromberg-Martin and Hikosaka, 2009, Gottlieb and Oudeyer, 2018], and efficiently explore complex mazes [Rosenberg et al., 2021]. Exploration is also a key challenge for artificially intelligent (AI) systems. Inspiration from animal curiosity has led to many recently developed methods that promote exploration through intrinsic reward signals. How effective has this translation from animals to AI been and how much room for improvement remains? One way to investigate these questions is to assess curious AI agents in scenarios that are informed by ethologically-relevant animal assays [Zador et al., 2023].

In this work, we bridge AI and animal exploration by developing virtual assays for AI agents that are inspired by existing, well-characterized assays of animal exploration. In turn, we can directly compare AI agent exploration with animal behavior, to identify gaps in AI performance and potential routes towards improvement. Moreover, by situating AI agents in ethologically relevant scenarios, we also better position ourselves to apply neuroscientific approaches towards understanding the neural computations that drive AI agent behavior [Merel et al., 2019].

Building on two detailed animal studies, we investigate novel object investigation [Ahmadlou et al., 2021], and spatial exploration in a labyrinth [Rosenberg et al., 2021]. As a starting point, we assess Plan2Explore [Sekar et al., 2020], a leading model-based, intrinsically-motivated, deep reinforcement learning (RL) agent that is compatible with visual input and performs well on exploration benchmarks.

Contributions (1) We construct two virtual environments for studying exploration by deep RL agents, based on animal exploration assays. (2) We characterize the exploratory behavior of Plan2Explore, compare it directly with mouse behavior, and identify gaps in its performance. (3) We investigate neural dynamics learned by Plan2Explore during object investigation.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

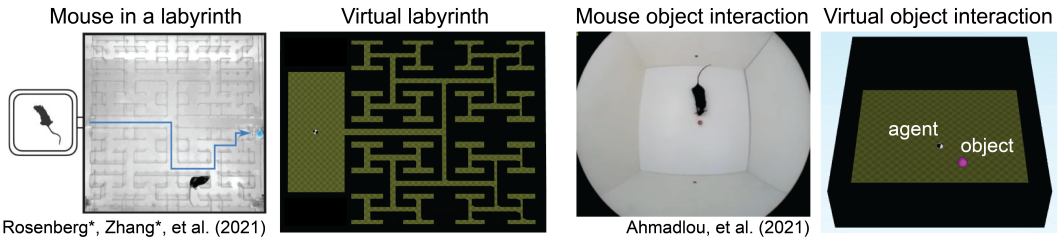

Figure 1: Virtual environments to study AI exploration, based on mouse labyrinth exploration [Rosenberg et al., 2021] and object interaction assay [Ahmadlou et al., 2021].

2 Related work

Intrinsic motivation We build on a large body of work related to using intrinsic reward to guide exploration [Schmidhuber, 2010], which has included notions of prediction loss [Pathak et al., 2017, Haber et al., 2018, Guo et al., 2022], novelty [Bellemare et al., 2016, Burda et al., 2018], learning progress [Kim et al., 2020, Lopes et al., 2012], and uncertainty [Sekar et al., 2020, Pathak et al., 2019]. Here we focus on Plan2Explore [Sekar et al., 2020], which uses latent disagreement to effectively promote exploration. We also build on neuroscientific investigations of exploration, including [Kidd and Hayden, 2015, Gottlieb and Oudeyer, 2018, White et al., 2019, Ogasawara et al., 2022].

Neurobehavioral investigation of AI systems We are inspired by work that bridges animals and AI, including use of neural analysis methods to understand computational systems [Zeiler and Fergus, 2014, Jonas and Kording, 2017, Olah et al., 2017, Merel et al., 2019], and use of deep neural networks to model animal behavior and neural activity [Yamins and DiCarlo, 2016, Richards et al., 2019, Cross et al., 2021, Doyle et al., 2023, Nayebi et al., 2023, Bonnen et al., 2023, Martinez et al., 2023].

3 Environments

We build 3D physics-simulated environments that provide egocentric, image-based input, using dm_control [Tunyasuvunakool et al., 2020] and its extensions [Kauvar et al., 2023]. See the Appendix for detailed description of the environments[1] and Plan2Explore agent[2].

Labyrinth Closely following the labyrinth design of Rosenberg et al. [2021], we constructed a symmetric maze with 64 end nodes, connected to a 'home' room. Many interesting features of mouse behavior were identified by Rosenberg et al. [2021] that provide opportunities for precisely comparing animal and AI behavior. First, mice spend a large amount of time in the maze and actively explore over 80% of that time. Second, mice efficiently explore the maze, as measured by the rate of visiting new nodes versus familiar nodes. Third, mouse behavior is strongly influenced by its current and very recent locations, and is well described by simple biases that guide decision making.

Novel object investigation Closely following the free-access double-choice design of Ahmadlou et al. [2021], we constructed a two-phase assay in which an agent has an initial opportunity to freely investigate one object, and then a second novel object is added into the environment. Investigation of each object is then compared. A number of interesting features of mouse behavior were identified by Ahmadlou et al. [2021] . First, mice are motivated to investigate the objects. Second, mice spend much more time investigating the novel object. Third, mice also spend more time 'deeply' investigating the novel object, as measured by behaviors such as biting, grabbing, and carrying.

4 Experiments

Labyrinth exploration We recorded exploration trajectories of 8 randomly initialized P2E agents. All agents successfully entered the labyrinth from the home room, and many explored a substantial fraction of the labyrinth (Figure 2a). Agents spent a similar fraction of time in the maze as mice, with a similar initial sharp increase followed by a slow decay (compare with Figure S2 of [Rosenberg

[1]Available as part of the adaptgym suite: `pip install adaptgym`
[2]Available at: `https://github.com/AutonomousAgentsLab/imol-explore-suite`

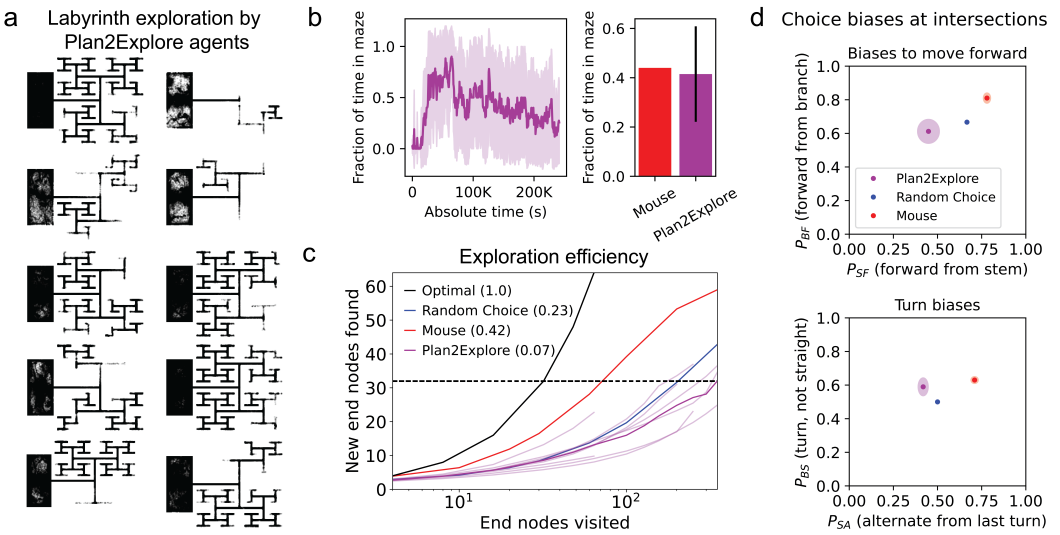

Figure 2: Comparison of labyrinth exploration by mice and Plan2Explore. a) P2E trajectories, across 15M timesteps (125 simulated hours). b) P2E spends a similar fraction of time in the maze as mice. c) Efficiency of exploration (all data except P2E is from [Rosenberg et al., 2021]). P2E is worse than mice, and random choice agent. d) Decision biases in exploration strategy. Mice bias toward proceed forward through intersections and alternating which direction they turn, but P2E does not.

et al., 2021]). Agents that randomly sampled the continuous action space failed to even enter the labyrinth. Surprisingly, though Plan2Explore explored, it does so very inefficiently, performing worse even than a 'random choice' agent that navigates between nodes by choosing a random direction at each intersection. In contrast, mice are much more efficient than the 'random choice' agent, but not quite optimal [Rosenberg et al., 2021]. Additionally, while mice exhibit a strong bias to move forward, turn if possible, and alternate turning directions, Plan2Explore agents exhibit less forward bias than even the 'random choice' agent and do not bias toward alternating turns.

Novel object investigation We recorded the trajectories of four P2E agents on the novel object assay, for 500K steps with a magenta ball, then 500K steps with both magenta and yellow balls. P2E agents interacted with the magenta ball during the first phase, and preferred the new ball during the second phase - similar to mice (Figures 3a, b). In contrast to mice, P2E's preference for the new ball was delayed. Similarly, the agent initially attends to the old ball before eventually prioritizing the new ball. Similar to mice, we found a bias towards deep exploration with the new ball, and shallow

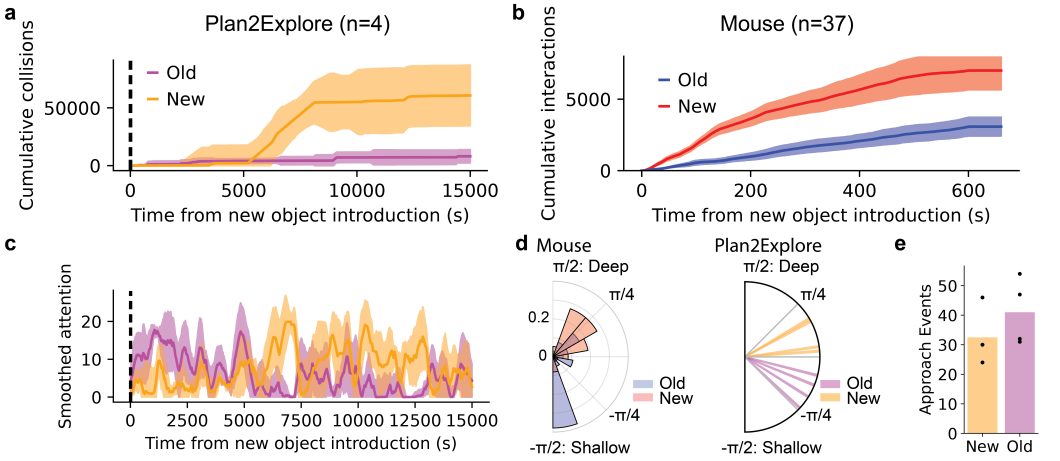

Figure 3: Comparison of novel object investigation by mice and Plan2Explore. a) Mouse and b) P2E agent interactions with the new and old objects. c) P2E agent attention over time. d) Bias towards deep vs. shallow investigation. e) P2E approach events to new and old ball.

exploration with the old ball, although the bias was less extreme than with mice. In contrast to mice, we did not find a preference for approach events towards the new ball.

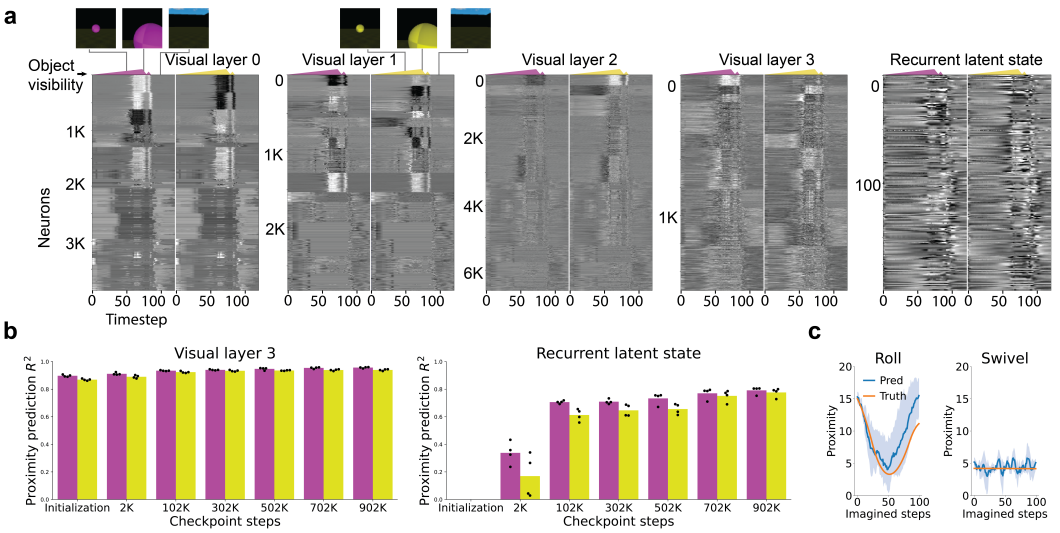

Figure 4: Neural analysis of object representations. a) Raster of clustered neural traces from visual and recurrent layers, comparing activations by magenta and yellow objects. b) Linear decoding of object proximity. c) Action-conditioned imagination, projected onto proximity dimension.

5 Neural analysis

Next, we sought to understand the neural computations underlying P2E's object investigation by probing the learned object representations. Inspired by object-vector coding in the medial entorhinal cortex [Høydal et al., 2019], one hypothesis is that the agent learns linearly decodable representations of key object features such as proximity. To study this, and to assess if these representations are present across various 'brain regions' (e.g. layers of the visual encoder and recurrent latent state), we generated a set of probe scenarios, ran agents on these scenarios using checkpoints from throughout training, and saved neural activation timeseries from the visual layers and world model latent state. In Figure 4a, we compare the neural activity corresponding to test sessions that are identical except for the object color (magenta vs. yellow), showing clear separation in population neural activity.

We used linear ridge regression to identify directions of neural population activity encoding the agent's proximity to each object. As evaluated with 7-fold cross-validation on 14K timesteps, proximity prediction is eventually very good at all layers, is better in the visual layers than in the recurrent latent state, and improves across checkpoints - particularly for the yellow (novel) object (Figures 4b, 5). These results provide insight into the process by which the world model learns to represent the environment. First, even though the visual layers contain clear information about the object, even at initialization, the world model needs experience and training to incorporate this information into its state representation. Second, the latent state learns a representation of the familiar object with information that is more specific to that object than to the novel object. As the novel object becomes more familiar, the representations of the novel and familiar objects converge toward having a similar degree of informativeness. The growth of encoded information about the novel object as a function of increased exposure to that object is particularly apparent in the jump in predictivity between steps 502K and 702K (the novel object was introduced at step 500K).

In addition, we assess the role of this 'proximity' neural direction in the world model's action-conditioned imagination. We find that rolling forward and backward (the first action space dimension) projects strongly onto this direction, but swiveling (the second dimension) does not (Figure 4c). This suggests the existence of a learned circuit that converts the first action dimension into shifts of neural activity along a specific direction and that could subserve the agent's ability to predict the outcome of imagined actions. By tracing the first action dimension's flow into the latent state, future work may be able to mechanistically identify a circuit in the network weights that converts action into imagined state prediction. Identification of such a circuit would provide insight into how P2E learns to link actions with changes in abstract state. It would also model how animal brains might learn to

mental time travel [Suddendorf and Corballis, 2007] and leverage corollary discharges to predict the consequences of executed actions [Guthrie et al., 1983, Crapse and Sommer, 2008].

6 Conclusion

In this work, we set out to directly compare intrinsically-motivated AI exploration with animal behavior. Plan2Explore agents exhibited some similarities to animal behavior, including a degree of novel object preference. However, there remain gaps in performance, including markedly reduced efficiency in labyrinth exploration, and slow initiation of novel object investigation. These may point to potential algorithmic improvements. For example, initial investigations suggest that P2E lacks allocentric localization in the labyrinth, and that exploration may be aided by adding in some capacity for allocentric spatial mapping, through architectural or training adjustments. Additionally, we take initial steps toward understanding the computations learned by Plan2Explore's neural circuits. We think this is an interesting direction for assessing neuroscience analysis methods (in settings more reminiscent of their original biological application than e.g. [Jonas and Kording, 2017]) and for gaining insight into how AI systems work and how they falter. In sum, we present tools and a strategy for directly using animal-inspiration to identify avenues for understanding and improving AI systems.

References

M. Ahmadlou, J. H. Houba, J. F. van Vierbergen, M. Giannouli, G.-A. Gimenez, C. van Weeghel, M. Darbanfouladi, M. Y. Shirazi, J. Dziubek, M. Kacem, et al. A cell type–specific cortico-subcortical brain circuit for investigatory and novelty-seeking behavior. *Science*, 372(6543): eabe9681, 2021.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

T. Bonnen, A. D. Wagner, and D. L. Yamins. Medial temporal cortex supports compositional visual inferences. *bioRxiv*, pages 2023–09, 2023.

E. S. Bromberg-Martin and O. Hikosaka. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, 63(1):119–126, 2009.

Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

T. B. Crapse and M. A. Sommer. Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600, 2008.

L. Cross, J. Cockburn, Y. Yue, and J. P. O'Doherty. Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, 109(4):724–738, 2021.

C. Doyle, S. Shader, M. Lau, M. Sano, D. L. Yamins, and N. Haber. Developmental curiosity and social interaction in virtual agents. *arXiv preprint arXiv:2305.13396*, 2023.

S. E. Glickman and R. W. Sroges. Curiosity in zoo animals. *Behaviour*, 26(1-2):151–187, 1966.

J. Gottlieb and P.-Y. Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):758–770, 2018.

Z. Guo, S. Thakoor, M. Pîslar, B. Avila Pires, F. Altché, C. Tallec, A. Saade, D. Calandriello, J.-B. Grill, Y. Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35:31855–31870, 2022.

B. L. Guthrie, J. D. Porter, and D. L. Sparks. Corollary discharge provides accurate eye position information to the oculomotor system. *Science*, 221(4616):1193–1195, 1983.

N. Haber, D. Mrowca, S. Wang, L. F. Fei-Fei, and D. L. Yamins. Learning to play with intrinsically-motivated, self-aware agents. *Advances in neural information processing systems*, 31, 2018.

D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

D. Hafner, K.-H. Lee, I. Fischer, and P. Abbeel. Deep hierarchical planning from pixels. *Advances in Neural Information Processing Systems*, 35:26091–26104, 2022.

D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

J. Hao, T. Yang, H. Tang, C. Bai, J. Liu, Z. Meng, P. Liu, and Z. Wang. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Ø. A. Høydal, E. R. Skytøen, S. O. Andersson, M.-B. Moser, and E. I. Moser. Object-vector coding in the medial entorhinal cortex. *Nature*, 568(7752):400–404, 2019.

E. Jonas and K. P. Kording. Could a neuroscientist understand a microprocessor? *PLoS computational biology*, 13(1):e1005268, 2017.

I. Kauvar, C. Doyle, L. Zhou, and N. Haber. Curious replay for model-based adaptation. *International Conference on Machine Learning*, 2023.

C. Kidd and B. Y. Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460, 2015.

K. Kim, M. Sano, J. De Freitas, N. Haber, and D. Yamins. Active world model learning with progress curiosity. In *International conference on machine learning*, pages 5306–5315. PMLR, 2020.

M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. *Advances in neural information processing systems*, 25, 2012.

D. M. Lydon-Staley, D. Zhou, A. S. Blevins, P. Zurn, and D. S. Bassett. Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature human behaviour*, 5(3): 327–336, 2021.

J. Martinez, F. Binder, H. Wang, N. Haber, J. Fan, and D. L. Yamins. Measuring and modeling physical intrinsic motivation. *arXiv preprint arXiv:2305.13452*, 2023.

J. Merel, D. Aldarondo, J. Marshall, Y. Tassa, G. Wayne, and B. Ölveczky. Deep neuroethology of a virtual rodent. *arXiv preprint arXiv:1911.09451*, 2019.

A. Nayebi, R. Rajalingham, M. Jazayeri, and G. R. Yang. Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes. *arXiv preprint arXiv:2305.11772*, 2023.

T. Ogasawara, F. Sogukpinar, K. Zhang, Y.-Y. Feng, J. Pai, A. Jezzini, and I. E. Monosov. A primate temporal cortex–zona incerta pathway for novelty seeking. *Nature neuroscience*, 25(1):50–60, 2022.

C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

D. Pathak, D. Gandhi, and A. Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

M. Rosenberg, T. Zhang, P. Perona, and M. Meister. Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *Elife*, 10:e66175, 2021.

J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE transactions on autonomous mental development*, 2(3):230–247, 2010.

R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. In *ICML*, 2020.

T. Suddendorf and M. C. Corballis. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and brain sciences*, 30(3):299–313, 2007.

E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, N. Heess, and Y. Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6: 100022, 2020. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100022. URL https://www.sciencedirect.com/science/article/pii/S2665963820300099.

J. K. White, E. S. Bromberg-Martin, S. R. Heilbronner, K. Zhang, J. Pai, S. N. Haber, and I. E. Monosov. A neural network for information seeking. *Nature communications*, 10(1):5168, 2019.

D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

A. Zador, S. Escola, B. Richards, B. Ölveczky, Y. Bengio, K. Boahen, M. Botvinick, D. Chklovskii, A. Churchland, C. Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

Appendix

A Environments (detailed description)

Our environments build on the MuJoCo-based Todorov et al. [2012] dm_control simulation framework Tunyasuvunakool et al. [2020] and its adaptgym extension Kauvar et al. [2023], which allows for easy construction of bespoke arenas populated with controllable objects. This framework supports 3D physics-simulated virtual environments that provide egocentric, image-based input to an agent. We run our environments with a physics simulation timestep of 0.5 ms and an action timestep of 30 ms. We developed two environments.

Labyrinth Closely following the labyrinth design of Rosenberg et al. [2021], we constructed a symmetric binary maze with 64 end nodes and a single entrance connected to a 'home' room. One advantageous feature of this design is that it provides many decision points that can be used for detailed behavior characterization. Navigating from the home room to any end node requires a sequence of six binary decisions. Moreover, with this design, mice can be left unattended for long stretches of time (e.g. overnight), allowing assessment of long-term exploratory behavior. A number of interesting features of mouse behavior were identified by Rosenberg et al. [2021]. First, mice spend a large amount of time exploring the maze: around half of a seven-hour period is spent inside the maze, and over 80% of that time is spent exploring the maze, even if there is an extrinsic reward located at one end node. Second, mice efficiently explore the maze, reaching all end nodes far more quickly than an agent that chooses a random direction at each intersection. Third, mouse behavior is strongly influenced by its current location and the 3 or so locations preceding it, and can be fairly well described by turning biases that guide local decision making. Overall, this environment provides an opportunity for precisely comparing animal and AI behavior along a number of detailed axes.

As in the mouse assay, the walls are black in our virtual labyrinth and offer no distinguishing features. Experiments with the mice demonstrated that they do not rely on odor to navigate the maze, and thus depend primarily on visual and tactile input. Our agents are only provided visual input.

We assessed exploration efficiency by plotting the number of new end nodes found as a function of total number of end nodes visited, and compared this to animal data from [Rosenberg et al., 2021]. Mouse behavior can also be fairly well characterized by a set of biases in the choices made at intersections, including a bias to move forward through each intersection (quantified by [Rosenberg et al., 2021] as P_{BF} and P_{SF}, the probabilities of moving forward out of a branch or stem, respectively), and a bias to turn when possible (P_BS), and to alternate turning directions (P_{SA}). We measured these biases in the virtual environment as well.

Novel object investigation Closely following the free-access double-choice design of Ahmadlou et al. [2021], we constructed an assay in which an agent is first provided an extended opportunity to freely investigate one object, and then a second, novel object is added into the environment and investigation of each object is compared. An advantageous feature of this design is that it probes the intrinsic motivation of the animal to investigate the different objects, as opposed to merely preference in a forced choice. A number of interesting features of mouse behavior were identified by Ahmadlou et al. [2021] . First, mice are motivated to investigate the objects. Second, spend much more time investigating the novel object. Third, mice spend much more time 'deeply' investigating the novel object, as measured by behaviors such as biting, grabbing, and carrying the object. Moreover, a neural substrate that modulates this motivated novel-object investigation was identified: inhibitory neurons in the medial zona incerta brain region.

In the virtual assay, the agent is placed in a square arena with black walls that is roughly 100 times bigger than the agent. The objects are spheres and are slightly larger than the agent. The familiar and novel objects differ only by color. In the experiments we present here, the familiar object is two-toned magenta and the novel object is two-toned yellow. Future experiments could counterbalance the colors and investigate the impact of altering the colors and properties of the object. Importantly, the agent can interact with the objects by colliding with them, allowing the agent to curiously investigate the physical dynamics associated with the object as it learns to model the world.

We quantify agent-object interactions as collisions between the agent and object, which are recorded by the environment. We measure attention of the agent towards an object by saying it is attending to the object if the center of mass of the object is within a 60 degree field of view cone from the agent.

We define a measure of 'deep' exploration as time periods in which the agent collides with the ball at least 50 % of the time, and compared this with shallow exploration, when the agent is near the ball and attending to it but not colliding with it. We also quantified approach events, defined as times when the agent turned toward the ball and began moving towards it.

In both environments, there are a number of notable differences from the animal experiments. First, our agent has a very simple embodiment: the agent is a sphere with a two dimensional continuous action space consisting of rolling forward (and backward) and turning. Second, the agent receives monocular visual input, and no other sensory input. Third, agents are initialized with random network weights when they are spawned. The first two simplifications minimize the necessity of pretraining–in contrast to the scenario, for example, where the agent had a complex embodiment that required substantial learning before locomotion was even possible. One strategy we use to address this lack of a pretraining phase, is to provide the agent longer sessions of exploration, giving it the opportunity to warm up its models. Notably, many of the key metrics for characterizing exploratory behavior do not depend on the absolute time of exploration. For example, in the labyrinth, the efficiency of visiting different end nodes depends only on the sequence in which nodes are visited, and in the novel object assay, the object preference depends only on the relative interaction time. In practice, we find that the agent does explore the maze and investigate the objects, and it does so on absolute simulated timescales within approximately an order of magnitude of the animals. Nevertheless, it is important to consider the essential influence of a mouse's life experience and ancestral evolution that shape its behavior, and how that differs from this setting.

B Agent (detailed description)

We used Plan2Explore [Sekar et al., 2020], as applied to DreamerV2 [Hafner et al., 2020]. While there are a multitude of exploration strategies for deep reinforcement learning [Hao et al., 2023], we selected Plan2Explore to study because it is one of the leading model-based agents guided by intrinsic motivation. Model-based agents have begun to exhibit impressive performance, surpassing model-free agents in many realms [Hafner et al., 2020, 2022, 2023] with improved sample efficiency and planning ability. Improving the exploration of model-based systems thus presents a pressing challenge.

Our Plan2Explore agent had a discrete latent space with a 200 dimension hidden state, 200 dimension deterministic state, and 32 x 32 discrete latent state. Additional hyperparameter settings were as follows: action repeat=2, prefill=1000, pretrain=100, grad heads=[decoder, reward], model lr=3e-4, actor lr=8e-5, critic lr=8e-5, actor ent=1e-4. For Plan2Explore disagreement, the number of ensemble models=10. The other hyperparameters were the default, including dataset batch=16 and length=50, image render size=[64, 64], actor and critic each 4 layer MLPs with 400 units and elu activation, imagination horizon=15, discount=0.99, discount lambda=0.95, slow target update=100, slow target fraction=1, slow baseline=True.

C Additional neural analysis

We note that while object features such as color and proximity must at some level be nonlinearly decodable (e.g. through the image decoder that is used to train the world model), the agent does not necessarily learn parsimonious linear dimensions representing these intuitively important features of the environment.

Here we present additional results, including the linear decoding performance of the first three visual layers. Additionally we show linear decoding from imagination of a test episode where the agent swivels back and forth (while maintaining the object in view), without rolling forward.

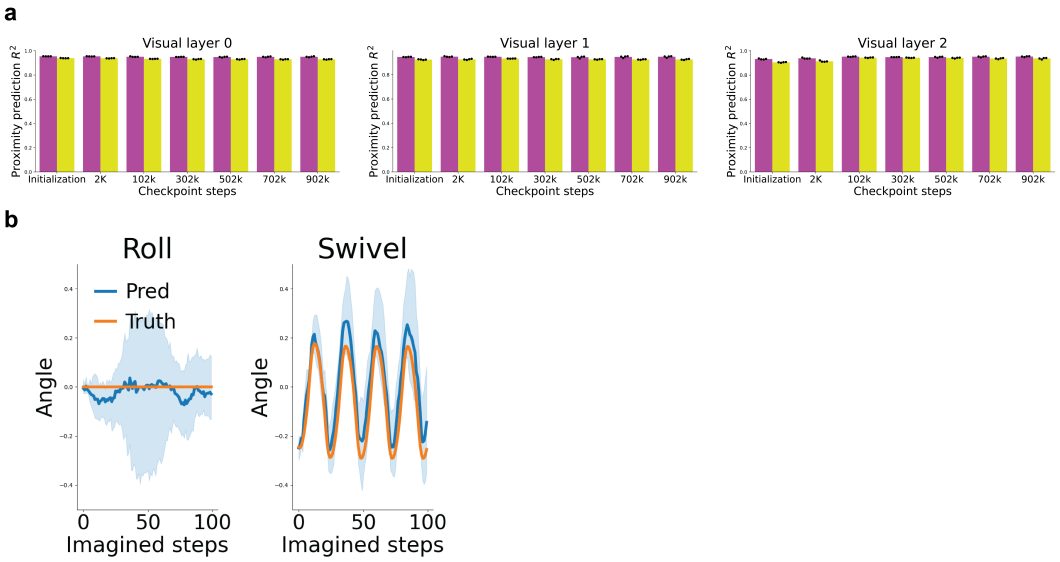

Figure 5: Effective linear decoding of object proximity from earliest visual layers is possible even with random initialization.