Imprinting in autonomous artificial agents using deep reinforcement learning

Donsuk Lee
Department of Informatics
Indiana University Bloomington
Bloomington, IN 47408
donslee@iu.edu

Samantha M. W. Wood
Department of Informatics
Indiana University Bloomington
Bloomington, IN 47408
sw113@iu.edu

Justin N. Wood
Department of Informatics
Indiana University Bloomington
Bloomington, IN 47408
woodjn@iu.edu

Abstract

Imprinting is a common survival strategy in which an animal learns a lasting preference for its parents and siblings early in life. To date, however, the origins and computational foundations of imprinting have not been formally established. What learning mechanisms generate imprinting behavior in newborn animals? Here, we used deep reinforcement learning and intrinsic motivation (curiosity), two learning mechanisms deeply rooted in psychology and neuroscience, to build autonomous artificial agents that imprint. When we raised multiple artificial agents together in the same environment, mimicking the early social experiences of newborn animals, the agents spontaneously developed imprinting behavior. Our results provide a pixels-to-actions computational model of animal imprinting. We show that domain-general learning mechanisms—deep reinforcement learning and intrinsic motivation—are sufficient for embodied agents to rapidly learn core social behaviors from unsupervised natural experience.

1 Introduction

Since the Nobel prize-winning work of Konrad Lorenz, imprinting has become one of the most well-known behaviors in the biological sciences [1–6]. Filial imprinting occurs during a short sensitive period immediately after birth, when an animal develops a lasting attachment to an object or agent. Many precocial avian species (e.g., chickens, ducks, geese) imprint to the first conspicuous objects they see after hatching [5, 7]. In naturalistic settings, these conspicuous objects are parents and siblings. Thus, imprinting allows animals to quickly learn to recognize caregivers and conspecifics, providing the survival benefits associated with group living. Evolution has equipped newborn brains with an intrinsic motivation system that drives animals to imprint. But what computational components underlie this behavior? What motivational and information-processing systems do newborn animals need to generate imprinting behavior from unsupervised natural experience?

While many researchers assume that imprinting requires specialized, domain-specific learning (e.g., [8]), others have argued that imprinting is an example of domain-general learning [9]. Neural evidence suggests that imprinting is specialized because it relies on an "imprinting circuit" extending from the visual wulst (VW) to the intermediate medial mesopallium (IMM) to the amygdala. The IMM has been linked specifically to imprinting [5, 10, 11]. However, this "imprinting circuit" is also

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

consistent with a general-learning account of imprinting because these brain regions are not merely engaged in imprinting. Each of these brain regions has domain-general functions and are engaged in a variety of cognitive tasks. For example, the IMM is also implicated in passive-avoidance learning [12] and memory for food retrieval [13]. The VW, IMM, and amygdala correspond to familiar mammalian homologs—-the visual cortex, association cortex, and amygdala, respectively—all of which are implicated in basic tasks like object recognition and preference formation. Further evidence that imprinting is based on domain-general learning processes comes from studies of imprinting preferences. Chicks are not innately destined to imprint to conspecifics, rather they can learn to imprint to a wide variety of naturalistic [14] and artificial stimuli [15]. Here, we test the hypothesis that imprinting behavior can emerge from domain-general learning, by equipping artificial agents solely with domain-general learning algorithms and testing whether the agents develop imprinting behavior.

We explored the possibility that two biologically inspired learning mechanisms, deep reinforcement learning and intrinsic motivation, provide the computational foundations for imprinting. We built artificial brains, embodied those brains in virtual animal bodies as "agents," and raised those agents in realistic virtual environments. Like real animals, our agents learned to perform actions from raw sensory inputs. Because the agents were active participants in the environment, their learning experiences were self-organized: the agents' actions determined what their learning inputs would be. Importantly, learning in our agents was not guided by any external rewards. Our agents were intrinsically motivated by prediction-based curiosity. Specifically, the agents learned to predict the consequences of their actions and actively search for informative experiences (those which produced high prediction errors given their current knowledge). Curiosity is an evolutionarily adaptive strategy for learning agents [16, 17] because curiosity encourages organisms to develop an abstract internal world model of their evolving environment [18].

We show that when learning is subject to a critical period (akin to imprinting in real animals), these artificial agents spontaneously imprint to one another. These results provide the first pixels-to-actions model of animal imprinting: an important step towards formalizing the computational mechanisms that underlie this foundational behavior.

2 Experiments and Results

We probed which computational components are sufficient to produce imprinting behavior by creating autonomous artificial agents—deep neural networks embodied in virtual animal bodies—and raising those agents in realistic virtual environments. Our agents received raw visual inputs and intrinsic rewards and performed actions in the environments. The agents' brains had two components: (i) an intrinsic motivation system that made predictions about the consequences of the agent's action and generated intrinsic rewards based on the prediction error, and (ii) a policy network that learned to select actions based on raw visual inputs and the intrinsic rewards. Both components were constructed using convolutional neural networks.

At each time step of the simulation, the agent received visual input from the environment and performed an action. While interacting with the environment, the agent continuously adjusted its internal model of environmental dynamics to predict how sensory inputs change in response to its own actions. Specifically, the agent learned to predict the next visual observation from the current visual input and its action at that time step. The intrinsic motivation system provided the error between the actual observation and the agent's prediction, as the intrinsic curiosity reward [19, 20]. The agent learned which actions to perform based on maximizing the cumulative intrinsic reward. Because the other agents were the least predictable parts of the environment, we hypothesized that the intrinsic curiosity reward would motivate the agents to be interested in one another and develop imprinting behavior. In the following experiments, we demonstrate that these artificial agents can learn to imprint to other agents (Experiment 1), imprint in complex, naturalistic environments (Experiment 2), and recognize agents based on shape and color cues (Experiment 3).

2.1 Experiment 1

In Experiment 1, we tested whether imprinting emerges when artificial agents are equipped with deep reinforcement learning and intrinsic motivation. To develop imprinting behavior, the agents needed to learn to move, detect and recognize other agents, and approach those agents.

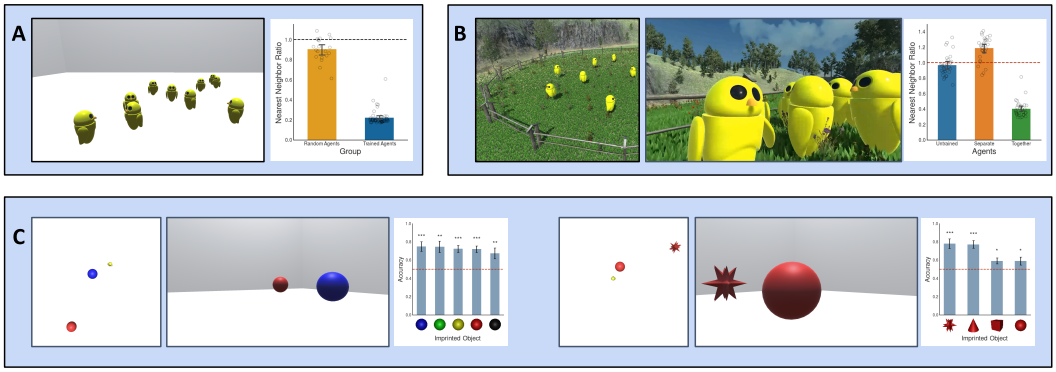


Figure 1: Environments and Results for Experiments 1-3. (A) Experiment 1: overhead view of the simple environment and test performance of random (untrained) and trained agents. (B) Experiment 2: overhead view of the natural environment, a sample agent's first-person view of the environment, and test performance of random (untrained) agents, agents raised (trained) separately, and agent raised (trained) together in a group. Only the agents raised in a group learned to imprint. (C) Experiment 3: overhead view of the environment, a sample agent's first-person view of the environment, and accuracy performance for trained agents, in the color recognition (left panel) and shape recognition (right panel) conditions.

We trained and tested 10 artificial agents simultaneously within a large cubic room with white walls (Fig.1A). The agents' actions were motivated entirely by intrinsic rewards: curiosity and a metabolic cost associated with movements.

After the Training Phase, the model parameters were fixed for the Test Phase (i.e., the agents did not receive any rewards during the Test Phase, and learning was discontinued). This was intended to mimic the critical period of imprinting in animals, in which newborn animals stop developing a preference for new imprinting objects after the first few days of life [5]. If the artificial agents learned to imprint during the Training Phase, then they should show imprinting behavior during the Test Phase, by attempting to reduce the distance between themselves and the other agents.

We measured the agents' imprinting behavior using the nearest neighbor index (NNI)[21]. NNI is an index of spatial dispersion of a group of individuals with NNI <1 indicating more clustered spacing than random distribution and NNI >1 indicating greater dispersion than random distribution.

The results are shown in Figure 1. The NNI was small ($M = 0.22$, $SD = 0.07$; compared to an NNI of 1.0 for a random spatial distribution of agents). This shows that our agents developed grouping behavior. To confirm that the agents developed more pronounced grouping behavior than a group of randomly moving agents, we also tested a group of random, untrained agents. Our trained agents were far more likely to group than random agents (mean difference = 0.63, $SD = 0.11$; independent samples t-test, $t(98) = 29.30$, $p < 10^{-42}$, Cohen's $d = 5.92$).

Experiment 1 indicates that deep reinforcement learning and intrinsic motivation can generate robust imprinting behavior in a simple environment. In the real world, however, newborn animals must learn to recognize group members in complex, natural environments. Natural environments introduce a number of challenges for biological and artificial agents. For instance, natural environments contain a variety of conspicuous features that could distract agents and lead them to imprint to incorrect objects (e.g., rocks or trees). Natural environments also make it more difficult to detect and recognize other agents because agents must be parsed from cluttered backgrounds [22, 23]. Can our artificial agents overcome these challenges and learn to imprint in naturalistic environments?

2.2 Experiment 2

The methods were identical to those used in Experiment 1, except for the environment in which we trained and tested the agents. To mimic the complexity of natural visual environments, we created a virtual grass field with common real-world objects, including trees, mountains, and insects (Fig.1B). Again, the agents learned to imprint, showing pronounced grouping behavior during the Test Phase

(NNI: *M* = 0.40, *SD* = 0.11). The trained agents had lower NNIs than random agents (mean difference = 0.56, *SD* = 0.13; *t*(58) = 17.22, $p < 10^{-22}$, Cohen's *d* = 4.52), confirming that deep reinforcement learning and intrinsic motivation are sufficient to generate imprinting behavior when embodied agents are raised (trained) in naturalistic visual environments.

Next, we examined the necessity of *social experiences* on the development of imprinting behavior in artificial agents. In the experiment described above, the artificial agents were 'reared' in groups with other agents (akin to animals, who are reared in groups with siblings). To explore whether this early social experience is necessary to develop imprinting behavior, we reared a second group of 10 artificial agents in the same naturalistic environment, but without social partners (other artificial agents). During training, these agents acquired ample experience with the naturalistic visual world, but did not acquire visual experience with social partners. After training, we then froze learning in the artificial agents and tested their imprinting behavior in a group (like the agents in the 'reared together' experiment described above). Thus, the agents trained separately were exposed to the same environmental features as the agents trained together, but only the agents trained together received experience with social partners during the Training Phase.

The agents reared separately had NNIs similar to random agents (*M* = 1.19, *SD* = 0.15) and were much more dispersed than the agents reared together (independent samples *t*-test, *t*(58) = 22.90, $p < 10^{-28}$, Cohen's *d* = 6.01). This result confirms that the development of imprinting behavior in artificial agents requires visual experience with social partners early in life, akin to the development of social preferences in newborn animals, who develop preferences for social partners encountered early in life [24–27].

While Experiments 1-2 demonstrate that artificial agents equipped with deep reinforcement learning and intrinsic motivation spontaneously learn behavioral signatures of imprinting in the wild, these experiments differ markedly from the methods used to probe imprinting in laboratory studies. For decades, the main way to test imprinting behavior in the lab has been to present animals with a single object to imprint to and then test their imprinting response with forced-choice trials. In Experiment 3, we tested whether curiosity-driven agents also demonstrate imprinting behavior in a laboratory-like setting that requires recognition based on color and shape features.

2.3 Experiment 3

To mimic laboratory tests of imprinting, we reared artificial agents with a single imprinted object (Training Phase), and then used a two-alternative forced-choice task to test whether the agents developed a preference for the imprinted object over novel objects (Test Phase).

On each test trial, the agent started in the center of the chamber and two objects were placed on opposite sides of the chamber. One object was the imprinted object, and the other object was identical to the imprinted object except for a change in color or shape (c.f., [28]) (Fig.1C). The objects rotated around a vertical axis with a constant angular speed. We measured the proportion of time the agents spent with the imprinted object versus the novel object on each trial. Performance was measured as mean accuracy (percent of time spent by the imprinted object versus the novel object) across the test trials.

The agents performed above chance level for all imprinted objects in both the Color Change (one-sample *t*-tests, all *P*s < 0.01) and Shape Change conditions (one-sample *t*-tests, all *P*s < 0.05). The artificial agents learned to recognize their imprinted object based on both color and shape features simply by being exposed to the object during training. From a machine-learning perspective, this was a challenging task because the agent only received "positive" examples (the imprinted object) during training and never received explicit labels for learning to distinguish the imprinted object from novel objects.

3 Discussion

We have shown that when artificial agents are equipped with deep reinforcement learning and intrinsic motivation, the agents can learn to solve the suite of tasks required for imprinting, including ego-motion, object recognition, and grouping behavior. Like real animals, our agents learned to imprint from raw visual inputs in naturalistic environments. Our agents also learned to imprint without external rewards or supervision, using intrinsic motivation (curiosity) to drive learning.

To our knowledge, this study provides the first pixels-to-actions model of animal imprinting.[1] For decades, researchers across computational neuroscience [31], cognitive science [32], and developmental psychology [33] have argued that task-performing models are essential for gaining a comprehensive and mechanistic understanding of the brain's learning mechanisms. By providing a task-performing (pixels-to-actions) model of animal imprinting, we formalize the mechanisms that underlie the entire learning process, from sensory inputs to behavioral outputs. As a result, these models can be compared against real animals, falsified, and refined.

We also show that two fundamental learning mechanisms—–deep reinforcement learning and intrinsic motivation—–are sufficient for artificial agents to rapidly learn to detect, recognize, and navigate towards social partners. Thus, imprinting behavior can develop in the complete absence of innate, domain-specific learning algorithms, which have long been thought to drive imprinting behavior in animals [8]. Our results provide computationally explicit evidence that it is not necessary to hardwire imprinting behavior into embodied agents to produce imprinting. Rather, generic learning algorithms can drive the development of imprinting. Neither reinforcement learning nor intrinsic motivation were initially designed to produce imprinting behavior in embodied agents. Nevertheless, when these learning algorithms are embodied and permitted to learn in naturalistic environments, animal-like imprinting spontaneously emerges. Imprinting behavior—with all of its survival benefits—may therefore be an emergent property of generic learning mechanisms adapting to the statistics of the natural visual (and social) world.

This finding implies that evolution would not have needed to 'discover' innate, domain-specific learning mechanisms to produce imprinting behavior. Rather, once evolution discovered generic (i.e., domain-general) learning mechanisms, animals would have rapidly learned to imprint during early postnatal development. Consequently, we speculate that domain-general algorithms (e.g., reinforcement learning and intrinsic motivation) provide the computational basis for core social behaviors like imprinting, collective behavior, and social preferences [34]. These generic learning algorithms provide a sufficient computational basis for embodied agents to rapidly learn domain-specific social knowledge through experience.

In sum, we present a pixels-to-actions model of animal imprinting. Artificial agents equipped with deep reinforcement learning and intrinsic motivation can learn the perceptual and cognitive abilities needed to imprint in naturalistic environments. Our results compliment a growing body of work using deep neural networks to model the visual [35], auditory [36], and motor [37] systems. Moreover, our results extend this approach to the study of imprinting: a behavior with deep historical roots in biology, ethology, and psychology.

References

[1] Konrad Z Lorenz. The companion in the bird's world. *The Auk*, 54(3):245–273, 1937.

[2] Lucia Regolin and Giorgio Vallortigara. Perception of partly occluded objects by young chicks. *Perception & psychophysics*, 57:971–976, 1995.

[3] Johan J Bolhuis. Early learning and the development of filial preferences in the chick. *Behavioural brain research*, 98(2):245–252, 1999.

[4] Justin N Wood. Newborn chickens generate invariant object representations at the onset of visual object experience. *Proceedings of the National Academy of Sciences*, 110(34):14000–14005, 2013.

[5] Gabriel Horn. Pathways of the past: the imprint of memory. *Nature Reviews Neuroscience*, 5(2):108–120, 2004.

[6] Antone Martinho III and Alex Kacelnik. Ducklings imprint on the relational concept of "same or different". *Science*, 353(6296):286–288, 2016.

[7] Eckhard H Hess. Imprinting in birds: Research has borne out the concept of imprinting as a type of learning different from association learning. *Science*, 146(3648):1128–1139, 1964.

[8] E. H. Hess. *Imprinting*. Van Nostrand Reinhold, New York, 1973.

[1] See [29] and [30] for early neural network models of imprinting that are not pixels-to-actions models.

[9] Patrick Bateson. Is imprinting such a special case? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 329(1253):125–131, 1990.

[10] Gabriel Horn. Visual imprinting and the neural mechanisms of recognition memory. *Trends in neurosciences*, 21(7):300–305, 1998.

[11] Anton Reiner, David J Perkel, Laura L Bruce, Ann B Butler, András Csillag, Wayne Kuenzel, Loreta Medina, George Paxinos, Toru Shimizu, Georg Striedter, et al. Revised nomenclature for avian telencephalon and some related brainstem nuclei. *Journal of Comparative Neurology*, 473(3):377–414, 2004.

[12] M Kossut and SPR Rose. Differential 2-deoxyglucose uptake into chick brain structures during passive avoidance training. *Neuroscience*, 12(3):971–977, 1984.

[13] . T.V. Smulders, S.W. Newman, and T.J. DeVoogd. Expression of immediate early genes during a food-hoarding task. In *Society for Neuroscience Abstracts*, 1997.

[14] Douglas A Spalding. Instinct. with original observations on young animals. *Eclectic Magazine: Foreign Literature*, 17:424, 1873.

[15] Paul Patrick Gordon Bateson. The characteristics and context of imprinting. *Biological Reviews*, 41(2):177–217, 1966.

[16] Satinder Singh, Richard L Lewis, Andrew G Barto, and Jonathan Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, 2010.

[17] Rachit Dubey and Thomas L Griffiths. Understanding exploration in humans and machines by formalizing the function of curiosity. *Current Opinion in Behavioral Sciences*, 35:118–124, 2020.

[18] Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *International conference on machine learning*, pages 5306–5315. PMLR, 2020.

[19] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[20] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[21] Philip J Clark and Francis C Evans. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453, 1954.

[22] Yuri Ostrovsky, Ethan Meyers, Suma Ganesh, Umang Mathur, and Pawan Sinha. Visual parsing after recovery from blindness. *Psychological Science*, 20(12):1484–1491, 2009.

[23] Samantha MW Wood and Justin N Wood. One-shot object parsing in newborn chicks. *Journal of Experimental Psychology: General*, 150(11):2408, 2021.

[24] Raymond E Engeszer, Michael J Ryan, and David M Parichy. Learned social preference in zebrafish. *Current Biology*, 14(10):881–884, 2004.

[25] Robert F Lachlan, Lucy Crooks, and Kevin N Laland. Who follows whom? shoaling preferences and social learning of foraging information in guppies. *Animal behaviour*, 56(1):181–190, 1998.

[26] Michael G Gaston, Robert Stout, and Roland Tom. Imprinting in guinea pigs. *Psychonomic Science*, 16(1):53–54, 1969.

[27] Eckhard H Hess. Imprinting: an effect of early experience, imprinting determines later social behavior in animals. *Science*, 130(3368):133–141, 1959.

[28] Justin N Wood. Newly hatched chicks solve the visual binding problem. *Psychological science*, 25(7):1475–1481, 2014.

[29] Randall C O'Reilly and Mark H Johnson. Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, 6(3):357–389, 1994.

[30] Patrick Bateson and Gabriel Horn. Imprinting and recognition memory: a neural net model. *Animal Behaviour*, 48(3):695–715, 1994.

[31] Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature neuroscience*, 21(9):1148–1160, 2018.

[32] Allen Newell. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Machine Intelligence*, pages 121–146. Routledge, 2012.

[33] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.

[34] Joshua McGraw, Donsuk Lee, and Justin N. Wood. Parallel development of social preferences in fish and machines. In *Proceedings of the Cognitive Science Society*, 2023.

[35] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[36] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[37] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.

[38] Timothy William Flynn, Sean M Connery, Michael A Smutok, R JORGE Zeballos, and Idelle M Weisman. Comparison of cardiopulmonary responses to forward and backward walking and running. *Medicine and science in sports and exercise*, 26(1):89–94, 1994.

[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

A Supplementary Material

A.1 Virtual environments

We created the virtual environments using the Unity game engine. For Experiment 1, we created an environment called "Simple World." Simple World was a white cubic chamber with 60×60×60 units (L×W×H) walls and a light source at the center of the chamber ceiling. For Experiment 2, we designed an environment called "Realistic World," an open terrain with natural objects including grass, mountains, and trees. Within Realistic World, we created an arena on a grassy field enclosed by a circular fence (diameter: 21 units). At the start of each trial, the agents spawned at random positions within the fenced arena. For Experiment 3, we used a smaller version of Simple World with 30×30×30 units walls.

A.2 Model

We created the artificial agents by embodying artificial brains in virtual animal bodies. The animal bodies were modelled after newborn chicks. Each agent had a yellow body, which fit into a cylinder measuring 3.5 units height and 1.2 units length (radius). Each agent received visual input (96×96 pixel resolution images) through an invisible forward-facing camera attached to its head. The agents could move forward or backward, rotate left or right, or remain stationary. The policy network generated the agent's actions as a pair of discrete variables: translation along the longitudinal axis and rotation around the vertical axis. Each action incurred a metabolic cost to incentivize the agents to move smoothly and efficiently. The action costs were chosen to reflect the fact that backward locomotion elicits a greater metabolic demand than forward locomotion [38]. The costs for forward motion, backward motion, rotation, and idling were -0.001, -0.01, -0.0005, and 0.0 respectively.

The artificial brains contained two biologically inspired components: (i) a policy network that selected actions based on perceived sensory states, and (ii) an intrinsic motivation system that produced rewards based on prediction errors (SI Figure 1). For the feature encoder of the intrinsic motivation network, we used the "Small" CNN (2 convolution layers), and for the feature encoder of the policy network, we used the "Large" CNN (15 convolution layers with skip connections). See SI Figure 2 for details.

The model's intrinsic motivation system was an Intrinsic Curiosity Module (ICM) adapted from [19]. The ICM learned a predictive model of environmental dynamics using self-supervision and produced curiosity rewards based on its prediction errors. The ICM consisted of three neural network components: a feature encoder φ, an inverse dynamics model g, and a forward dynamics model f, parametrized by θ_φ, θ_g, and θ_f, respectively. Given a transition $(s_t,\, a_t,\, s_{t+1})$, the visual observations were encoded into representations $x_t = \varphi(s_t; \theta_\varphi)$ and $x_{t+1} = \varphi(s_{t+1}; \theta_\varphi)$ by the feature encoder. The inverse model learned to predict actions $\hat{a}_t = g(x_t, x_{t+1}; \theta_g)$ from the features of two consecutive observations. The forward model learned to predict the features of the observation at time $t+1$, $\hat{x}_{t+1} = f(x_t, a_t; \theta_f)$, from the features x_t and action a_t at time t. The parameters θ_φ, θ_g, and θ_f were jointly optimized to minimize the weighted average of inverse and forward dynamics prediction errors:

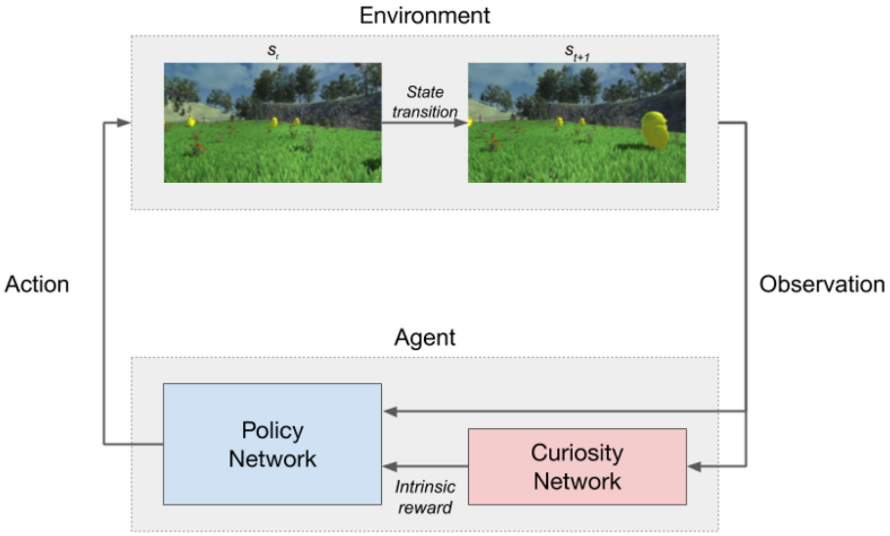
$$(1-\alpha)L_i[g(x_t, x_{t+1}; \theta_g), a_t] + \alpha L_f[f(x_t, a_t; \theta_f), x_{t+1}]$$

where α is a weighting factor, L_i is the negative log-likelihood of the true action a_t according to the inverse model's prediction, and L_f is the squared error between the forward model's prediction and the features x_{t+1}. The forward model's prediction error was multiplied by a scaling factor and provided to the policy network as the intrinsic curiosity reward.

The feature encoder of the ICM was a convolutional neural network (CNN) followed by a fully-connected layer with 128 units. The inverse dynamics model consisted of a hidden layer with 256 hidden units and two output heads to predict translation and rotation actions. The forward dynamics model was a 2-layer MLP with 256 hidden units and 128 output units.

The policy network π, parameterized by θ_p, received visual observation s_t at time t and sampled an action a_t from a stochastic policy $\pi(s_t; \theta_p)$. In our case with a discrete action space, $\pi(s_t; \theta_p)$ was a categorical probability distribution across possible actions. Given a transition (s_t, a_t, s_{t+1}) between t and $t+1$, the reward function was defined as:

$$R_t(s_t, a_t, s_{t+1}) = R_m(a_t) + R_c(s_t, a_t, s_{t+1})$$



Supplementary Figure 1: Our agents received visual observations and performed actions in virtual environments. No external rewards were provided to the agents. The artificial brains contained two components: (1) a curiosity network that produced intrinsic rewards based on prediction errors, and (2) a policy network that selected actions based on perceived sensory states. The curiosity network learned to make predictions about future observations based on the current observation and action. The prediction error was provided to the policy network as an intrinsic reward. The policy network learned a policy (a function mapping an observation to an action) that maximized cumulative intrinsic reward.

where R_m is a metabolic cost of action a_t and R_c is a curiosity reward output by the intrinsic motivation system taking the transition as input. The policy network was optimized to maximize the expected sum of rewards $E[\Sigma_t \gamma^t R_t]$ with discount factor γ.

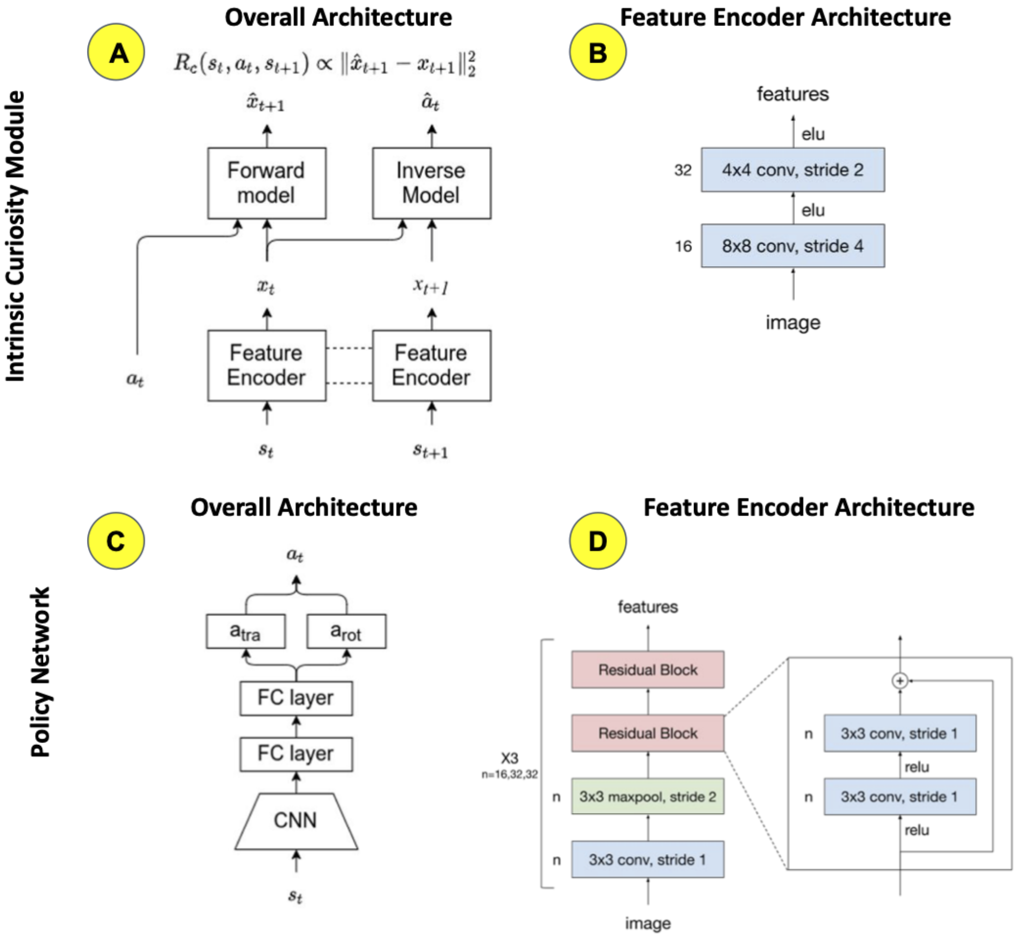
In the policy network, visual inputs were processed through a CNN feature encoder followed by two fully-connected layers with 128 hidden units. Then, the outputs from the last hidden layer were passed into two separate output layers to produce a vector of probabilities for each translation action (forward, backward, no translation) and each rotation action (clockwise, counterclockwise, no rotation).

In each experiment, all of the trained agents had the same brain architecture. However, each agent's brain started with different random initialization of the connection weights within the architecture, and each agent's brain was shaped by the particular visual experiences that the agent acquired during the Training Phase.

We used an off-the-shelf implementation from the Unity ML-Agents Toolkit version 0.10.1. The ML-Agents Toolkit provides a variety of neural network components (e.g., CNNs, recurrent networks) that can be combined together to produce different types of deep reinforcement learning models.

A.3 Training

At the beginning of the Training Phase, the agents were spawned at random positions and orientations within their environment. In Experiment 3, one agent and one object were spawned at random positions and orientations. Each training episode lasted for 1,000 timesteps. The agents were reset at random positions and orientations at the beginning of each episode. We trained the agents for 1,000 episodes in Simple World (Experiments 1 and 3) and 1,200 episodes in Realistic World (Experiment 2). The agents were trained to optimize the sum of their intrinsic curiosity reward and metabolic cost using Proximal Policy Optimization (PPO) [39]. We scaled the intrinsic curiosity module's output (its prediction error) by a factor of 0.1 for the curiosity reward. We used the following hyperparameters in all of our experiments: γ (discount rate) = 0.99, λ (Generalized Advantage Estimate regularization) =



Supplementary Figure 2: Model architectures. (A) The curiosity network consisted of three components: a feature encoder, an inverse model and a forward model. The feature encoder mapped visual inputs onto a feature space. The forward and inverse models operated on the feature space to predict forward and backward dynamics. The intrinsic curiosity rewards were proportional to the prediction errors of the forward model. (B) The Small visual encoders in the curiosity network was a 2-layer convolutional neural network with ELU activations (C) The policy network consisted of a convolutional neural network followed by two fully connected layers and two output heads for translation and rotation actions. (D) The Large visual encoder in the policy network was a 15-layer convolutional neural network with residual connections and ReLU activations.

0.95, β (entropy regularization) = 0.001, batch size = 256, buffer size = 2560, learning rate = 0.001. The learning rate decayed linearly, reaching 0 at the end of training. This decay in learning rate across the Training Phase was designed to mimic the critical period found in filial imprinting.

A.4 Testing

After training, the neural network weights were fixed, so the networks did not continue to learn during the Test Phase. Each test episode consisted of 2,000 timesteps. We tested our agents for 50 episodes in Experiment 1, 30 episodes in Experiment 2, and 35 episodes per test condition in Experiment 3. At every timestep, we recorded the position of each agent in X,Y coordinates for computing the Nearest Neighbor Index (NNI).

A.5 Effects of different architectures

To assess whether imprinting can also develop in a variety of other artificial brains, we also conducted a follow-up experiment where we varied four hyperparameters in the artificial brain: (i) the architecture of the policy network feature encoder (2-layer, 4-layer, or 15-layer neural network), (ii) the architecture of the intrinsic motivation network feature encoder (2-layer, 4-layer, or 15-layer neural network), (iii) the strength of the intrinsic reward (.01, 0.1, or 1.0), and (iv) the resolution of the retinal input (64×64, 96×96, or 128×128 pixel resolution). By varying these hyperparameters, we could explore whether imprinting emerges in small, medium, and large neural networks and study how different components of the model shape imprinting behavior. All of the agents were trained and tested in the realistic environment from Experiment 2, since this environment most closely resembled the environments encountered by animals in nature.

Two notable findings emerged. First, the agents learned to imprint with all but one of the artificial brains (independent samples *t*-tests versus random agents, $P\text{s} < 10^{-12}$). Even small neural networks were sufficient to produce imprinting behavior, perhaps explaining why animals with relatively small brains (e.g., chicks, fish, insects) still develop grouping behavior. Second, the only architecture in which imprinting did not emerge was the low curiosity architecture. Agents with the lowest curiosity strength (0.01) showed no evidence of imprinting (independent samples *t*-tests versus random agents, $t(58) = 0.45$, $p = .65$, Cohen's $d = 0.12$), whereas the agents with moderate (0.1) and strong (1.0) curiosity strength showed robust evidence of imprinting ($P\text{s} < 10^{-22}$). Thus, artificial agents need moderate to strong curiosity to develop imprinting behavior in this class of model architectures. This finding suggests, perhaps counterintuitively, that curiosity (seeking novelty) drives imprinting (a preference for the familiar). Note, however, that curiosity-driven deep reinforcement learning must be subject to a critical period to produce imprinting behavior; if learning stays "turned on," the agents will continue to seek novel experiences.