Intrinsically Motivated Social Play in Virtual Infants

Chris Doyle
Stanford University
crd@stanford.edu

Sarah Shader
Pomona College
sdsb2018@mymail.pomona.edu

Michelle Lau
Stanford University
mblau@stanford.edu

Megumi Sano
Stanford University
megsano@stanford.edu

Daniel L. K. Yamins
Stanford University
yamins@stanford.edu

Nick Haber
Stanford University
nhaber@stanford.edu

Abstract

Infants explore their complex physical and social environment in an organized way. To gain insight into what intrinsic motivations may help structure this exploration, we create a virtual infant agent and place it in a developmentally-inspired 3D environment with no external rewards. The environment has a virtual caregiver agent with the capability to interact contingently with the infant agent in ways that resemble play. We test intrinsic reward functions that are similar to motivations that have been proposed to drive exploration in humans: surprise, uncertainty, novelty, and learning progress. The reward functions that are proxies for novelty and uncertainty are the most successful in generating diverse experiences and activating the environment contingencies. We also find that learning a world model in the presence of an attentive caregiver helps the infant agent learn how to predict scenarios with challenging social and physical dynamics. Our findings provide insight into how curiosity-like intrinsic rewards and contingent social interaction lead to social behavior and the creation of a robust predictive world model.

1 Introduction

Infants are born into a complex set of social and physical phenomena. Infants must figure out how to control their bodies and learn how people and objects respond to their actions. Infants' exploration is not random, they explore their world in a structured way [Gopnik et al., 1999].

A compelling hypothesis is that the motivation to explore may be linked to a desire to improve the accuracy of predictions about the world. Working to improve these predictions (the agent's "world model") can create a self-generated learning curriculum, through a cycle of evaluating deficiencies in the model, seeking out information, updating the model, and gaining new capabilities [Schmidhuber, 2010]. Researchers have found evidence that suggests violations of expectation catalyze learning [Stahl and Feigenson, 2015], and that learning progress is an important component for task selection [Ten et al., 2021]. Children appear sensitive to the discriminability of hypotheses and explore longer when hypotheses are harder to distinguish [Siegel et al., 2021]. Stimulus novelty may also play a role in curiosity-driven exploration [Poli et al., 2022]. Children can effectively explore diverse scenarios, including both physical and social phenomena. Intrinsic reward functions implemented in reinforcement learning contexts are more fragile and can be susceptible to white-noise fixation [Oudeyer et al., 2007, Schmidhuber, 2010, Pathak et al., 2017], or may not lead to meaningful behavior diversity.

Previous work showed that intrinsic rewards in virtual agents lead to exploration in a physical context [Haber et al., 2018] and a preference for viewing animate objects in a protosocial context [Kim et al., 2020], but it did not include complex social contingencies or a sophisticated embodiment for the

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

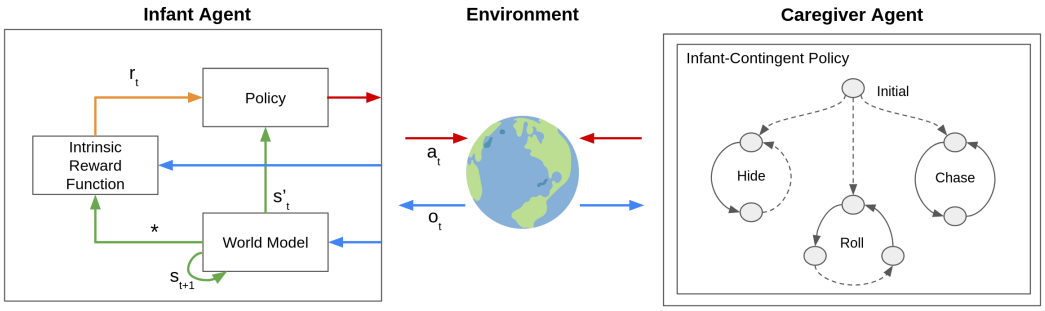

Figure 1: The infant agent's world model, reward function, and policy interact to drive infant actions over time. The intrinsic reward function takes different inputs depending on the choice of function. The state diagram (right) outlines the caregiver's static policy. Dotted lines indicate transitions that depend on the infant state, and solid lines indicate those that do not. The top circle is the starting state, where the caregiver waits for the infant to point, and the three branches are unlocked by the infant pointing at one of the objects in the room.

agent. We extend the work in these directions to evaluate if a curiosity-like intrinsic reward function can generate social behavior in virtual agents and to examine the effect of contingency on how a virtual agent learns social and physical dynamics.

Contributions (1) We introduce a developmentally-inspired virtual 3D environment with an embodied infant agent and a caregiver agent that can engage in complex, contingent, social behaviors. (2) We show that through reinforcement learning, the intrinsically-motivated infant agent generates temporally variable, social, play-like behavior within our environment, in the absence of extrinsic reward. When motivated by a novelty reward or an uncertainty reward, the agent builds world models that can make good predictions about experiences beyond their own. (3) We show that a high level of contingency in the caregiver agent corresponds with the infant agent learning to make better predictions about challenging scenarios involving caregiver and object dynamics.

2 Environment and infant agent

The 3D virtual environment contains an infant agent, a caregiver agent, and two movable balls. The infant has a body and two arms, each having a shoulder and elbow joint. The infant is controlled through a set of discrete actions. Its observations include proprioceptive information and an object-centric representation of the environment, limited by its field of view. The caregiver is controlled by a script. At the beginning of each episode, the caregiver waits for the infant to point to an object in the room. Based on the object pointed to, the caregiver initiates an activity: "hide and seek," "roll to infant," or "chase the ball." Details about the environment and behaviors are in Appendix A.

The infant agent is a reinforcement learning agent that has three major components: a forward predictive world model, an intrinsic reward function, and a policy (Figure 1). Each agent has one of five reward functions: Adversarial [Achiam and Sastry, 2017], Disagreement [Pathak et al., 2019], Random Network Distillation (RND) [Burda et al., 2018], δ-progress [Graves et al., 2017], and γ-progress [Kim et al., 2020]. These correspond to "surprise," "uncertainty," "novelty," and two versions of "learning progress." Details about the agent and intrinsic rewards are in Appendix B.

3 Experiment 1: Compare exploration across intrinsic reward functions

We investigate two questions: what type of behavior diversity arises from different reward functions and which reward functions lead the infant agent to learn a robust world model.

3.1 Method

Infant agents are assigned one of the intrinsic reward functions and trained for 20M steps. Three random seeds are run for each reward function. After training, we evaluate the agent's behavior diversity and world model. To assess behavior diversity, we measure (1) the entropy of different components of infant observations, (2) the frequency that the infant activates contingencies built into

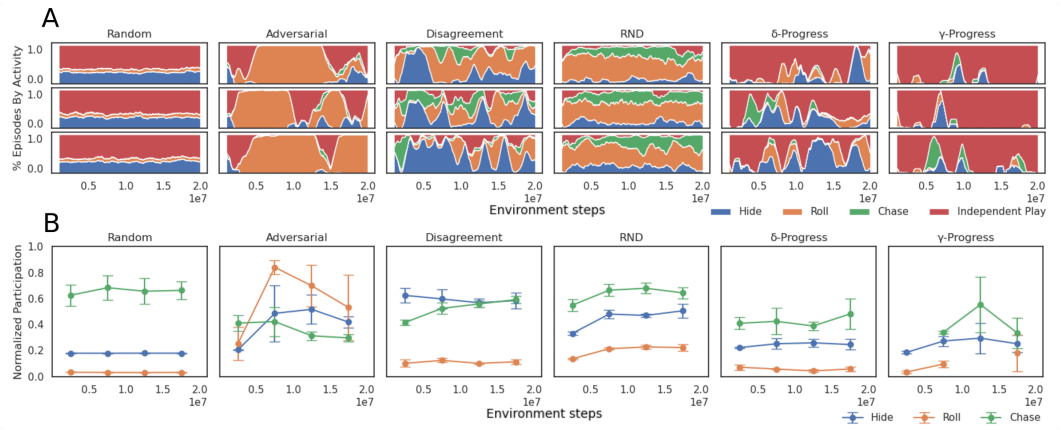

Figure 2: A. The proportion of episodes that the infant agent activates one of the contingent behaviors (Hide, Roll, Chase) or if they do not activate a behavior (Independent Play). Three random seeds are shown for each reward function. B. The average participation by behavior is calculated across 5M steps. The standard error is shown across the three random seeds. Values are normalized by dividing them by the 99th percentile of the metric across training.

the caregiver agent, and (3) the level of participation in the contingent caregiver activities. Details of this evaluation are in Appendix C. To assess the world model, we measure world model prediction error on sets of validation cases. Details of the world model evaluation are in Appendix D.

3.2 Results

Disagreement and RND yield the greatest diversity of experience and the most robust world models Disagreement and RND generated higher entropy (Table 1) than all other intrinsic reward functions across the Location, Orientation, and Attention components of state. The signals also generated a larger number of total contingency activations (Table 1) than all the other signals. The world models from Disagreement and RND agents have lower prediction error on validation cases than the world models from other agents (Figure 5A).

Although RND and Disagreement both generate a high diversity of states, the behaviors have different temporal structures (Figure 2A). RND has a relatively stable split of behavior activations, whereas Disagreement yields varying proportions of behavior activations over time. Reward functions seem to generate behavior patterns that are distinguishable from each other and consistent across seeds.

Agents using the Adversarial reward function focus on hitting the ball in the Roll behavior Agents driven by the adversarial signal frequently activate the Roll behavior. Within the Roll behavior the agent spends most of the time hitting the ball back and forth between its arms. This behavior contributes to the lower entropy observed in Table 1 and the high Roll participation in Figure 2B.

When the world models of agents with different intrinsic motivation signals are evaluated on validation cases from the Adversarial agent's experiences, the prediction error is high (Figure 5B). The prediction error is high even for the Adversarial agent that collected and trained on the experiences. The agent found situations that are difficult to model accurately with the current world model class.

4 Experiment 2: Vary the frequency the caregiver responds to the infant

Social interaction is critical for typical development in children [Kaler and Freeman, 1994]. We use our simulated environment to analyze how the infant agent's understanding of the world is affected when it interacts with a less attentive caregiver compared with a more attentive one.

4.1 Method

Infant agents are trained for 10M steps using the Disagreement reward function, selected based on the high state entropy and low world model prediction error observed in Experiment 1. The analysis is limited to one reward function due to computational constraints. In contrast to Experiment 1, the

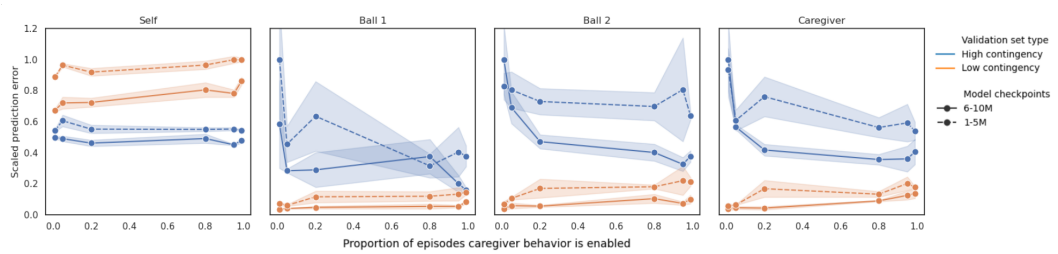

Figure 3: Prediction error by observation component shown plotted against the proportion of time caregiver contingency is enabled during training. Lower values indicate better accuracy. The error is decomposed into that due to infant orientation, position, and arm configuration ("Self"), Ball 1, Ball 2, and Caregiver predictions. High contingency refers to the set of validation cases created from agents where caregiver contingency was enabled 95% of the time or more, and low contingency refers to where caregiver contingency was enabled 5% of the time or less. The results are shown for world model checkpoints taken from 1M to 5M training steps and 6M to 10M training steps and plotted with the standard error across three seeds.

caregiver response to the infant "pointing" is stochastic. A flag is set with some probability at the beginning of each episode to determine if the caregiver will respond to the infant. In a training run, the probability is fixed to be one of: 1%, 5%, 20%, 80%, 95%, or 99%. We refer to 95% and 99% as high-contingency (HC) and 1% and 5% as low-contingency (LC).

We create validation sets from the experiences of infants trained with HC and LC caregivers and test world models trained with different levels of contingency on those sets. We decompose the prediction error into that due to predictions about Ball 1, Ball 2, the Caregiver, and the "Self," which includes infant orientation, position, and arm configuration.

4.2 Results

Increasing caregiver contingency shifts prediction difficulty from proprioceptive inputs to external dynamics In the absence of very frequent caregiver interaction, agents focus on proprioceptive exploration. This manifests as the "Self" component being harder to predict in validation cases with LC caregivers than with HC caregivers. When caregiver interaction is frequent, the caregiver facilitates challenging external dynamics scenarios through its complex behavior patterns. A HC caregiver corresponds with more difficult to predict ball and caregiver components (Figure 3).

High levels of caregiver contingency have a positive, asymmetric effect on world model accuracy In an environment with a high level of contingency, world model performance improves more on HC validation sets than it deteriorates on LC validation sets. There is a decrease in error on the Ball 1, Ball 2, and Caregiver components on HC validation sets as the level of contingency increases. The caregiver agent facilitates challenging scenarios, which provide valuable experiences for the infant agents to learn from. In these same components, there is an increase in loss on LC validation sets with higher levels of caregiver contingency, but that increase is small on an absolute basis (Figure 3).

5 Discussion and future work

A human infant's ability to cause change in the world is amplified by an attentive caregiver, who responds to coos, cries, reaches and points. In our environment the caregiver is likely to facilitate the infant agent's first experience seeing a ball move. This amplification is visible in our results: the prediction difficulty of external dynamics increases with a highly-contingent caregiver. Caregivers can also provide a dense intrinsic reward. Social behavior is a difficult prediction problem, so each interaction will yield more data to challenge the infant's understanding of the world. For signals that depend on a world model, this will motivate exploration. Our results support this: multiple intrinsic reward functions frequently activate caregiver behaviors.

In this environment basic social interaction and contingency activation can arise without requiring a specific module, social intrinsic reward, or extrinsic reward. It can arise from curiosity, implemented as an information-maximizing motivation. However, there is evidence that humans do have prosocial

biases. We know that neonates prefer face-like patterns [Reynolds and Roth, 2018] and some argue that humans are equipped with unique social motivations [Tomasello, 2019].

Our work raises questions around how curiosity may interact with social motivation and if social motivation may be necessary for effective exploration in some environments. Further experiments could help explore these issues. One limitation of the current environment is that much of the complexity is in the social interactions. By introducing an environment with multiple, similarly complex asocial and social interactions, we could investigate the performance of various curiosity-like motivations in the presence of asocial distractors and characterise the extent to which agents attend to social vs. asocial phenomena, similar to the analysis done in Kim et al. [2020].

An important extension of this work is to compare artificial and human trajectories on matched environments. Real infant walking trajectories have been analyzed to understand exploration patterns and state coverage [Hoch et al., 2019]. A direct comparison to artificial agents may lead to (1) a better characterisation of infant exploratory motivations and patterns and (2) a better understanding of the limitations of intrinsically motivated reinforcement learning agents.

References

J. Achiam and S. Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

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

M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.

A. Gopnik, A. N. Meltzoff, and P. K. Kuhl. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co, 1999.

A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. PMLR, 2017.

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.

J. E. Hoch, S. M. O'Grady, and K. E. Adolph. It's the journey, not the destination: Locomotor exploration in infants. *Developmental science*, 22(2):e12740, 2019.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar, et al. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*, 2018.

S. R. Kaler and B. Freeman. Analysis of environmental deprivation: Cognitive and social development in romanian orphans. *Journal of Child Psychology and Psychiatry*, 35(4):769–781, 1994.

S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*, 2018.

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.

P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

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.

J. Piaget, M. Cook, et al. *The origins of intelligence in children*, volume 8. International Universities Press New York, 1952.

F. Poli, M. Meyer, R. B. Mars, and S. Hunnius. Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition*, 225:105119, 2022.

G. D. Reynolds and K. C. Roth. The development of attentional biases for faces in infancy: A developmental systems perspective. *Frontiers in psychology*, 9:222, 2018.

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

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020.

M. H. Siegel, R. W. Magid, M. Pelz, J. B. Tenenbaum, and L. E. Schulz. Children's exploratory play tracks the discriminability of hypotheses. *Nature communications*, 12(1):3598, 2021.

A. E. Stahl and L. Feigenson. Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230):91–94, 2015.

A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

H. Tang, R. Houthooft, D. Foote, A. Stooke, O. Xi Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

A. Ten, P. Kaushik, P.-Y. Oudeyer, and J. Gottlieb. Humans monitor learning progress in curiosity-driven exploration. *Nature communications*, 12(1):5972, 2021.

M. Tomasello. *Becoming human: A theory of ontogeny*. Harvard University Press, 2019.

6 Appendix A: Environment Details

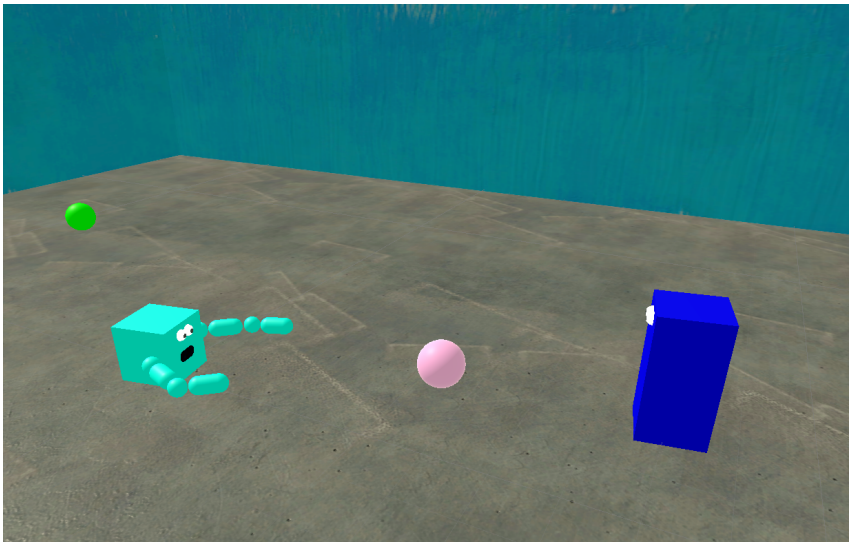

Figure 4: The environment is a room that contains an infant agent (teal), caregiver agent (dark blue), and two movable balls (pink and green)

Our 3D virtual environment is created in Unity and uses the ML-Agents framework [Juliani et al., 2018]. We use episodes of 2,000 timesteps over 200 in-environment seconds. At the end of an episode, the environment is reset to its starting state.

The setting is a closed room containing two ball objects, a caregiver agent, and an infant agent, pictured in Figure 4. The Unity physics engine allows the objects to respond to forces applied to them by the infant's body and arms. The balls can also be picked up and thrown by the caregiver.

6.1 Infant

The infant has two arms with shoulder and elbow joints. The arms can only move in the plane parallel to the floor and are at a height they can collide with the ball. At each timestep, the infant can choose one of 13 actions: do nothing, turn left/right, move forward/back, or rotate any of the four arm joints clockwise/counterclockwise.

The infant has a partial view of the environment: it receives an indicator as to whether each object is in its field of view (120°forward), and if the object is in view, its position, orientation, and velocity. It receives proprioceptive information giving the positions and orientations of its arms and its body, and the value of a hit sensor on each arm.

6.2 Caregiver

The caregiver agent can move around the room and pick up and throw the balls. It is controlled by a script that begins each episode watching the infant agent and waiting for the infant to "point" to an object. Pointing is determined by the infant orienting their body toward an object, with an arm pointed straight forward, and holding that position for five timesteps. If the infant points toward an object or the caregiver, a branch of the script is activated. Pointing toward the caregiver activates the "hide and seek" branch (Hide), pointing toward the pink ball activates the "roll to infant" branch (Roll), and pointing toward the green ball activates the "chase the ball" branch (Chase). At the end of an episode, the environment is reset and the caregiver waits for the infant to point again. The high-level state diagram is shown in Figure 1.

6.2.1 Hide and seek

The caregiver selects a point in the area behind the infant and moves there. When it arrives, it waits for the infant to look in its direction, at which point the caregiver selects a new point to move to behind the infant.

6.2.2 Roll to infant

The caregiver retrieves the pink ball (Ball 1), moves a target distance from the infant, then looks at the infant. The caregiver waits for the infant to look in its direction. Once that occurs, the caregiver rolls the ball to the infant and waits for a fixed period before retrieving the ball and starting again.

6.2.3 Chase the ball

The caregiver continually retrieves the green ball (Ball 2) and throws it forward. This cycle causes the ball to be thrown around the room, bouncing off the walls and floors.

6.2.4 Independent play

If no object is pointed toward, the caregiver will remain looking at the infant for the entire episode.

7 Appendix B: Infant Design

The infant agent has three primary components that drive its behavior over time: a world model, an intrinsic reward function, and a policy (Figure 1). The components are intended to be only high level analogies to human capabilities, they do not represent a mechanistic hypothesis. We choose implementations that are computationally tractable and have been successful in various reinforcement learning contexts.

7.1 World model

The objective of the infant agent's world model is to predict the next observation given the history of observations and actions. We create a latent dynamics model that attempts to model changes in the underlying the environment state.

7.1.1 Architecture

The world model uses a two-layer LSTM [Hochreiter and Schmidhuber, 1997]. In addition to maintaining the hidden states of h and c in the LSTM, the world model maintains a belief state b that contains an estimate of position, orientation, and velocity for all objects in the environment. The b component has a dimensionality equal to that of an observation o, less the object-visibility indicator dimensions. The augmentation allows the infant agent to maintain a belief about objects outside its visual frame, a capability that humans have [Piaget et al., 1952]. We refer to the tuple of world model states (h, c, b) as s. Changes to b are predicted using an MLP decoder on h. Delta prediction for physical quantities has been successful in fully-observed physics prediction [Battaglia et al., 2016, Chang et al., 2016] and we adapt it for a partially observed setting.

7.1.2 Training

The world model is supervised on rollouts of length $L = 30$. We use a stored hidden state and burn-in to help prediction accuracy [Kapturowski et al., 2018]. The world model is recurrently applied $s_{t+1} \leftarrow f_\theta(s_t, a_t)$ using the action sequence from the replay buffer. Predicted observations $\hat{o}$ are directly read out of the b state. The world model loss is the square error of the visible components of the observation over the rollout.

$$\mathcal{L}(\hat{o}_{1\ldots L}, o_{1\ldots L}) = \sum_{i=1}^{L} \left(\sum_{j=1}^{\dim(o_i)} (\hat{o}_{i,j} - o_{i,j})^2 \mathbb{1}_{\text{visible}}(o_{i,j}) \right) \tag{1}$$

Algorithm 1 Agent algorithm

Input total episodes E, episode length T, world model training iterations per episode M, batch size N, sequence training length L, intrinsic reward function $\mathcal{R}$
Initialize replay buffer R = $\emptyset$, parameters for world model θ, policy ϕ, and intrinsic reward ψ
for episode $= 1, 2, \ldots E$ **do**
    Initialize belief b and LSTM hidden states (h, c)
    for $t = 1, 2, \ldots T$ **do**
        Observe o_t
        Update s_t to s'_t with information from o_t
        $a_t \sim \pi_\phi(a|s'_t)$
        $s_{t+1} \leftarrow f_\theta(s'_t, a_t)$
        Take action a_t
    end for
    Add collected tuples of (o, a, s) to replay buffer R
    Calculate reward r_t for steps $1...T$ using $\mathcal{R}$
    Update ϕ with PPO, update ψ as applicable
    for $i = 1, 2, \ldots, M$ **do**
        Sample N sequences with length L from R
        Calculate $\mathcal{L}_{\text{WM}}$ on batch and update θ
    end for
end for

7.2 Intrinsic Reward Functions

7.2.1 Adversarial

A violation of expectation can be framed as the prediction from a world model being significantly different from the observed outcome. This surprise-based intrinsic reward can be formulated as a function of the prediction error [Achiam and Sastry, 2017, Pathak et al., 2017, Schmidhuber, 2010]. We use the model loss as the reward.

7.2.2 Disagreement

Being uncertain about the outcome of an action can be interpreted as there being variance around a prediction of the future. Uncertainty has been formulated as the variance of predictions across an ensemble of trained world models [Pathak et al., 2019, Sekar et al., 2020]. We use $K = 10$ models in the ensemble. Because of memory and training time constraints, the recurrent dynamics model is not replicated. Instead, the ensemble members are MLPs that predict the next observation from the state s and action.

7.2.3 Random Network Distillation (RND)

Novel stimuli can indicate the potential for learning. In environments with a discrete state space, a reward that is a decreasing function of visit counts can be effective in incentivizing exploration [Strehl and Littman, 2008]. Although that approach is not directly applicable to continuous spaces, methods such as pseudo-counts [Tang et al., 2017] and Random Network Distillation [Burda et al., 2018] can be used.

7.2.4 Learning Progress

An agent may try to pursue experiences that are likely to improve its understanding of the world. One approach to estimating this is to evaluate recent learning progress on that topic, that is, the magnitude of improvement between a previous world model and the current one. This has been implemented as δ-progress [Achiam and Sastry, 2017, Graves et al., 2017] and γ-progress [Kim et al., 2020]. The difference between the methods is how the previous world model is defined: δ-progress uses a world model from δ steps ago and γ-progress updates the weights of the old model by performing a weighted average of old weights with current weights.

7.3 Policy learning

We modify Proximal Policy Optimization (PPO) [Schulman et al., 2017], a model-free reinforcement learning algorithm, to have the learned policy be based on the world model state s instead of observations.

8 Appendix C: Behavior Analysis

We consider behavior diversity in three ways: state coverage, social contingency activation, and level of contingency participation.

To evaluate state coverage, we independently consider four components of the infant agent's observations: its location within the room, its orientation, its pose, and what objects, animate or inanimate, are visible to it. We calculate the normalized entropy as the entropy of a discretized version of the component relative to a uniform distribution (Table 1).

Activating and participating in the social contingencies is particularly important because it allows the infant agent to unlock new parts of the state space. We look at the proportion of episodes where the behavior is activated over time by seed (Figure 2A) and in aggregate (Table 1). For each of the activated behaviors we identified a metric that corresponds to "participation" in the activities: within the Hide behavior, the number of times the infant finds the caregiver; within the Roll behavior, the number of times the infant hits the ball; within the Chase behavior, the frequency that the infant is looking at the caregiver when the ball is thrown. Participation is shown in Figure 2B.

Table 1: Comparison of Normalized Entropy and Behavior Activation. Mean and standard error over 3 seeds.

	Normalized Entropy				Behavior Activation			
Agent	Location	Orientation	Pose	Attention	Hide	Roll	Chase	Total
Random	5 ± 0	56 ± 0	100 ± 0	62 ± 0	30 ± 0	9 ± 0	0 ± 0	39 ± 0
Adversarial	45 ± 5	79 ± 2	76 ± 4	71 ± 1	7 ± 2	64 ± 6	2 ± 0	73 ± 4
Disagreement	93 ± 0	100 ± 0	99 ± 0	95 ± 0	38 ± 5	35 ± 7	14 ± 4	87 ± 3
RND	87 ± 1	98 ± 0	99 ± 0	93 ± 0	15 ± 2	53 ± 2	23 ± 1	91 ± 1
δ-Progress	40 ± 7	88 ± 1	94 ± 2	80 ± 1	23 ± 6	14 ± 2	4 ± 2	42 ± 7
γ-Progress	38 ± 4	80 ± 3	90 ± 3	74 ± 1	7 ± 1	3 ± 0	4 ± 3	14 ± 3

9 Appendix D: World Model Evaluation

9.1 World model performance evaluation

We assess the robustness of a world model by evaluating its predictions on trajectories it has not been trained on. We test it on experiences collected by agents with different seeds and different intrinsic reward functions, and on experiences collected by manually programmed agents.

For each agent, we create a set of validation cases from its lifetime experience by uniformly sampling 2000 trajectory segments from its history. In a round-robin fashion, we test the world model from each agent against the validation case sets for each other agent, including different seeds and different intrinsic reward functions. We score the model on each validation case set by calculating the average total model loss over a 10-step rollout. The average loss on the validation sets from other agents is the "Agent-Generated" validation case set loss. The "Manually-Generated" validation case set is created from the history of an agent that runs a fixed script. The script directs it to find a ball, move toward it, and hit it with an arm. The world models are scored on these validation cases to calculate the "Manually-Generated" validation case set loss. Both the "Agent-Generated" and "Manually-Generated" losses are in Figure 5A.

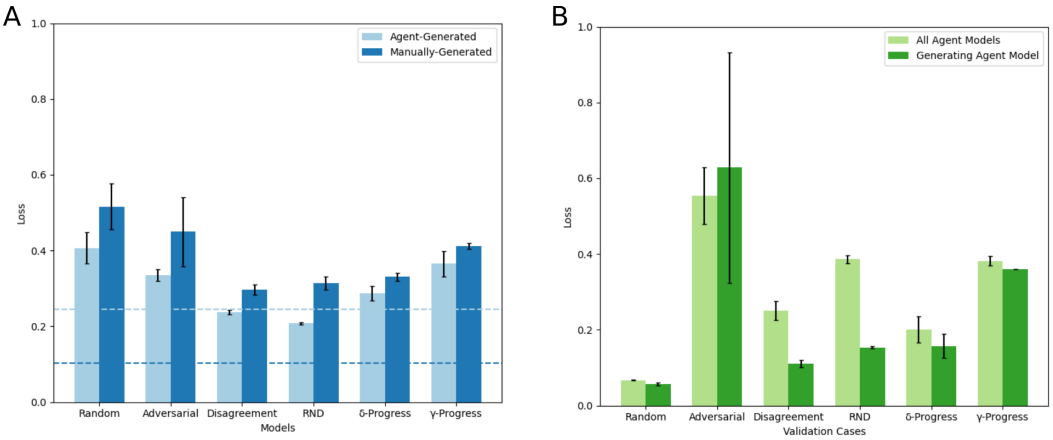

Figure 5: A. World model loss on validation cases generated from the experience of other agents and validation cases that were created manually. Lower values indicate better accuracy. Agent-Generated cases include validation sets from all seeds of all intrinsic reward functions. The horizontal lines are the average loss if each validation case is predicted using the world model from the agent that generated the data (an estimate of good performance in our model class). B. World model loss on a validation case set. All Agent Models is the average loss on a set across all agents. Generating Agent Model is the loss on the set using the world model from the agent that generated the set.