Developmental Curiosity and Social Interaction in Virtual Agents

Anonymous CogSci submission

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. These generic reward functions lead the infant agent to explore its environment and discover the contingencies that are embedded into the caregiver agent. 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. Taken together, our findings provide insight into how curiosity-like intrinsic rewards and contingent social interaction lead to dynamic social behavior and the creation of a robust predictive world model.

Keywords: curiosity; intrinsic motivation; world models; reinforcement learning; contingency; development

Introduction

Infants are born into a complex set of social and physical phenomena. At the center of their world are caregivers who smile at them, change their diapers, point at things, and sing songs, and around them there are bouncing balls, falling block towers, and spinning tops. Infants must figure out how to control their bodies and learn how the world responds to their actions. Infants' exploration of this rich environment is not random, they explore their world in a structured way (Gopnik, Meltzoff, & Kuhl, 1999).

Over time, children develop an understanding of their world. Infants are sensitive to social contingency, the reactions of others to their actions (Nadel, Carchon, Kervella, Marcelli, & Réserbat-Plantey, 1999) and the level of responsiveness of a partner (Bigelow & Rochat, 2006). Infants have expectations about how people will respond to their actions (Tronick, Als, Adamson, Wise, & Brazelton, 1978) and how objects will behave (Stahl & Feigenson, 2015).

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 & Feigenson, 2015), and that learning progress is an important component for task selection (Ten, Kaushik, Oudeyer, & Gottlieb, 2021). Children appear sensitive to the discriminability of hypotheses and explore longer when hypotheses are harder to distinguish (Siegel, Magid, Pelz, Tenenbaum, & Schulz, 2021). Stimulus novelty may also play a role in curiosity-driven exploration (Poli, Meyer, Mars, & Hunnius, 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, Kaplan, & Hafner, 2007; Schmidhuber, 2010; Pathak, Agrawal, Efros, & Darrell, 2017), or may not lead to meaningful behavior diversity.

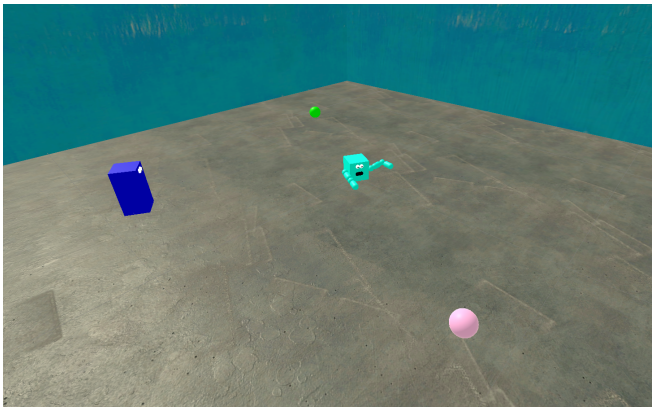

Figure 1: The environment is a room that contains a infant agent (teal), caregiver agent (dark blue), and two movable balls (pink and green)

Previous work showed that intrinsic rewards lead to exploration in a physical context (Haber, Mrowca, Wang, Fei-Fei, & Yamins, 2018) and a preference for viewing animate objects in a protosocial context (Kim, Sano, De Freitas, Haber, & Yamins, 2020), but it did not include complex social contingencies or a sophisticated embodiment for the 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.

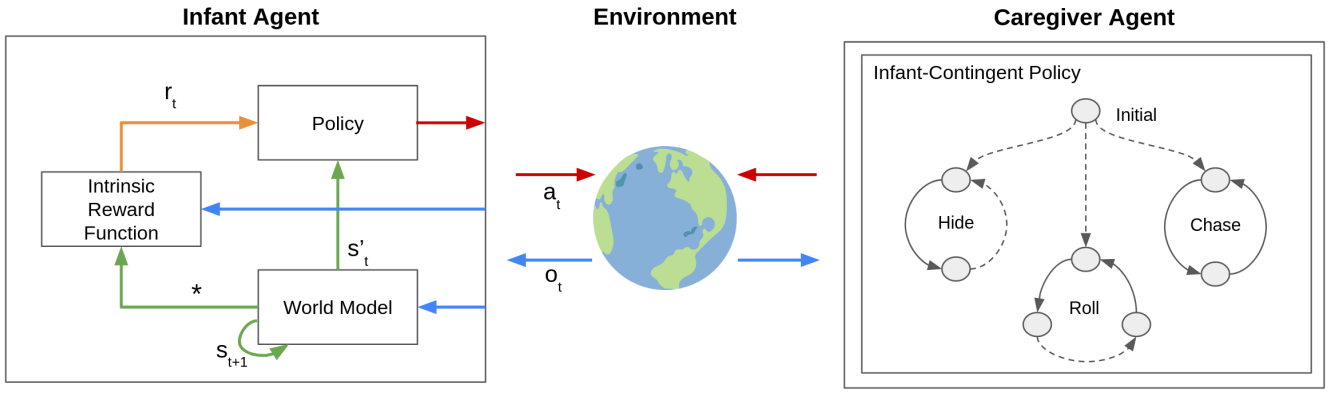

Figure 2: The infant agent's world model, reward function, and policy interact to drive infant actions over time. The intrinsic reward function takes different inputs depending on the choice of function. The state diagram (right) outlines the caregiver's static policy. Dotted lines indicate transitions 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.

We summarize our contributions as follows.

- 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 with the infant.
- We describe an infant agent with an intrinsic reward function inspired by motivations hypothesized to be present in humans. The infant agent learns to pursue intrinsic rewards through reinforcement learning. We show that the 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.
- 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.

Environment

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 1. 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.

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 partial observability: 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.

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 2.

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.

Roll to infant The caregiver retrieves the pink ball, 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.

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

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 $1 \ldots T$ 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

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

Infant Agent

The infant agent has three primary components that drive its behavior over time: a world model, an intrinsic reward function, and a policy (Figure 2).

World model

The objective of the infant agent's world model is to accurately 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.

Model Architecture The world model uses a two-layer LSTM (Hochreiter & Schmidhuber, 1997). We supplement the hidden states of h and c in the LSTM with a hidden state b. The belief state b contains an estimate of position, orientation, and velocity for each object, whether the object is currently in view or not. Together we refer to the combination of world model hidden 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, Pascanu, Lai, Jimenez Rezende, et al., 2016; Chang, Ullman, Torralba, & Tenenbaum, 2016) and we adapt it for a partially observed setting.

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, Ostrovski, Quan, Munos, & Dabney, 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. The world model loss is the square error of the visible components of the observation over the length of the rollout.

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

Intrinsic reward functions

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. We use the model loss as the reward. (Achiam & Sastry, 2017) (Pathak et al., 2017) (Schmidhuber, 2010).

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, Gandhi, & Gupta, 2019) (Sekar et al., 2020). Our implementation has $K = 10$ 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 current state and action.

Random Network Distillation (RND) Novel stimuli can indicate the potential for learning. In environments with a discrete state space, using a reward that is a decreasing function of visit counts can be effective in incentivizing exploration (Strehl & Littman, 2008). However, that approach is not directly applicable to continuous state spaces. Approaches for continuous spaces include pseudo-counts (Tang et al., 2017) and Random Network Distillation (Burda, Edwards, Storkey, & Klimov, 2018).

Learning Progress An intuitive learning strategy is to pursue experiences that are likely to improve the agent's 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 & Sastry, 2017; Graves, Bellemare, Menick, Munos, & Kavukcuoglu, 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 the old model weights with the current weights. We evaluate both reward functions.

Policy learning

We modify Proximal Policy Optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017), a model-free reinforcement learning algorithm, to have the learned policy be based on the world model state s instead of observations.

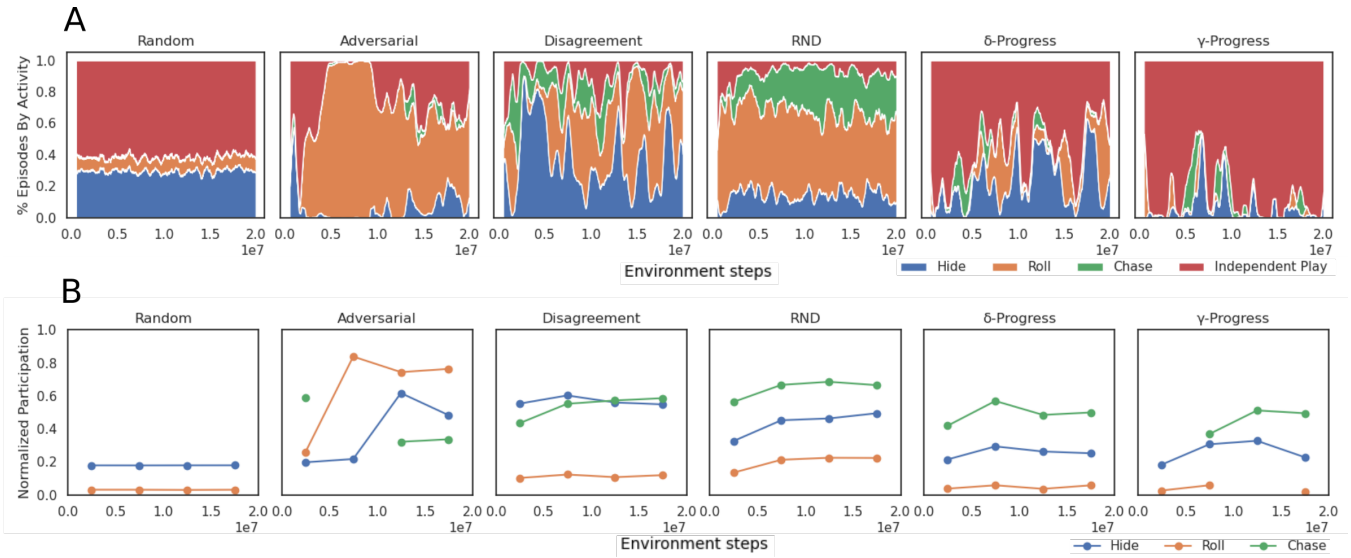

Figure 3: A. The proportion of episodes that the infant agent activates one of the contingent behaviors (Hide, Roll, Chase) or if they do not activate an behavior (Independent Play). B. Participation is shown as the average of 5M training steps, placed in the middle of the step range it was averaged over. The values are normalized by dividing them by the 99th percentile value of these metrics across training. A and B are averaged over 3 seeds.

Experiment 1: Compare exploration across intrinsic reward functions

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

Method

We run three seeds of infant agents for each reward function for 20M steps. We record all experience to perform behavior analysis, and evaluate the world model at the end of training.

Behavior diversity 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 distribution relative to the entropy of 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 the proportion of episodes where the behavior is activated. For each of the activated activities 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.

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, which may occupy a very different part of the trajectory space than the autonomous agents.

For each agent, we create a set of validation cases from its lifetime experience by uniform sampling 2000 trajectory segments. 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.

Table 1: Normalized entropy of infant state components. Calculated out of 100. Mean and standard error, n=3

Agent	Location	Orientation	Pose	Attention
Random	5 ± 0	56 ± 0	100 ± 0	62 ± 0
Adversarial	45 ± 5	79 ± 2	76 ± 4	71 ± 1
Disagreement	93 ± 0	100 ± 0	99 ± 0	95 ± 0
RND	87 ± 1	98 ± 0	99 ± 0	93 ± 0
δ-progress	40 ± 7	88 ± 1	94 ± 2	80 ± 1
γ-progress	38 ± 4	80 ± 3	90 ± 3	74 ± 1

Results

Disagreement and RND yield the greatest diversity of experience and acquire 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 (Location: $p < 0.005$; Orientation: $p < 0.01$; Attention: $p < 0.001$; t-test with fdr-bh correction). They also generate a larger number of total activations (Table 2) than the random agent, δ-progress, and γ-progress ($p < 0.01$), and RND shows higher total and Chase activations than Adversarial ($p < 0.05$).

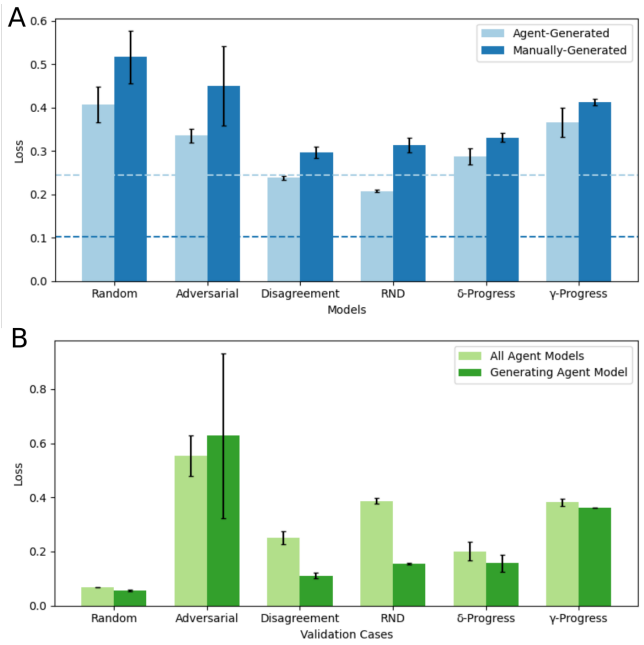

Figure 4: A. World model loss on validation cases generated from the experience of other agents and validation cases that were created manually. Lower values indicate better accuracy. Agent-Generated cases include validation sets from all seeds of all intrinsic reward functions. The horizontal lines 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.

This behavior diversity likely contributes to the large spread in Figure 4B between 1) the model loss from other agents' world models on experiences from Disagreement and RND, and 2) the low loss that Disagreement and RND can achieve on its own validation cases. This implies that the cases are predictable by this model class, but that the other models have not learned the dynamics yet.

The world model loss from Disagreement and RND agents is significantly lower than Random, Adversarial, and γ-progress agents when predicting the outcome of experiences that other instrinsically-motivated agents collected ($p < 0.05$; t-test with fdr-bh correction and one high-loss outlier removed in Adversarial and γ-progress). Interestingly, δ-progress also does well on our world model evaluation despite having fewer contingency activations, lower participation, and lower state entropy.

Although RND and Disagreement both generate a high diversity of states, they have different temporal structures in behavior (Figure 3A). RND has a relative stable split of activity activations after 5M steps. Disagreement appears to have phases of activity preference, generating different proportions of behavior activations over time.

Table 2: Percent of episodes where behavior activation occurs over agent training. Mean and standard error, n=3.

Agent	Hide	Roll	Chase	Total
Random	30 ± 0	9 ± 0	0 ± 0	39 ± 0
Adversarial	7 ± 2	64 ± 6	2 ± 0	73 ± 4
Disagreement	38 ± 5	35 ± 7	14 ± 4	87 ± 3
RND	15 ± 2	53 ± 2	23 ± 1	91 ± 1
δ-Progress	23 ± 6	14 ± 2	4 ± 2	42 ± 7
γ-Progress	7 ± 1	3 ± 0	4 ± 3	14 ± 3

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. The described behavior contributes to the lower entropy observed in Table 1 and the high Roll participation in Figure 3B.

The intrinsic reward function seeks out high loss situations. The agent succeeds and the validation cases made from the adversarial agent's experience have in the highest average loss across models (Figure 4). Even with the benefit of the experience, the loss remains high for the world model of the agent that collected it. The agent found situations that are difficult to model accurately with the current world model class.

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

In our environment, we want to understand the effect of different levels of caregiver contingency. How is the infant agent's understanding of the world affected by being in an environment with a caregiver that frequently responds to their actions compared with one that rarely responds?

Method

Agents are trained for 10M steps with the Disagreement reward function. In contrast to Experiment 1, the 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 is sensitive to the infant. Different settings for this probability are tested: 1%, 5%, 20%, 80%, 95%, and 99%. We refer to 95% and 99% as high-contingency (HC) and 1% and 5% as low-contingency (LC).

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

Results

Increasing caregiver contingency corresponds to a shift in prediction difficulty from proprioceptive inputs to external dynamics The "Self" component is harder to predict in validation cases with LC caregivers than with HC caregivers. This holds for both agents trained with LC and HC caregivers. In contrast, the presence of a HC caregiver corresponds with more difficult to predict ball and caregiver components (Figure 5). In the absence of very frequent caregiver interaction,

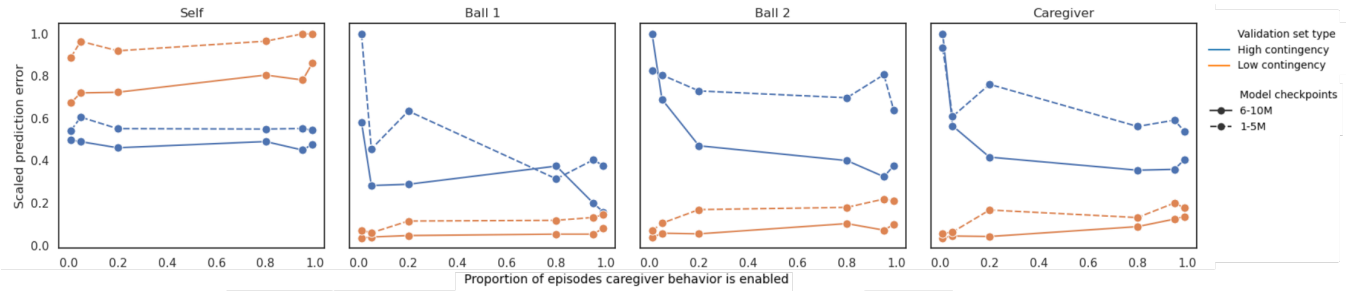

Figure 5: Prediction error by observation component shown plotted against the proportion of time caregiver was enabled while the agent trained. Lower values indicate better accuracy. We decompose the loss into that due to infant orientation, position, and arm configuration ("Self"), Ball 1, Ball 2, and Caregiver predictions. High contingency is the set of validation cases created from agents the caregiver behaviors enabled 95% of the time or more, and low contingency is from agents with the caregiver behaviors 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.

agents focus on proprioceptive exploration. When present, the caregiver facilitates challenging external dynamics scenarios though its complex behavior patterns.

There is a net benefit to world model accuracy from high levels of contingency Training in an environment with a HC caregiver yields substantial improvements in accuracy on challenging ball and caregiver dynamics, and only a small decrease in accuracy on scenarios with a LC caregiver.

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. As noted in the previous result, 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. This asymmetry appears to be persistent across training, it appears in early checkpoints, averaged from 1M to 5M steps, and later checkpoints, averaged from 6M to 10M steps.

The effect of increasing contingency on Self components appears nearly symmetric, possibly as a consequence of the previous result: the presence of a HC caregiver doesn't correspond to challenging examples, the difficulty decreases.

Discussion and future work

We find that basic social interaction and contingency activation can arise without requiring a specific module, social intrinsic reward, or extrinsic reward. Contingency sensitivity can arise from curiosity, implemented here as an information-maximizing intrinsic reward function.

An infant's ability to cause change in the world is amplified by an attentive caregiver. In our environment the caregiver agent is likely to facilitate the infant agent's first experience seeing a ball move. The amplification is the case for real infants, caregivers are responsible for a huge amount of action in their world, frequently in response to the infant's actions like cooing, crying, reaching or pointing. This amplification is visible in our results, as the prediction difficulty of external dynamics increases with a highly-contingent caregiver.

Contingent caregiver behavior provides a dense intrinsic reward for curiosity signals in a usually sparse environment. Social behavior is a very rich and difficult prediction problem, so each interaction will yield more data to challenge their understanding of the world. For signals that depend on a world model, this going to strongly motivate exploration. This idea is supported in our results: multiple intrinsic reward functions frequently activate caregiver behaviors.

Disagreement and RND motivated robust exploration in different ways: we observed different temporal patterns in behavior activation and different participation levels. Other intrinsic motivations were not as successful. Learning progress reward functions generated less state coverage and less robust world models. A possible challenge for learning progress is that rewards may be sparse if the world model makes occasional step-like improvements in quality but not frequent small improvements. Adversarial rewards led the infant agent to repeatedly hit the ball during the Roll behavior because the model loss remained high. This is an instance of the white noise problem, a known issue with the signal. Humans, in contrast, have mechanisms to avoid fixation on a single activity — boredom, for example, has been considered as motivating a wider diversity of experience (Bench & Lench, 2013).

Pixel-based observations and a richer caregiver agent would offer additional challenges for a world model. The agent has the potential to be extended with explicit representations of other agents' beliefs or with model-based reinforcement learning so that the predictive capabilities of the world model could be leveraged to plan future actions.

Our modeling approach allows us to generate trajectories of the virtual infant agent. Real infant walking trajectories have been analyzed to understand exploration patterns and state coverage (Hoch, O'Grady, & Adolph, 2019). One important next step is to compare artificial and human trajectories on matched environments. This may lead to a better characterisation of infant exploratory motivations and patterns as well as insight into the observed diversity of exploration in children.

References

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

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

Bench, S. W., & Lench, H. C. (2013). On the function of boredom. *Behavioral sciences*, *3*(3), 459–472.

Bigelow, A. E., & Rochat, P. (2006). Two-month-old infants' sensitivity to social contingency in mother–infant and stranger–infant interaction. *Infancy*, *9*(3), 313–325.

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

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

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

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

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

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

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

Juliani, A., Berges, V.-P., Teng, E., Cohen, A., Harper, J., Elion, C., ... others (2018). Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*.

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

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

Nadel, J., Carchon, I., Kervella, C., Marcelli, D., & Réserbat-Plantey, D. (1999). Expectancies for social contingency in 2-month-olds. *Developmental science*, *2*(2), 164–173.

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

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

Pathak, D., Gandhi, D., & Gupta, A. (2019). Self-supervised exploration via disagreement. In *International conference on machine learning* (pp. 5062–5071).

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

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

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

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., & Pathak, D. (2020). Planning to explore via self-supervised world models. In *International conference on machine learning* (pp. 8583–8592).

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

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

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

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

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

Tronick, E., Als, H., Adamson, L., Wise, S., & Brazelton, T. B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child psychiatry*, *17*(1), 1–13.