Active World Model Learning with Progress Curiosity

Kuno Kim [1] **Megumi Sano** [1] **Julian De Freitas** [2] **Nick Haber** [*3] **Daniel Yamins** [*1 4]

Abstract

World models are self-supervised predictive models of how the world evolves. Humans learn world models by curiously exploring their environment, in the process acquiring compact abstractions of high bandwidth sensory inputs, the ability to plan across long temporal horizons, and an understanding of the behavioral patterns of other agents. In this work, we study how to design such a curiosity-driven Active World Model Learning (AWML) system. To do so, we construct a curious agent building world models while visually exploring a 3D physical environment rich with distillations of representative real-world agents. We propose an AWML system driven by γ-Progress: a scalable and effective learning progress-based curiosity signal. We show that γ-Progress naturally gives rise to an exploration policy that directs attention to complex but learnable dynamics in a balanced manner, as a result overcoming the "white noise problem". As a result, our γ-Progress-driven controller achieves significantly higher AWML performance than baseline controllers equipped with state-of-the-art exploration strategies such as Random Network Distillation and Model Disagreement.

1. Introduction

Imagine yourself as an infant in your parent's arms, sitting on a playground bench. You are surrounded by a variety of potentially interesting stimuli, from the constantly whirring merry-go-round, to the wildly rustling leaves, to your parent's smiling and cooing face. After briefly staring at the motionless ball, you grow bored. You consider the merry-go-round a bit more seriously, but its periodic motion is ultimately too predictable to keep your attention long. The leaves are quite entertaining, but after watching their random motions for a while, your gaze lands on your parent. Here you find something really interesting: you can anticipate, elicit, and guide your parents' changes in expression as you engage them in a game of peekaboo. Though just an infant, you have efficiently explored and interacted with the environment, in the process gaining strong intuitions about how different things in your world will behave.

Here, you have learned a powerful *world model* — a self-supervised predictive model of how the world evolves, both due to its intrinsic dynamics and your actions. Through learning world models, humans acquire compact abstractions of high bandwidth sensory inputs, the ability to plan across long temporal horizons, and the capacity to anticipate the behavioral patterns of other agents. Devising algorithms that can efficiently construct such world models is an important goal for the next generation of socially-integrated AI and robotic systems.

A key challenge in world model learning is that real-world environments contain a diverse range of dynamics with varying levels of learnability. The inanimate ball and periodic merry-go-round display dynamics that are easy to learn. On the other end of the spectrum, stimuli such as falling leaves exhibit random noise-like dynamics. Lying in a "sweet spot" on this spectrum are animate agents that have interesting and complex yet learnable dynamics, e.g. your parent's expressions and play offerings. Balancing attention amidst a sea of stimuli with diverse dynamics in a way that maximizes learning progress is a challenging problem. Particularly difficult is the *white noise problem* (Schmidhuber, 2010; Pathak et al., 2017; Burda et al., 2018b; Pathak et al., 2019), perseverating on unlearnable stimuli rather than pursuing learnable dynamics. Thus, it is a natural hypothesis that behind the infant's ability to learn powerful world models must be an equally powerful *active learning* algorithm that directs its attention to maximize learning progress.

In this work, we formalize and study Active World Model Learning (AWML) – the problem of determining a directed exploration policy that enables efficient construction of better world models. To do so, we construct a progress-driven curious neural agent performing AWML in a custom-built 3D virtual world environment. Specifically, our contribu-

[*]Equal contribution [1]Department of Computer Science, Stanford University [2]Department of Psychology, Harvard University [3]Graduate School of Education, Stanford University [4]Department of Psychology, Stanford University. Correspondence to: Kuno Kim <khkim@cs.stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

tions are as follows:

1. We construct a 3D virtual environment rich with agents displaying a wide spectrum of realistic stimuli behavior types with varying levels of learnability, such as static, periodic, noise, peekaboo, chasing, and mimicry.

2. We formalize AWML within a general reinforcement learning framework that encompasses curiosity-driven exploration and traditional active learning.

3. We propose an AWML system driven by γ-Progress: a novel and scalable learning progress-based curiosity signal. We show that γ-Progress gives rise to an exploration policy that overcomes the white noise problem and achieves significantly higher AWML performance than state-of-the-art exploration strategies — including Random Network Distillation (RND) (Burda et al., 2018b) and Model Disagreement (Pathak et al., 2019).

2. Related Works

World Models A natural class of world models involve forward dynamics prediction. Such models can directly predict future video frames (Finn et al., 2016; Wang et al., 2018; Wu et al., 2019), or latent feature representations such as 3D point clouds (Byravan & Fox, 2017) or object-centric, graphical representations of scenes (Battaglia et al., 2016; Chang et al., 2016; Mrowca et al., 2018). Action-conditioned forward-prediction models can be used directly in planning for robotic control tasks (Finn & Levine, 2017), as performance-enhancers for reinforcement learning tasks (Ke et al., 2019), or as "dream" environment simulations for training policies (Ha & Schmidhuber, 2018). In our work, we focus on forward dynamics prediction with object-oriented representations.

Active Learning and Curiosity A key question the agent is faced with is how to choose its actions to efficiently learn the world model. In the classical *active learning* setting (Settles, 2011), an agent seeks to learn a supervised task with costly labels, judiciously choosing which examples to obtain labels for so as to maximize learning efficiency. More recently, active learning has been implicitly generalized to self-supervised reinforcement learning agents (Schmidhuber, 2010; Oudeyer et al., 2013; Jaderberg et al., 2016). In this line of work, agents typically self-supervise a world model with samples obtained by curiosity-driven exploration. Different approaches to this general idea exist, many of which are essentially different approaches to estimating future *learning progress* — e.g. determining which actions are likely to lead to the highest world model prediction gain in the future. One approach is the use of *novelty* metrics, which measure how much a particular part of the environment has been explored, and direct agents into under-explored parts of state-space. Examples include count-based and psuedo-count-based methods (Strehl & Littman, 2008; Bellemare et al., 2016; Ostrovski et al., 2017), Random Network Distillation (RND) (Burda et al., 2018b), and *empowerment* (Mohamed & Rezende, 2015). Novelty-based approaches avoid the difficult world model progress estimation problem entirely by not depending at all on a specific world model state, and relying on novelty as a (potentially inconsistent) proxy for expected learning progress.

The simplest idea that takes into account the world model is *adversarial* curiosity, which estimates current world model error and directs agents to take actions estimated to maximize this error (Stadie et al., 2015; Pathak et al., 2017; Haber et al., 2018). However, adversarial curiosity is especially prone to the *white noise problem*, in which agents are motivated to waste time fruitlessly trying to solve unsolvable world model problems, e.g. predicting the dynamics of random noise. The white noise problem can to some degree be avoided by solving the world-modeling problem in a learned latent feature space in which degeneracies are suppressed (Pathak et al., 2017; Burda et al., 2018a).

Directly estimating learning progress (Oudeyer et al., 2007; 2013) or *information gain* (Houthooft et al., 2016) avoids the white noise problem in a more comprehensive fashion. However, such methods have been limited in scope because they involve calculating quantities that cannot easily be estimated in high-dimensional continuous action spaces. *Surprisal* (Achiam & Sastry, 2017) and model disagreement (Pathak et al., 2019) present computationally-tractable alternatives to information gain, at the cost of the accuracy of the estimation. For comprehensive reviews of intrinsic motivation signal choices, see (Aubret et al., 2019; Linke et al., 2019). In this work, we present a novel method for estimating learning progress that is "consistent" with the original prediction gain objective while also scaling to high-dimensional continuous action-spaces.

Cognitive Science Research in cognitive science suggests that humans are active world model learners from a young age. Infants appear to actively gather information from their environment by attending to objects in a highly non-random manner (Smith et al., 2019), devoting more attention to objects that violate their expectations (Stahl & Feigenson, 2015). They are also able to self-generate learning curricula, with preference to stimuli that are complex enough to be interesting but still predictable (Kidd et al., 2012). Interestingly, infants seem to particularly attend to spatiotemporal kinematics indicative of *animacy*, such as efficient movement towards targets (Gergely et al., 1995) and contingent behavior between agents (Frankenhuis et al., 2013). Our work shows that a similar form of animate attention naturally arises as a result of optimizing for learning progress.

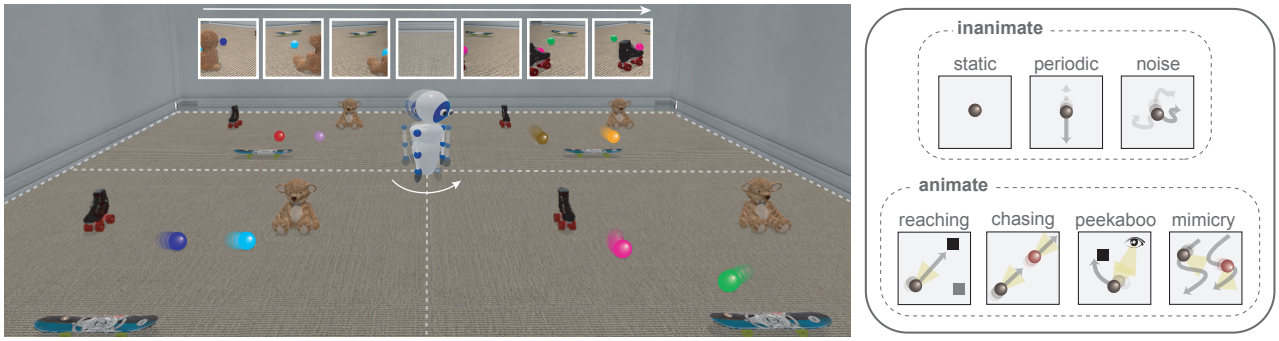

Figure 1. **Virtual environment.** Our 3D virtual environment is a distillation of key aspects of real-world environments. The *curious agent* (white robot) is centered in a room, surrounded by various *external agents* (colored spheres) contained in different quadrants, each with dynamics that correspond to a realistic inanimate or animate behavior (right box). The curious agent can rotate to attend to different behaviors as shown by the first-person view images at the top. See https://bit.ly/31vg7v1 for videos.

3. Virtual World Environment

To faithfully replicate real-world algorithmic challenges, we design our 3D virtual environment to preserve the following key properties of real-world environments:

1. *Diverse dynamics.* Agents operate under a diverse set of dynamics specified by agent-specific programs. An agent's actions may depend on those of another agent resulting in complex intedependent relationships.

2. *Partial observability.* At no given time do we have full access to the current state of every agent in the environment. Rather, our learning is limited by what lies within our field of view.

3. *Contingency.* How much we learn is contingent on how we, as embodied agents, choose to interact with the environment.

Concretely, Our virtual environment consists of two main components, a *curious agent* and various *external agents*.

The **curious agent**, embodied by an avatar, is fixed at the center of a room (Figure 1). Just as a human toddler can control her gaze to visually explore her surroundings, the agent is able to partially observe the environment based on what lies in its field of view (see top of Figure 1). The agent can choose from 9 actions: rotate $12°, 24°, 48°$, or $96°$, to the left/right, or stay in its current orientation.

The **external agents** are spherical avatars that each act under a hard-coded policy inspired by real-world inanimate and animate stimuli. An *external agent behavior* consists of either one external agent, e.g reaching, or two interacting ones, e.g chasing. Since external agents are devoid of surface features, the curious agent must learn to attend to different behaviors based on spatiotemporal kinematics alone. We experiment with external agent behaviors (see Figure 1, right) including static, periodic, noise, reaching, chasing, peekaboo, and mimicry. The animate behaviors have deterministic and stochastic variants, where the stochastic variant preserves the core dynamics underlying the behavior, albeit with more randomness. See https://bit.ly/31vg7v1 for video descriptions of the environment and external agent behaviors.

We divide the room into four quadrants, each of which contains various auxiliary objects (e.g teddy bear, roller skates, surfboard) and one external agent behavior. The room is designed such that the curious agent can see at most one external agent behavior at any given time. This design is key in ensuring partial observability, such that the agent is faced with the problem of allocating attention between different external agent behaviors in an efficient manner.

4. Active World Model Learning

In this section, we formalize Active World Model Learning (AWML) as a Reinforcement Learning (RL) problem that generalizes conventional Active Learning (AL). We then derive γ-Progress, a scalable progress-based curiosity signal with algorithmic and computational advantages over previous progress-based curiosity signals.

Consider an agent in an environment $\mathcal{E} = (\mathcal{S}, \mathcal{A}, P, P_0)$ where $\mathcal{S}, \mathcal{A}$ are state, action spaces, $P : \mathcal{S} \times \mathcal{A} \to \Omega(\mathcal{S})$ is the transition dynamics, $\Omega(\mathcal{S})$ is the set of probability measures on $\mathcal{S}$, and P_0 is the initial state distribution. The agent's goal is to optimize a sequence of data collection decisions in order to learn a model ω_θ of a target function $\omega : \mathcal{X} \to \Omega(\mathcal{Y})$ with as few data samples as possible. We model this agent's decision making process as an infinite horizon Markov Decision Process (MDP) $\mathcal{M} = (\bar{\mathcal{S}}, \mathcal{A}, \bar{P}, \bar{P}_0, c)$. Intuitively, $\mathcal{M}$ is a meta-MDP that jointly characterizes the evolution of the environment, collected data history, and the model as the agent makes data collection decisions to optimize the learning objective encoded by the MDP reward. Specifically, $\bar{\mathbf{s}} \in \bar{\mathcal{S}} = \mathcal{S} \times \mathcal{H} \times \Theta$ is a meta-state that decomposes into $\bar{\mathbf{s}} = (\mathbf{s}, H, \theta)$ where $\mathbf{s} \in \mathcal{S}$ is

an environment state, $H = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1 ...\} \in \mathcal{H}$ is the history of visited environment state-actions, and $\theta \in \Theta$ is the current model parameters. Intuitively H is a raw form of the data collected so far which can be post-processed to yield a training set for ω_θ. $\mathbf{a} \in \mathcal{A}$ is a data collection decision, e.g where to move next in the environment, and the meta-dynamics $\bar{P} : \bar{\mathcal{S}} \times \mathcal{A} \to \Omega(\bar{\mathcal{S}})$ is defined as:

$$(\mathbf{s}', H', \theta') \sim \bar{P}(\cdot|\bar{\mathbf{s}} = (\mathbf{s}, H, \theta), \mathbf{a}) \text{ where}$$
$$\mathbf{s}' \sim P(\mathbf{s}, \mathbf{a}),\ H' = H \cup \{\mathbf{a}, \mathbf{s}'\},\ \theta' \sim P_\ell(H', \theta)$$

where $P_\ell : H \times \Theta \to \Omega(\Theta)$ is transition function for the model parameters, e.g a (stochastic) learning algorithm which updates the parameters on the history of data. In words, $\bar{P}$ steps the environment state $\mathbf{s}$ according to the environment dynamics P, appends the history with new data, and updates the model ω_θ on the augmented history. The meta-initial state distribution is $\bar{P}_0(\bar{\mathbf{s}} = (\mathbf{s}, H, \theta)) = P_0(\mathbf{s})\mathbb{1}(H = \{\})q(\theta)$ where $\mathbb{1}$ is the indicator function and $q(\theta)$ is a prior distribution over the model parameters. $c : \mathcal{S}^* \times \mathcal{A}^* \to \mathbb{R}$ is the cost function which encodes the learning objective of the agent. For example, $c(\bar{\mathbf{s}} = (\mathbf{s}, H, \theta)) = \mathcal{L}_\mu(\theta) = \mathbb{E}_\mu[\mathcal{L}(\theta, \mathbf{x}, \mathbf{y})]$ encodes the goal of an agent seeking to minimize a loss function $\mathcal{L}$ on data $(\mathbf{x}, \mathbf{y}) \sim \mu$ while training on a minimal number of data samples. A policy $\pi : \mathcal{S} \to \Omega(\mathcal{A})$ maps states to action distributions and an optimal policy $\pi^* = \arg\max_\pi J(\pi)$ achieves the highest performance $J(\pi) = -\mathbb{E}_\pi[\sum_{t=0}^{\infty} \beta^t c(\bar{\mathbf{s}})]$ where $0 \leq \beta \leq 1$ is a discount factor. Overall, the MDP $\mathcal{M}$ is constructed from an environment $\mathcal{E} = (\mathcal{S}, \mathcal{A}, P, P_0)$, a target function ω, a learning algorithm P_ℓ, and a prior parameter distribution $q(\theta)$. Appendix B shows how several variants of conventional active learning can be recovered by appropriately defining the MDP $\mathcal{M}$ in relation to $\omega, \mathcal{X}, \mathcal{Y}$.

We now formalize the Active World Model Learning (AWML) problem using this general active learning framework. Formally, AWML aims to find an optimal policy for the meta MDP $\mathcal{M}$ with the additional constraint that $\mathcal{X}, \mathcal{Y}$ are arbitrary length sequences of environment states and actions. In words, the target function ω is constrained to be a self supervised predictor on trajectories sampled from the environment. We henceforth aptly refer to ω_θ as the world model. Consider a simple AWML problem of learning the forward dynamics, i.e $\mathcal{X} = \mathcal{S} \times \mathcal{A}, \mathcal{Y} = \mathcal{S}, P = \omega$, and

$$c(\bar{\mathbf{s}}, \mathbf{a}, \bar{\mathbf{s}}') = \mathcal{L}_\mu(\theta) - \mathcal{L}_\mu(\theta') \tag{1}$$

where $\bar{\mathbf{s}} = (\mathbf{s}, H, \theta), \bar{\mathbf{s}}' = (\mathbf{s}', H', \theta')$, with a discount factor $\beta < 1$ to encourage minimal interactions with the environment. Recall that $\theta' = P_\ell(H \cup \{\mathbf{a}, \mathbf{s}'\}, \theta)$ is the updated model parameters after collecting new data $\{\mathbf{a}, \mathbf{s}'\}$. Thus $-c(\bar{\mathbf{s}}, \mathbf{a}, \bar{\mathbf{s}}')$ measures the reduction in world model loss as a result of obtaining new data $\{\mathbf{a}, \mathbf{s}'\}$, i.e the *prediction gain*. Unfortunately, evaluating prediction gain at every environment step involves repeatedly computing $\mathcal{L}_\mu(\theta) - \mathcal{L}_\mu(\theta')$ which is typically intractable as many samples are needed estimate the expectation. This bottleneck necessitates an efficiently computable proxy reward to estimate Eq. 1. Thus, solving the AWML problem entails solving two subproblems: Proxy reward design and Reinforcement Learning (RL). See Appendix C to see how a variety of curiosity signals proposed in prior work can be interpretted as (Stadie et al., 2015; Burda et al., 2018b; Pathak et al., 2019; Achiam & Sastry, 2017) solution attempts to the former problem. We now motivate γ-Progress by first outlining the limitations of a previously proposed progress-based curiosity signal, δ-Progress.

δ-Progress (Achiam & Sastry, 2017; Graves et al., 2017) curiosity measures how much better the current new model θ_{new} is compared to an old model θ_{old}.

$$\mathcal{L}_\mu(\theta) - \mathcal{L}_\mu(\theta') \simeq \log \frac{\omega_{\theta_{new}}(\mathbf{s}'|\mathbf{s}, \mathbf{a})}{\omega_{\theta_{old}}(\mathbf{s}'|\mathbf{s}, \mathbf{a})} \tag{2}$$

The choice of $\theta_{new}, \theta_{old}$ is crucial to the efficacy of the progress reward. A popular approach (Achiam et al., 2017; Graves et al., 2017) is to choose

$$\theta_{new} = \theta_k, \quad \theta_{old} = \theta_{k-\delta}, \quad \delta > 0 \tag{3}$$

where θ_k is the model parameter after k update steps using P_ℓ. Intuitively, if the progress horizon δ is too large, we obtain an overly optimistic approximation of future progress. However if δ is too small, the agent may prematurely give up on learning hard transitions, e.g where the next state distribution is very sharp. In practice, tuning the value δ presents a major challenge. Furthermore, the widely pointed out (Pathak et al., 2019) limitation of δ-Progress is that the memory usage grows $\mathcal{O}(\delta)$, i.e one must store δ world model parameters $\theta_{k-\delta}, ..., \theta$. As a result it is intractable in practice to use $\delta > 3$ with deep neural net models.

γ-Progress (Ours) We propose the following choice of $\theta_{new}, \theta_{old}$ to overcome both hurdles:

$$\theta_{new} = \theta, \quad \theta_{old} = (1 - \gamma) \sum_{i=1}^{k-1} \gamma^{k-1-i} \theta_0 \tag{4}$$

In words, the old model is a weighted mixture of past models where the weights are exponentially decayed into the past. γ-Progress can be interpreted as a noise averaged progress signal. Conveniently, γ-Progress can be implemented with a simple θ_{old} update rule:

$$\theta_{old} \leftarrow \gamma \theta_{old} + (1 - \gamma)\theta_{new} \tag{5}$$

Similar to Eq. 3, we may control the level of optimism towards expected future loss reduction by controlling the progress horizon γ, i.e a higher γ corresponds to a more optimistic approximation. γ-Progress has key practical advantages over δ-Progress: γ is far easier to tune than δ, e.g

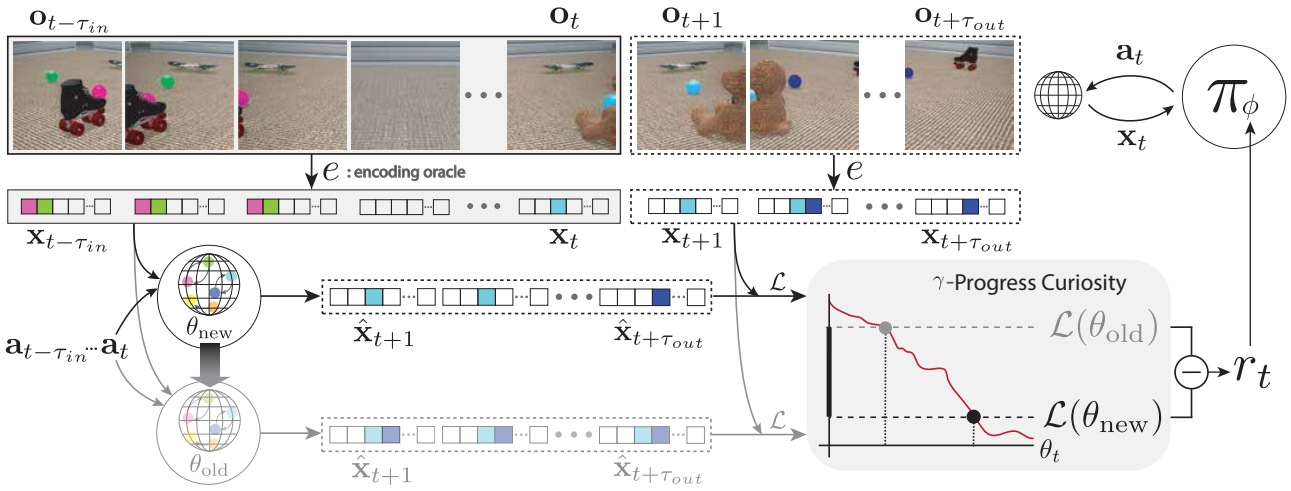

Figure 2. **Active World Model Learning with γ-Progress** The curious agent consists of a *world model* and a *progress-driven controller*. The curious agent's observations $\mathbf{o}_t$ are passed through an encoding oracle e that returns an object-oriented representation $\mathbf{x}_t$ containing the positions of external agents that are in view, auxiliary object positions, and the curious agent's orientation. Both the new (black) and old (gray) models take as input $\mathbf{x}_{t-\tau_{in}:t}$ and predict $\hat{\mathbf{x}}_{t:t+\tau_{out}}$. The old model weights, θ_{old}, are slowly updated to the new model weights θ_{new}. The controller, π_ϕ, is optimized to maximize γ-Progress reward: the difference $\mathcal{L}(\theta_{old}) - \mathcal{L}(\theta_{new})$.

we use a single value of γ throughout all experiments, and memory usage is constant with respect to γ. Crucially, the second advantage enables us to tune the progress horizon so that the model does not prematurely give up on exploring hard transitions. The significance of these practical advantages will become apparent from our experiments.

5. Methods

In this section we describe practical instantiations of the two components in our AWML system: a *world model* which fits the forward dynamics and a *controller* which chooses actions to maximize γ-Progress reward. See Appendix E for full details on architectures and training procedures.

World Model As the focus of this work is not to resolve the difficulty of representation learning from high-dimensional visual inputs, we assume that the agent has access to an oracle encoder $e : \mathcal{O} \to \mathcal{X}$ that maps an image observation $\mathbf{o}_t \in \mathcal{O}$ to a disentangled object-oriented feature vector $\mathbf{x}_t = (\mathbf{x}_t^{ext}, \mathbf{x}_t^{aux}, \mathbf{x}_t^{ego})$ where $\mathbf{x}_t^{ext} = (\tilde{\mathbf{c}}_t, \mathbf{m}_t) = (\tilde{\mathbf{c}}_{t,1}, \ldots \tilde{\mathbf{c}}_{t,n_{ext}}, \mathbf{m}_{t,1}, \ldots, \mathbf{m}_{t,n_{ext}})$ contains information about the external agents; namely the observability masks $\mathbf{m}_{t,i}$ ($\mathbf{m}_{t,i} = 1$ if external agent i is in curious agent's view at time t, else $\mathbf{m}_{t,i} = 0$) and masked position coordinates $\tilde{\mathbf{c}}_{t,i} = \mathbf{c}_{t,i}$ if $\mathbf{m}_{t,i} = 1$ and else $\tilde{\mathbf{c}}_{t,i} = \hat{\mathbf{c}}_{t,i}$. Here, $\mathbf{c}_{t,i}$ is the true global coordinate of external agent i and $\hat{\mathbf{c}}_{t,i}$ is the model's predicted coordinate of external agent i where $i = 1, \ldots, n_{ext}$. Note that the partial observability of the environment is preserved under the oracle encoder since it provides coordinates only for external agents in view. $\mathbf{x}_t^{aux}$ contains coordinates of auxiliary objects, and $\mathbf{x}_t^{ego}$ contains the ego-centric orientation of the curious agent.

Algorithm 1 AWML with γ-Progress

Require: progress horizon γ, step sizes η_ω, η_Q
Initialize θ_{new}, ϕ
for $k = 1, 2, \ldots$ **do**
  Update policy: $\pi_\phi \leftarrow \epsilon\text{-}greedy(Q_\phi, \epsilon - 0.0001)$
  Sample $(\mathbf{x}, \mathbf{a}, c) \sim \pi_\phi$ and place in Buffer $\mathcal{B}$
  where $c = \mathcal{L}(\theta_{new}, \mathbf{x}, \mathbf{a}) - \mathcal{L}(\theta_{old}, \mathbf{x}, \mathbf{a})$
  for $j = 1, \ldots, M$ **do**
    Sample batch $b_j \sim \mathcal{B}$
    Update new world model:
    $\theta_{new} \leftarrow \theta_{new} - \eta_\omega \cdot \nabla_{\theta_{new}} \mathcal{L}(\theta_{new}, b_j)$
    Update old world model:
    $\theta_{old} \leftarrow \gamma\theta_{old} + (1-\gamma)\theta_{new}$
    Update Q-network with DQN (Mnih et al., 2015):
    $\phi \leftarrow \text{DQN}(\phi, b_j, \eta_Q)$
  end
end

Our world model ω_θ is an ensemble of component networks $\{\omega_{\theta^k}\}_{k=1}^{N_{cc}}$ where each ω_{θ^k} independently predicts the forward dynamics for a subset $I_k \subseteq \{1, \ldots, \dim(\mathbf{x}^{ext})\}$ of the input dimensions of $\mathbf{x}^{ext}$ corresponding to a minimal behaviorally interdependent group. For example, $\mathbf{x}^{ext}_{t:t+\tau,I_k}$ may correspond to the masked coordinates and observability masks of the chaser and runner external agents for times $t, t+1, \ldots, t+\tau$. We found that such a "disentangled" architecture outperforms a simple entangled architecture (see Appendix D for details). We assume $\{I_k\}_{k=1}^{n_{cc}}$ is given as prior knowledge but future work may integrate dependency graph learning into our pipeline. A component network ω_{θ^k} takes as input $(\mathbf{x}^{ext}_{t-\tau_{in}:t,I_k}, \mathbf{x}^{aux}_{t-\tau_{in}:t}, \mathbf{x}^{ego}_{t-\tau_{in}:t}, \mathbf{a}_{t-\tau_{in}:t+\tau_{out}})$,

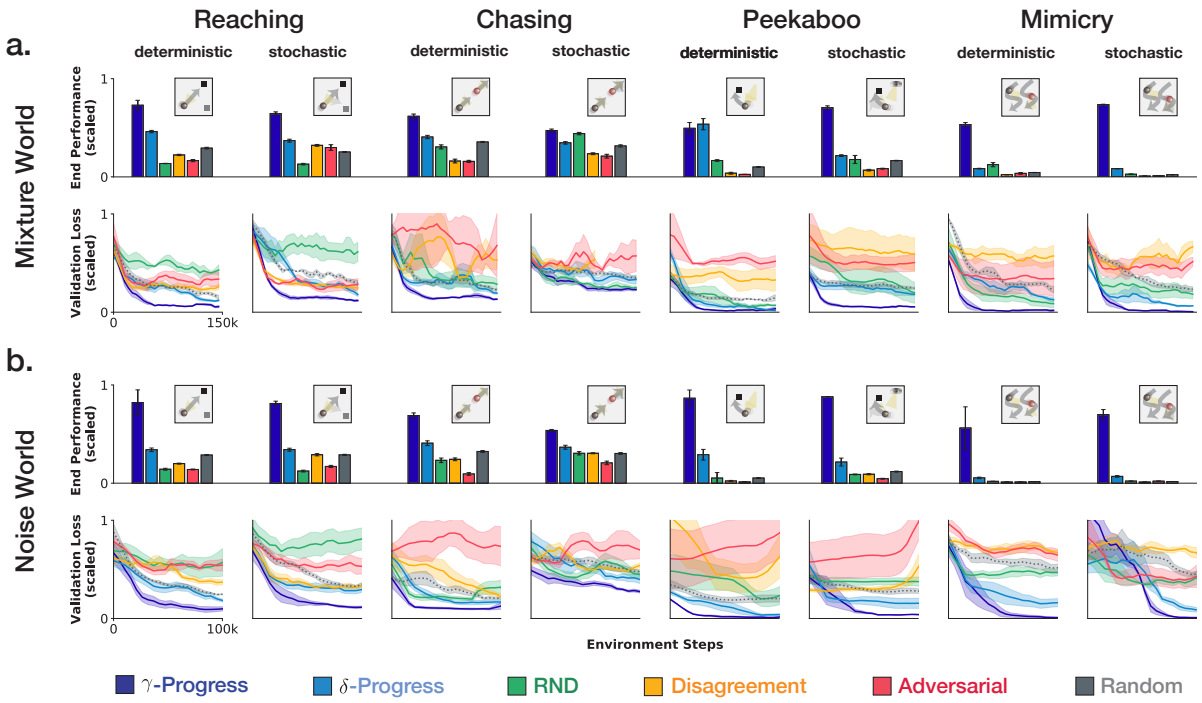

Figure 3. **AWML Performance**. The animate external agent is varied across experiments according to the column labels. End performance is the mean of the last five validation losses. Sample complexity plots show validation losses every 5000 environment steps. Error bars/regions are standard errors of the best 5 seeds out of 10. (a). *Mixture World*: γ-Progress achieves lower sample complexity than all baselines on $7/8$ behaviors while tying with RND on stochastic chasing. Notably, γ-Progress also outperforms all baselines in end performance on $6/8$ behaviors. (b). *Noise World*: γ-Progress is more robust to white noise than baselines and achieves lower sample complexity and higher end performance on $8/8$ behaviors. Baselines frequently perform worse than random due to noise fixation

where $\mathbf{a}$ denotes the curious agent's actions, and outputs $\hat{\mathbf{x}}^{ext}_{t:t+\tau_{out},I_k}$. The outputs of the component network are concatenated to get the final output $\hat{\mathbf{x}}^{ext}_{t:t+\tau_{out}} = (\hat{\mathbf{c}}_{t:t+\tau_{out}}, \hat{\mathbf{m}}_{t:t+\tau_{out}})$. The world model loss is:

$$\mathcal{L}(\theta, \mathbf{x}_{t-\tau_{in}:t+\tau_{out}}, \mathbf{a}_{t-\tau_{in}:t+\tau_{out}}) = \sum_{t'=t}^{t+\tau_{out}} \sum_{i=1}^{N_{ext}} \mathbf{m}_{t',i} \cdot \|\hat{\mathbf{c}}_{t',i} - \tilde{\mathbf{c}}_{t',i}\|_2 + \mathcal{L}_{ce}(\hat{\mathbf{m}}_{t',i}, \mathbf{m}_{t',i})$$

where $\mathcal{L}_{ce}$ is cross-entropy loss. We parameterize each component network ω_{θ^k} with a two-layer Long Short-Term Memory (LSTM) network followed by two-layer Multi Layer Perceptron (MLP). The number of hidden units are adapted to the number of external agents being modeled.

The Progress-driven Controller Our controller π_ϕ is a two-layer fully-connected network with 512 hidden units that takes as input $\mathbf{x}_{t-2:t}$ and outputs estimated Q-values for 9 possible actions which rotate the curious agent at different velocities. π_ϕ is updated with the DQN (Mnih et al., 2013) learning algorithm using the cost:

$$c(\mathbf{x}_t) = \mathcal{L}(\theta_{new}, \mathbf{x}_{t-\tau_{in}-\tau_{out}:t}, \mathbf{a}_{t-\tau_{in}-\tau_{out}:t}) - \mathcal{L}(\theta_{old}, \mathbf{x}_{t-\tau_{in}-\tau_{out}:t}, \mathbf{a}_{t-\tau_{in}-\tau_{out}:t}) \quad (6)$$

with $\gamma = 0.9995$ across all experiments.

Table 1. Mean ratio of baseline end performance over Random baseline end performance (standard error in parentheses)

	MIXTURE WORLD	NOISE WORLD
γ-PROGRESS	**7.83** (3.57)	**13.79** (5.29)
δ-PROGRESS	2.2 (0.51)	2.46 (0.55)
RND	1.25 (0.25)	0.85 (0.10)
DISAGREEMENT	0.62 (0.10)	0.76 (0.06)
ADVERSARIAL	0.62 (0.09)	0.59 (0.10)

6. Results

We evaluate the AWML performance of γ-Progress on two metrics: *end performance* and *sample complexity*. End performance is the inverse of the the final world loss after a larger number of environment interactions, and intuitively measures the "consistency" of the proxy reward with respect to the true reward. Sample complexity measures the rate of reduction in world model loss $\mathcal{L}_\mu(\theta)$ with respect to the number of environment interactions. The samples from the validation distribution μ correspond to core validation cases we crafted for each behavior. On the reaching behaviors, for example, we validate the world model loss with objects spawned at new locations. For details on each behavior-specific validation case and metric computation, we refer readers to Appendix F.

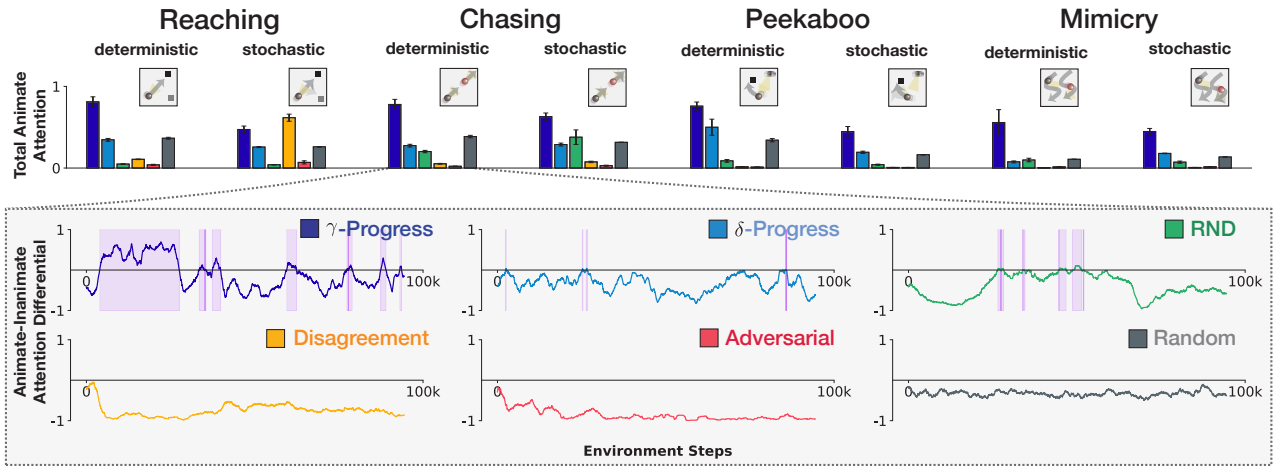

Figure 4. **Attention Patterns**. a) The bar plot shows the total animate attention, which is the ratio between the number of time steps an animate external agent was visible to the curious agent, and the number of time steps a noise external agent was visible. The time series plots in the zoom-in box show the differences between mean attention to the animate external agents and the mean of attention to the other agents in a 500 step window, with periods of animate preference highlighted in purple. Results are averaged across 5 runs. γ-Progress displays strong animate attention while baselines are either indifferent, e.g δ-Progress, or fixating on white noise, e.g Adversarial. b) Fraction of indifference and white noise failures, out of eight tasks.

Experiments are run in two virtual worlds: Mixture and Noise. In the Mixture world, the virtual environment is instantiated external agents spanning four representative types: static, periodic, noise, and animate. This set up is a natural distillation of a real-world environment containing a wide spectrum of behaviors. In the Noise world, the environment is instantiated with three noise agents and one animate agent. This world strain tests the noise robustness of γ-Progress. For each world, we run separate experiments in which the animate external agents are varied amongst the deterministic and stochastic versions of reaching, chasing, peekaboo, and mimicry agents (see Section 3). We compare the AWML performance of the following methods:

γ-Progress (Ours) is our proposed variant of progress curiosity which chooses θ_{old} to be a geometric mixture of all past models as in Eq. 4.

δ-Progress (Achiam & Sastry, 2017; Graves et al., 2017) is the δ-step learning progress reward from Eq. 3 with $\delta = 1$. We found that $\delta = 1, 2, 3$ perform similarly and any $\delta > 3$ is impractical due to memory constraints.

RND (Burda et al., 2018b) is a novelty-based method that trains a predictor neural net to match the outputs of a random state encoder. States for which the predictor networks fails to match the random encoder are deemed "novel", and thus receive high reward.

Disagreement (Pathak et al., 2019) is the disagreement based method from Eq. 9 with $N = 3$ ensemble models. We found that $N = 3$ performs best out of $N \in \{1, 2, 3\}$ and $N > 3$ is impractical due to memory constraints.

Adversarial (Stadie et al., 2015; Pathak et al., 2017) is the prediction error based method from Eq. 8. We use the ℓ_2 prediction loss of the world model as the reward.

Random chooses actions uniformly at random among the 9 possible rotations.

6.1. AWML Performance

Fig. 3a shows end performance (first row) and sample complexity (second row) in the Mixture world, and Fig. 3b shows the same for the Noise World. In the Mixture world, we see that γ-Progress has lower sample complexity than δ-Progress, Disagreement, Adversarial, and Random baselines on all $8/8$ behaviors and outperforms RND on $7/8$ behaviors while tying on stochastic chasing. In the Noise world, we see that γ-Progress has lower sample complexity than all baselines on all $8/8$ behaviors. See Table 1 for aggregate end performance, and `https://bit.ly/31vg7v1` for visualizations of model predictions.

6.2. Attention control analysis

Figure 4 shows the ratio of attention to animate vs other external agents for each behavior in the Mixture world as well as example animate-inanimate attention differential timeseries (for the Noise world, see Appendix G). The γ-Progress agents spend substantially more time attending to animate agents than do alternative policies. This increased animate-inanimate attention differential often corresponds to a characteristic attentional "bump" that occurs early as the γ-Progress curious agent focuses on animate external agents quickly before eventually "losing interest" as prediction accuracy is achieved. Strong animate attention emerges

for $7/7$ behaviors when using γ-Progress. Please see appendix H for a more in-depth of analysis of how attention, and particular early attention, predicts performance and how curiosity signal predicts attention.

Table 2. **Failure modes** Fraction of indifference and white noise failures, out of eight external agent behaviors.

	INDIFFERENCE	NOISE FIXATION
γ-PROGRESS	**0/8**	**0/8**
δ-PROGRESS	7/8	0/8
RND	2/8	4/8
DISAGREEMENT	0/8	7/8
ADVERSARIAL	0/8	8/8

Baselines display two distinct modes that lead to lower performance (Table 2). The first is *attentional indifference*, in which it finds no particular external agent interesting — more precisely, we say that an curiosity signal choice displays attentional indifference if its average animate/inanimate ratio in the Mixture world is within two standard deviations of Random policy's, thus achieving no directed exploration. δ-Progress frequently had attentional indifference as the new and old world model, separated by a fixed time difference, were often too similar to generate a useful curiosity signal.

The second failure mode is *white noise fixation*, where the observer is captivated by the noise external agents — more precisely, we say that a curiosity signal choice displays white noise fixation if its average animate/inanimate ratio in the Noise world is more than two standard deviations below Random policy's. RND suffers from white noise fixation due to the fact that our noise behaviors have the most diffuse visited state distribution. We also observe that for noise behaviors, a world model ensemble does not collectively converge to a single mean prediction, and as a result Disagreement finds the noise behavior highly interesting. Finally, the Adversarial baseline fails since noise behaviors yield the highest prediction errors. The white noise failure mode is particularly detrimental to sample complexity, with RND, Disagreement, and Adversarial, as evidenced by their below-Random performance in the Noise world.

7. Future directions

In this work, we propose and test γ-Progress, a computationally-scalable progress curiosity signal that enables an agent to efficiently train a world model in agent-rich environments while overcoming the white noise problem. Although in this work we limited the agent's ways of interacting with the world to visual attention, future works may explore larger action spaces. Another exciting avenue of research is investigating how to learn the world model priors used in this work such as the visual encoder and disentanglement prior. Finally, we hope see follow-up works

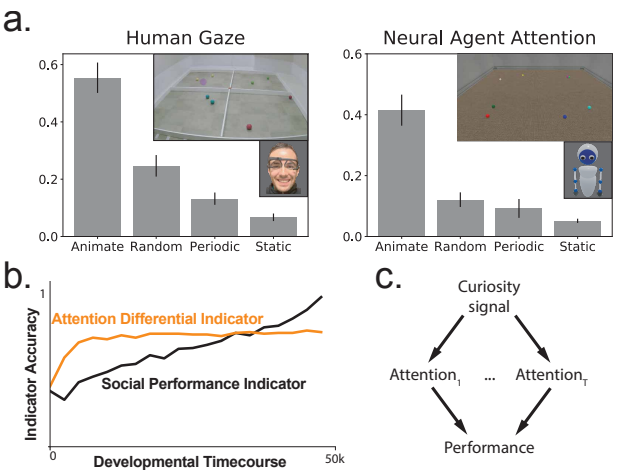

Figure 5. **Modeling human behavior**. (a) A pilot human behavior study. (b) Accuracy of early indicators of final performance, as a function of time, and (c) factor analysis: curiosity signal determines attention, which determines final performance.

applying γ-Progress to other domains such as video games.

We also see several key next steps at the intersection of AI and cognitive science.

Human behavior How might AWML-optimized agents model human behavior? We have run a pilot human subject experiment (Figure 5a) in which we conveyed static, periodic, animate, and noise stimuli to twelve human participants via spherical robots while we tracked their gaze. In aggregate, gaze is similar to γ-Progress attention. In follow-up work, we aim to make a finer model comparison to the behavior of humans shown a diverse array of stimuli.

Early indicator analysis Eventually, we would like to use curiosity-driven learning as a model for intrinsic motivation in early childhood. In this interpretation, the attention timecourse is a readily observable behavioral metric, and performance represents some more difficult-to-obtain measure of social acuity. Variation in curiosity signal would, in this account, be a latent correlate of developmental variability. For example, Autism Spectrum Disorder is characterized by both differences in low-level facial attention (Jones & Klin, 2013) and high-level social acuity (Hus & Lord, 2014). Motivated by this, we sought to determine whether attention could be used as an early indicator of performance. We thus train two models: (1) $\text{PERF}_{\leq T}$, which takes performance before time T as input, and (2) $\text{ATT}_{\leq T}$, which takes attention before time T as input. As seen in Figure 5b, $\text{ATT}_{\leq T}$ is throughout most of the timecourse a more accurate indicator than direct measurement of early-stage model performance itself. The overall situation is conveyed by the factor diagram Figure 5c — for further details, see Appendix H.1. As models grow in their capacity to model early childhood learning, translating such an analysis into a real-world experimental population could lead to substantial improvements

in diagnostics of developmental variability.

Acknowledgements

This work was supported by the McDonnell Foundation (Understanding Human Cognition Award Grant No. 220020469), the Simons Foundation (Collaboration on the Global Brain Grant No. 543061), the Sloan Foundation (Fellowship FG-2018- 10963), the National Science Foundation (RI 1703161 and CAREER Award 1844724), the DARPA Machine Common Sense program, the IBM-Watson AI Lab, and hardware donation from the NVIDIA Corporation.

References

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

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. *arXiv preprint arXiv:1705.10528*, May 2017.

Astington, J. W., Harris, P. L., and Olson, D. R. *Developing theories of mind*. CUP Archive, 1990.

Aubret, A., Matignon, L., and Hassas, S. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J., et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pp. 4502–4510, 2016.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. Large-scale study of curiosity-driven learning. In *arXiv:1808.04355*, 2018a.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation, 2018b.

Byravan, A. and Fox, D. Se3-nets: Learning rigid body motion using deep neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 173–180. IEEE, 2017.

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

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, May 2016.

Finn, C. and Levine, S. Deep visual foresight for planning robot motion. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 2786–2793. IEEE, 2017.

Finn, C., Goodfellow, I., and Levine, S. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, pp. 64–72, 2016.

Frankenhuis, W. E., House, B., Barrett, H. C., and Johnson, S. P. Infants' perception of chasing. *Cognition*, 126(2): 224–233, 2013.

Gergely, G., Nádasdy, Z., Csibra, G., and Bíró, S. Taking the intentional stance at 12 months of age. *Cognition*, 56 (2):165–193, 1995.

Graves, A., Bellemare, M. G., Menick, J., Munos, R., and Kavukcuoglu, K. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1311–1320. JMLR. org, 2017.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Haber, N., Mrowca, D., Wang, S., Fei-Fei, L., and Yamins, D. L. Learning to play with intrinsically-motivated self-aware agents. In *Advances in Neural Information Processing Systems*, 2018.

Houthooft, R., Chen, X., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. Vime: Variational information maximizing exploration. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1109–1117. Curran Associates, Inc., 2016.

Hus, V. and Lord, C. The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores. *Journal of autism and developmental disorders*, 44(8):1996–2012, 2014.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Jones, W. and Klin, A. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504(7480):427–431, 2013.

Ke, N. R., Singh, A., Touati, A., Goyal, A., Bengio, Y., Parikh, D., and Batra, D. Learning dynamics model in reinforcement learning by incorporating the long term future. *arXiv preprint arXiv:1903.01599*, 2019.

Kidd, C., Piantadosi, S. T., and Aslin, R. N. The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, 7(5):e36399, 2012.

Linke, C., Ady, N. M., White, M., Degris, T., and White, A. Adapting behaviour via intrinsic reward: A survey and empirical study. *arXiv preprint arXiv:1906.07865*, 2019.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

Mohamed, S. and Rezende, D. J. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pp. 2125–2133, 2015.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L., Tenenbaum, J. B., and Yamins, D. L. K. Flexible neural representation for physics prediction. *arXiv preprint arXiv:1806.08047*, June 2018.

Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2721–2730. JMLR. org, 2017.

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

Oudeyer, P.-Y., Baranes, A., and Kaplan, F. Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In *Intrinsically motivated learning in natural and artificial systems*, pp. 303–365. Springer, 2013.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. *arXiv preprint arXiv:1705.05363*, 2017.

Pathak, D., Gandhi, D., and Gupta, A. Self-supervised exploration via disagreement. *arXiv:1906.04161*, 2019.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978.

Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990 – 2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, Sept 2010.

Schmidhuber, J. Unsupervised minimax: Adversarial curiosity, generative adversarial networks, and predictability minimization. *arXiv preprint arXiv:1906.04493*, 2019.

Settles, B. *Active Learning*, volume 18. Morgan & Claypool Publishers, 2011.

Seung, H. S., Opper, M., and Sompolinsky, H. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294. ACM, 1992.

Smith, K., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J., and Ullman, T. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems*, pp. 8983–8993, 2019.

Stadie, B., Levine, S., and Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

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

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

Wang, Y., Gao, Z., Long, M., Wang, J., and Yu, P. S. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. *arXiv preprint arXiv:1804.06300*, 2018.

Wellman, H. M. *The child's theory of mind.* The MIT Press, 1992.

Wu, Y.-H., Fan, T.-H., Ramadge, P. J., and Su, H. Model imitation for model-based reinforcement learning. *ArXiv*, abs/1909.11821, 2019.