frontiers in Neurorobotics

METHODS
published: 22 January 2019
doi: 10.3389/fnbot.2018.00088

Boredom-Driven Curious Learning by Homeo-Heterostatic Value Gradients

Yen Yu, Acer Y. C. Chang and Ryota Kanai*

Araya, Inc., Tokyo, Japan

This paper presents the Homeo-Heterostatic Value Gradients (HHVG) algorithm as a formal account on the constructive interplay between boredom and curiosity which gives rise to effective exploration and superior forward model learning. We offer an instrumental view of action selection, in which an action serves to disclose outcomes that have intrinsic meaningfulness to an agent itself. This motivated two central algorithmic ingredients: devaluation and devaluation progress, both underpin agent's cognition concerning intrinsically generated rewards. The two serve as an instantiation of homeostatic and heterostatic intrinsic motivation. A key insight from our algorithm is that the two seemingly opposite motivations can be reconciled—without which exploration and information-gathering cannot be effectively carried out. We supported this claim with empirical evidence, showing that boredom-enabled agents consistently outperformed other curious or explorative agent variants in model building benchmarks based on self-assisted experience accumulation.

Keywords: curiosity, boredom, goal-directedness, intrinsic motivation, outcome devaluation, satiety, homeostatic motivation, heterostatic motivation

OPEN ACCESS

Edited by:
Vieri Giuliano Santucci,
Istituto di Scienze e Tecnologie della Cognizione (ISTC), Italy

Reviewed by:
Erhard Wieser,
Technische Universität München, Germany
James Danckert,
University of Waterloo, Canada

****Correspondence:***
Yen Yu
yen.yu@araya.org

Received: *29 April 2018*
Accepted: *19 December 2018*
Published: *22 January 2019*

Citation:
Yu Y, Chang AYC and Kanai R (2019) Boredom-Driven Curious Learning by Homeo-Heterostatic Value Gradients. Front. Neurorobot. 12:88. doi: 10.3389/fnbot.2018.00088

1. INTRODUCTION

In this study, we present an instrumental view of action selection, in which an action serves to disclose outcomes that have intrinsic meaningfulness—i.e., that hold epistemic values—to an agent itself. The implication of this statement is twofold: (1) for agents whose innate goal appeals to their own knowledge gain, the occurrence of curiosity rests upon the devaluation of known knowledge (and hence goal-directedness); (2) boredom—consequential to devaluation—and curiosity entail a mutually reinforcing cycle for such kind of (meaningful) disclosure to ensue.

Animal studies have shown that learning stimulus-response (S-R) associations through action-outcome reinforcement is but one facet of instrumental behavior. Internally, animals may build models that assign values to reappraise experienced outcomes. This expands the landscape of instrumental behavior to include the stimulus-outcome-response (S-O-R) learning system—or goal-directed learning (Balleine and Dickinson, 1998). Goal-directed behavior is known in both empirical and computational approaches to support adaptive and optimal action selection (Adams and Dickinson, 1981; Adams, 1982; Mannella et al., 2016). Central to such behavioral adaptiveness is devaluation. This means for a given action-outcome pair the associated reinforcing signal is no longer monotonic. Instead, an outcome value will change with reappraisals in accordance with an agent's internal goal.

One classic paradigm of devaluation manipulates an agent's level of satiation based on food accessibility, leading to altered behavioral patterns. In the context of epistemic disclosure, an analogy can be drawn between devaluation and the emergence of boredom, in which one's assimilation of knowledge reduces the value of similar knowledge in future encounters.

The relationship between boredom and outcome devaluation has a long history in psychological research. Empirical findings indicated that boredom is reportedly accompanied by negative affective experiences, suggesting that experienced outcomes are intrinsically evaluated and considered as less valuable (Perkins and Hill, 1985; Vodanovich et al., 1991; Fahlman et al., 2009; van Tilburg and Igou, 2012; Bench and Lench, 2013).

Psychophysiological studies also demonstrated that boredom plays an active role in eliciting information-seeking behaviors. Subjects showing higher levels of reported boredom are accompanied by increased autonomic arousal, such as heart rate and galvanic skin response. These findings are in line with our key notion that boredom intrinsically and actively drives learning behaviors (Berlyne, 1960; London et al., 1972; Harris, 2000). Note, however, that this notion is contested and a matter of unsettled debate (e.g., Eastwood et al., 2012; Fahlman et al., 2013; Merrifield and Danckert, 2014; Danckert et al., 2018). It is therefore worth pointing out that boredom may be accompanied by a low arousal state (Barmack, 1939; Geiwitz, 1966; Mikulas and Vodanovich, 1993; Pattyn et al., 2008; Vogel-Walcutt et al., 2012).

A finding by Larson (1990) invites the speculation that a task set may interact with boredom, thereby modifying a subject's behavioral pattern to follow either low or high arousal states. This means boredom may merely signal a state of disengagement. Whether an agent's cognitive resources can be freely allocated to re-engage another task inherently depends upon the existence of a prohibiting condition. Larson's (1990) participants, who reported boredom and were later rated with low scores in creative writing, were by design not allowed to disengage from the essay-writing task. Other theories, on the other hand, suggested that boredom is associated with increase in creativity (Schubert, 1977, 1978; Harris, 2000).

We thus postulate that, in the absence of any *a priori* cognitive or behavioral constraints, a state of boredom is followed by an attempt to diversify one's experience. That is, boredom begets exploration. This is in line with Vodanovich and Kass's (1990) notion of boredom in "inspiring a search for change and variety" and Zuckerman's (2008) "sensation-seeking." Sensation-seeking (Zuckerman, 1971, 2008; Kass and Vodanovich, 1990; Dahlen et al., 2005) is categorized as a personality trait, tightly linked to boredom susceptibility (Zuckerman et al., 1978). High sensation seekers get bored more easily, suggesting that individuals susceptible to boredom are predisposed to seek novel sensations. As a result, a learner who is also a novelty-seeker may have a world model that generalizes better. In our framework, receiving novel sensations is formalized as planning to visit states where an agent can effectively learn faster (i.e., the agent gets bored quicker). This effect is then treated as an intrinsic reward, prompting an agent to continue experiencing the state before the reward is depleted.

A recent computational modeling tapped into a similar theme (Gomez-Ramirez and Costa, 2017), where boredom facilitates exploration. However, our work differs from that of Gomez-Ramirez and Costa (2017) in that our model permits a simple form of agency (by having an action policy) and focuses on learning. Additionally, their exploration may favor predictable state space, whereas our agent will treat high predictability as an intrinsically non-rewarding state.

Finally, in psychology studies, the term boredom usually comes under two distinct constructs: a state of boredom and boredom proneness (Elpidorou, 2014, 2017; Mugon et al., 2018). Boredom proneness is regarded as the psychological predisposition of an individual to experience boredom which poses a systematic impact on one's social and psychological well-being. By contrast, a state of boredom is seen as a transient, regulatory signal that prompts one's behaviors into alignment with its goal-directedness (Elpidorou, 2017). In this sense, our model conceptually encompasses the function of the state boredom regulatory signal.

Curiosity, irrespective of being a by-product of external goal-attainment or an implicit goal in and of an agent itself, is often ascribed as a correlate of information-seeking behavior (Gottlieb et al., 2013). Behaviors exhibiting curious quality are observed in humans and animals alike, suggesting an universal role of curiosity in shaping one's fitness in terms of survival chance. Though the exact neural mechanism underlying the emergence of curious behavior still remains obscure, current paradigms have their focus on (1) novelty disclosure and (2) uncertainty reduction aspects of information-seeking (Bellemare et al., 2016; Friston et al., 2017; Ostrovski et al., 2017; Pathak et al., 2017). Indeed, both aspects can be argued to improve agent's fitness in epistemic landscape if the agent elects to incorporate the novelty or uncertainty.

Both boredom and curiosity are tightly connected to the notion of intrinsic motivation. Specifically, the occurrence of boredom and curiosity can be mapped to homeostatic and heterostatic motivations, respectively. The homeostatic and heterostatic motivations as two important classes of intrinsic motivation have been extensively reviewed in Oudeyer and Kaplan (2009). Simply, a homeostatic motivation drives a system to compensate perturbations in order to reach some equilibrial state. A heterostatic motivation is the opposite of a homeostatic motivation. A system that is driven by heterostatic motivations will self-perturb out of its equilibrium. In our formalism, predictive model learning and policy learning, each respectively induces boredom and curiosity, suggesting that the two classes of motivation can in fact be complementary when the two learning tasks are carried out concurrently. Our contribution thus pertains to the reconciliation of homeo-heterostatic motivations.

2. MARKOV DECISION PROCESS

In what follows, we briefly review preliminaries for the ensuing algorithm. We focus on well-established themes surrounding typical reinforcement learning, including Markov Decision Process and value gradients as a policy optimisation technique.

In Markov Decision Process (MDP) one considers the tuple $(S, A, R, P, \pi, \gamma)$. S and A are spaces of real vectors whose member, $\boldsymbol{s} \in S$ and $\boldsymbol{a} \in A$, represent states (or sensor values) and actions. R is some reward function defining the mapping $R : S \times A \rightarrow \mathbb{R}$. The probabilities associated with states and actions are given by the forward model $P(S'|A = \boldsymbol{a}, S = \boldsymbol{s})$ and

the action policy $\pi(A|S = \boldsymbol{s})$. Throughout the paper we use the 'prime' notation, e.g., $\boldsymbol{s}'$, to represent one time step into the future: $\boldsymbol{s}' = \boldsymbol{s}(t+1)$.

The goal of MDP is to optimally determine the action policy π^* such that the expected cumulative reward over a finite (or infinite) horizon is maximized. Considering a finite horizon problem with discrete time, $t \in [0, T]$, this is equivalent to $\pi^* = \arg\max_\pi \mathbb{E}_{\boldsymbol{a}\sim\pi}\left[\sum_{t=0}^{T} \gamma^t R(\boldsymbol{s}(t), \boldsymbol{a}(t))\right]$, where $\gamma \in [0, 1]$ is the discount factor.

Many practical approaches for solving MDP often resort to approximating state-action value $q(\boldsymbol{a}, \boldsymbol{s})$ or state value $v(\boldsymbol{s})$ functions (Sutton and Barto, 1998; Mnih et al., 2013; Heess et al., 2015; Lillicrap et al., 2015). These value functions are given in the Bellman equation

$$\begin{aligned} v(\boldsymbol{s}) &= \mathbb{E}_{\pi(\boldsymbol{a}|\boldsymbol{s})}\Big[R(\boldsymbol{a}, \boldsymbol{s}) + \gamma q(\boldsymbol{a}, \boldsymbol{s})\Big] \\ &= \mathbb{E}_{\pi(\boldsymbol{a}|\boldsymbol{s})}\Big[R(\boldsymbol{a}, \boldsymbol{s}) + \gamma \mathbb{E}_{P(\boldsymbol{s}'|\boldsymbol{a},\boldsymbol{s})}[v(\boldsymbol{s}')]\Big] \end{aligned} \tag{1}$$

When differentiable forward model and reward function are both available, policy gradients can be analytically estimated using value gradients (Fairbank and Alonso, 2012; Heess et al., 2015).

3. HOMEO-HETEROSTATIC VALUE GRADIENTS

This section describes formally the algorithmic structure and components of the Homeo-Heterostatic Value Gradients, or HHVG. The naming of HHVG suggests its connections with homeostatic and heterostatic intrinsic motivations. A detailed review on homeostatic and heterostatic motivations are given in Oudeyer and Kaplan (2009). Briefly, a homeostatic motivation encourages an organism to occupy a set of predictable, unsurprising states (i.e., a *comfort zone*). Whereas, a heterostatic motivation does the opposite; curiosity belongs to this category.

The algorithm offers a reconciliation between the two seemingly opposite qualities and concludes with their cooperative nature. Specifically, the knowledge an organism maintains about its comfort zone helps instigate outbound heterostatic drives. In return, satisfying heterostatic drives broadens the organism's extent of comfort zone. As a consequence, the organism not only improves its fitness in terms of homeostatic outreach but also becomes effectively curious.

3.1. Nomenclature and Notations

It is instructive to overview the nomenclature of the algorithm. We consistently associate homeostatic motivation with the emergence of *boredom*, which reflects the result of having incorporated novel information into one's knowledge, thereby diminishing the novelty to begin with. This is conceptually compatible with outcome *devaluation* or induced satiety in instrumental learning. *Devaluation progress* is therefore referred to as one's epistemic achievement. That is, the transitioning of a priori knowledge to one of having assimilated otherwise unknown information. The devaluation progress is interpreted as an instantiation of intrinsic reward. The drive to maintain steady rewards conforms to a heterostatic motivation.

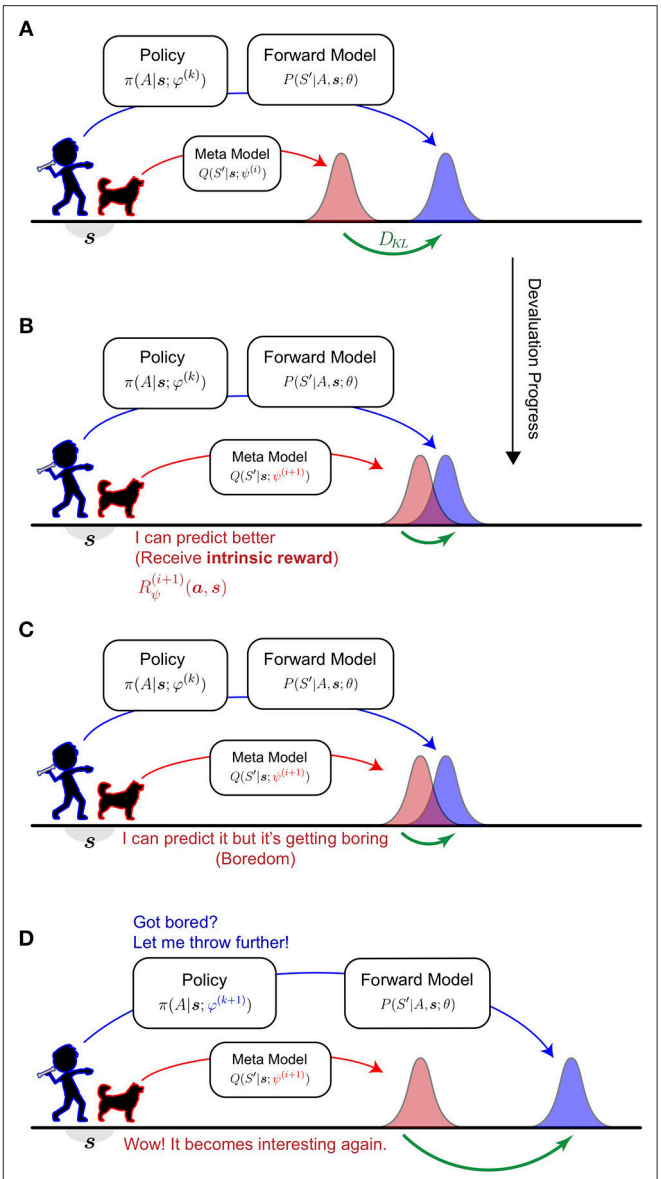

FIGURE 1 | Intuitive understanding of the Homeo-Heterostatic Value Gradients (HHVG) algorithm. **(A)** The algorithm can be interpreted as the cooperative interplay between a thrower (kid; blue) and a catcher (dog; red). The thrower is equipped with a forward model that estimates its aiming and is controlled by an action policy. Without knowing the thrower's policy, the catcher (meta-model), in order to make good catches, infers where the thrower is aiming on average. **(B)** The catcher is interested in novel, unpredicted throws. Whenever the catcher improves its predictive power some intrinsic reward (devaluation progress) is generated. **(C)** As the catcher progresses further, similar throws become highly predictable, thus inducing a sense of boredom. **(D)** To make the interplay interesting again, the thrower is driven to devise new throws, so that the catcher can afford to make further progress. By repeating **(A,B)** the thrower has attempted diverse throws and known well about its aim. At the same time, the catcher will assume a vantage point for any throw.

The notation $\mathcal{L}(\cdot)$ consistently denotes loss functions throughout the paper; any variables on which the loss function

depends are always made explicit. There are occasions where we abbreviated the loss function to avoid clutters. A definition such as $\mathcal{L}_{mm}(\psi) := \mathcal{L}(\boldsymbol{a}, \boldsymbol{s}; \psi, \theta)$ is then given upon first appearance. Here, the subscript *mm* indicates *meta-model*. One may tell in this example that the symbols $\boldsymbol{a}$, $\boldsymbol{s}$, and θ on the right hand side are temporarily omitted. This means the optimisation procedure for the meta-model concerns only the parameter ψ. Similarly, this applies to $\mathcal{L}_{fm}$, $\mathcal{L}_{vf}$, and $\mathcal{L}_{ap}$, where the subscripts stand for *forward model*, *value function*, and *action policy*. The symbol $\mathcal{N}$ is reserved for Normal distribution.

3.2. Intuition

An intuitive understanding of HHVG is visualized in **Figure 1**. Imagine the interplay between a thrower and their counterpart—a catcher. The catcher anticipates where the thrower is aiming and makes progress by improving its prediction. The thrower, on the other hand, keeps the catcher engaged by devising novel aims. Over time, the catcher knows well what the thrower is capable of, whilst the thrower has attempted a wide spectrum of pitches.

In the algorithm, the thrower is represented by a forward model attached to a controller (policy) and the catcher a "meta-model." We unpack and report them individually. Procedural information is summarized in **Algorithm 1**.

3.3. Forward Model

We start by specifying at current time the state and action sample as $\boldsymbol{s}$ and $\boldsymbol{a}$. The forward model describes the probability distribution over future state S', given $\boldsymbol{s}$, $\boldsymbol{a}$, and parameter θ.

$$P(S'|A = \boldsymbol{a}, S = \boldsymbol{s}; \theta) \tag{2}$$

The entropy associated with S', conditioned on $\boldsymbol{s}$ and $\boldsymbol{a}$, gives a measure of the degree to which S' is informative on average. We referred to this measure as one of *interestingness*. Note this is a different concept from the "interestingness" proposed by Schmidhuber (2008), which is the first-order derivative of compressibility.

3.4. Boredom, Outcome Devaluation, and Meta-Model

Boredom, in common understanding, is perhaps not unfamiliar to most people under the situation of being exposed to certain information which one has known well by heart. It is the opposite of being interested. In the current work, we limited the exposure of information to those being disclosed by one's actions.

To mark the necessity of boredom, we first identify the limitation of a naive instantiation of curiosity; then, we show that the introduction of boredom serves to resolve this limitation.

Consider the joint occurrence of future state S' and action A: $P(S', A|S = \boldsymbol{s}; \theta, \varphi)$. This can be derived from the product rule of probability using $P(S'|A = \boldsymbol{a}, S = \boldsymbol{s}; \theta)$ (as shown Equation 2) and action policy $\pi(A|S = \boldsymbol{s}; \varphi)$, parametrised by φ (action policy is revisited in section 3.6).

A naive approach to curiosity is by optimizing the action policy, such that A is predictive of maximum *interestingness* (see section 3.3) about the future.

However, this naive approach would certainly lead to the agent behaving habitually and, as a consequence, becoming obsessive about a limited set of outcomes. In other words, a purely interestingness-seeking agent is a darkroom agent (see section 3.7; also Friston et al., 2012 for related concept).

Algorithm 1 Homeo-heterostatic value gradients

Variables
- outer loop time t
- gradient step counter ℓ, i, j, k
- state $\boldsymbol{s}^t := \boldsymbol{s}(t)$ and action $\boldsymbol{a}^t := \boldsymbol{a}(t)$
- learning rate $\lambda^\theta, \lambda^\psi, \lambda^\nu, \lambda^\varphi$
- discount factor γ
- experience pool $\mathcal{D}$

Models and parameters
- forward model $P(S'|\boldsymbol{s}, \boldsymbol{a}; \theta)$
- meta-model $Q(S'|\boldsymbol{s}; \psi)$
- value approximator $v(\boldsymbol{s}; \nu)$
- action policy $\pi(A|\boldsymbol{s}; \varphi)$

Objectives
- forward-model learning $\mathcal{L}_{fm}(\theta)$
- meta-model learning $\mathcal{L}_{mm}(\psi)$ ▷ Eq.4
- value learning $\mathcal{L}_{vf}(\nu)$ ▷ Eq.6
- policy learning $\mathcal{L}_{ap}(\varphi)$ ▷ Eq.8

for $t = 0 \ldots T$ **do**
- From $\boldsymbol{s}^t$, sample action $\boldsymbol{a}^t \sim \pi(\cdot|\boldsymbol{s}^t; \varphi)$
- Perform $\boldsymbol{a}^t$ and advance to $\boldsymbol{s}^{t+1}$
- Insert tuple $\left(\boldsymbol{s}^t, \boldsymbol{a}^t, \pi(\boldsymbol{a}^t|\boldsymbol{s}^t), \boldsymbol{s}^{t+1}\right)$ into $\mathcal{D}$
- Sample $\mathcal{D}$ and train forward model:
  - $\mathcal{L}_{fm}(\theta) := \mathcal{L}(\boldsymbol{s}', \boldsymbol{a}, \boldsymbol{s}; \theta) = \|\boldsymbol{s}' - f(\boldsymbol{a}, \boldsymbol{s}; \theta)\|^2$ ▷ Eq.14
  - $\theta^{(\ell+1)} \leftarrow \theta^{(\ell)} - \lambda_\theta \nabla_\theta \mathcal{L}_{fm}(\theta^{(\ell)})$
- Value learning (M updates, see **Algorithm 2**)
- Sample $\mathcal{D}$ and perform devaluation:
  - $\psi^{(i+1)} \leftarrow \psi^{(i)} - \lambda_\psi \nabla_\psi \mathcal{L}_{mm}(\psi^{(i)})$
- Sample $\mathcal{D}$ and train action policy:
  - evaluate $R_\psi^{(i+1)} = \mathcal{L}_{mm}(\psi^{(i)}) - \mathcal{L}_{mm}(\psi^{(i+1)})$
  - evaluate $\nu' = v(\boldsymbol{s}'; \nu^{(j+M)})$
  - $w \leftarrow \pi(\boldsymbol{a}|\boldsymbol{s}; \varphi^{(k)})/\pi(\boldsymbol{a}|\boldsymbol{s}; \varphi^{(<k)})$
  - $\varphi^{(k+1)} \leftarrow \varphi^{(k)} + \lambda^\varphi \nabla_\varphi w \mathcal{L}_{ap}(\varphi^{(k)})$ given $R_\psi^{(i+1)}$, ν'

Algorithm 2 Fitted Policy Evaluation [cf. Heess et al. (2015)]

Given
- outer loop time t
- experience pool $\mathcal{D}$
- value function $v(\boldsymbol{s}; \nu^{(j)})$
- gradient step counter i, j, k

Clone parameter $\tilde{\nu} \leftarrow \nu^{(j)}$

for $m = 1 \ldots M$ **do**
- Sample $\left(\boldsymbol{s}^\tau, \boldsymbol{a}^\tau, \pi(\boldsymbol{a}^\tau|\boldsymbol{s}^\tau; \varphi^{(<k)}), \boldsymbol{s}^{\tau+1}\right)$ from $\mathcal{D}$ $(\tau < t)$
- Evaluate $R_\psi^{(i+1)} = \mathcal{L}_{mm}(\psi^{(i)}) - \mathcal{L}_{mm}(\psi^{(i+1)})$
- $y = R_\psi^{(i+1)} + \gamma v(\boldsymbol{s}^{\tau+1}; \tilde{\nu})$
- $w = \pi(\boldsymbol{a}^\tau|\boldsymbol{s}^\tau; \varphi^{(k)})/\pi(\boldsymbol{a}^\tau|\boldsymbol{s}^\tau; \varphi^{(<k)})$
- Apply updates $\nu^{(j+m)} \leftarrow \nu^{(j+m-1)} - \nabla_\nu \frac{w}{2}\left(y - v(\boldsymbol{s}; \nu^{(j+m-1)})\right)^2$
- Every C updates, $\tilde{\nu} \leftarrow \nu^{(j+m)}$

Such obsession with limited outcomes poses a caveat—the agent has no recourse to inform itself about prior exposure of similar sensations. If the agent is otherwise endowed with this capacity, namely, by assimilating previous experiences into summary statistics, an ensuing sense of boredom would be induced. The induction of boredom essentially causes the agent to value the same piece of information less, thus changing the agent's perception toward interestingness. If the agent were to pursue the same interestingness-seeking policy, a downstream effect of boredom would drive the agent to seek out other information that could have been known. This conception amounts to an implicit goal of *devaluating* known outcomes.

To this end, we introduce the following meta-model Q to represent *a priori* knowledge about the future. Note that Q is a conditional probability function over S' and is not to be confused with a state-action value function $q(\boldsymbol{a}, \boldsymbol{s})$ in MDP. The meta-model, parametrised by ψ, is an approximation to the *true* marginalization of joint probability $P(S', A|S = \boldsymbol{s}; \theta, \varphi)$ over A:

$$\begin{aligned} Q(S'|S = \boldsymbol{s}; \psi) &\approx P(S'|S = \boldsymbol{s}; \theta, \varphi) \\ &= \sum_A \Big[P(S', A|\boldsymbol{s}; \theta, \varphi)\Big] \\ &= \sum_A \Big[P(S'|A, \boldsymbol{s}; \theta)\pi(A|\boldsymbol{s}; \varphi)\Big] \end{aligned} \tag{3}$$

We associate the occurrence of boredom, or, synonymously, outcome devaluation, with minimizing the devaluation objective with respect to ψ. The devaluation objective is given by the Kullback-Leibler (KL) divergence:

$$\begin{aligned} \mathcal{L}_{mm}(\psi) &:= \mathcal{L}(\boldsymbol{a}, \boldsymbol{s}; \psi, \theta) \\ &= D_{KL}\left[P(\boldsymbol{s}'|\boldsymbol{a}, \boldsymbol{s}; \theta)\middle\|Q(\boldsymbol{s}'|\boldsymbol{s}; \psi)\right] \end{aligned} \tag{4}$$

3.5. Devaluation Progress, Intrinsic Reward, and Value Learning

Through the use of KL-divergence in Equation 4, we emphasize the complementary nature of devaluation in relation to a knowledge-gaining process. That is to say, devaluation results in information gain for the agent. This, in fact, can be regarded as cognitively rewarding and, thus, serves to motivate our definition of intrinsic reward.

One rewarding scenario happens when $Q(S'|\boldsymbol{s}; \psi)$ has all the information there is to be possessed by A about S'. A is therefore rendered redundant. One may speculate, at this point, the agent could opt for inhibiting its responses. Disengaging actions potentially saves energy which is rewarding in biological sense. This outcome is in line with the "opportunity cost model" proposed by Kurzban et al. (2013). In their model, boredom is seen as a resource regulatory signal which drives an agent to disengage the current task and curb the computational cost. As a consequence, the occurrence of boredom may encourage re-allocation of computational processes to alternative higher-value activities (Kurzban et al., 2013).

Alternatively, the agent may attempt to develop new behavioral repertoires, bringing into S' new information (i.e., novel outcomes) that is otherwise unknown to Q. The ensuing sections will focus on this line of thinking.

From Equation 4, we construct the quantity *devaluation progress* to represent an intrinsically motivated reward. The devaluation progress is given by the difference between KL-divergences before and after devaluation [as indicated by the superscript $(i+1)$]:

$$\begin{aligned} R_\psi^{(i+1)}(\boldsymbol{a}, \boldsymbol{s}) &:= \mathcal{L}(\boldsymbol{a}, \boldsymbol{s}; \psi^{(i)}, \theta) - \mathcal{L}(\boldsymbol{a}, \boldsymbol{s}; \psi^{(i+1)}, \theta) \\ &= \mathcal{L}_{mm}(\psi^{(i)}) - \mathcal{L}_{mm}(\psi^{(i+1)}), \end{aligned} \tag{5}$$

Here, we write $R_\psi^{(i+1)}(\boldsymbol{a}, \boldsymbol{s})$ in accordance with notational convention in reinforcement learning, where reward is typically a function of state and action. Subscript ψ indicates the dependence of R on meta model parameter.

Having established the intrinsic reward, value learning is such that the value function approximator $v(\boldsymbol{s}; \nu)$ follows the Bellman equation $v(\boldsymbol{s}) = \mathbb{E}_{\boldsymbol{a}}[R(\boldsymbol{a}, \boldsymbol{s}) + \gamma \mathbb{E}_{\boldsymbol{s}'}[v(\boldsymbol{s}')]]$. In practice, we minimize the objective with respect to ν:

$$\begin{aligned} \mathcal{L}_{vf}(\nu) &:= \mathcal{L}(\boldsymbol{s}', \boldsymbol{a}, \boldsymbol{s}; \nu) \\ &= \left\|y - v(\boldsymbol{s}; \nu)\right\|^2 \\ y &= R_\psi^{(i+1)}(\boldsymbol{a}, \boldsymbol{s}) + \gamma v(\boldsymbol{s}'; \tilde{\nu}) \end{aligned} \tag{6}$$

3.6. Policy Optimisation

We define action policy at state $S = \boldsymbol{s}$ as the probability distribution over A with parameter φ:

$$\pi(A|S = \boldsymbol{s}; \varphi) \tag{7}$$

Our goal is to determine the policy parameter φ that maximizes the expected sum of future discounted rewards. One approach is by applying Stochastic Value Gradients (Heess et al., 2015) and maximizes the value function. We thus define our policy objective as follows (notice the negative sign; we used a gradient update rule that defaults to minimization):

$$\begin{aligned} \mathcal{L}_{ap}(\varphi) &:= \mathcal{L}(\boldsymbol{s}', \boldsymbol{a}, \boldsymbol{s}; \theta, \psi^{(i)}, \psi^{(i+1)}, \nu, \varphi) \\ &= -\mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s};\varphi)}\Big[R_\psi^{(i+1)}(\boldsymbol{a}, \boldsymbol{s}) + \gamma \mathbb{E}_{\boldsymbol{s}'\sim P(\cdot|\boldsymbol{a},\boldsymbol{s};\theta)}\big[v(\boldsymbol{s}'; \nu)\big]\Big] \end{aligned} \tag{8}$$

3.7. Remarks on Homeostatic and Heterostatic Regulations

Oudeyer and Kaplan (2009) outlined the distinctions between two important classes of intrinsic motivation: homeostatic and heterostatic. A homeostatic motivation is one that can be satiated, leading to a certain equilibrium behaviorally; whereas a heterostatic motivation topples the agent, thus preventing it from occupying habitual states.

Our algorithm entails regulations relating to both classes of intrinsic motivation. Specifically, the devaluation objective (Equation 4) realizes the homeostatic aspect due to its connection with induced satiety. On the other hand, the devaluation progress (Equation 5) introduced for policy optimisation instantiates a heterostatic drive to agent's behavioral pattern.

Heterostasis is motivated by the agent pushing itself toward novelty and away from devalued, homeostatic states (as revealed at the end of this section in Equation 13). This statement is shown formally by replacing the reward $R_\psi^{(i+1)}(\boldsymbol{a},\boldsymbol{s})$ in Equation 8, with Equation 5. We then arrived at the following form involving expected KL-divergence:

$$\begin{aligned}
&- \mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s};\varphi)}\Big[D_{KL}[P(\boldsymbol{s}'|\boldsymbol{a},\boldsymbol{s};\theta)\|Q(\boldsymbol{s}'|\boldsymbol{s};\psi^{(i)})] \\
&- D_{KL}[P(\boldsymbol{s}'|\boldsymbol{a},\boldsymbol{s};\theta)\|Q(\boldsymbol{s}'|\boldsymbol{s};\psi^{(i+1)})]\Big] \\
&- \mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s};\varphi)}\mathbb{E}_{\boldsymbol{s}'\sim P(\cdot|\boldsymbol{a},\boldsymbol{s};\theta)}\Big[\nu(\boldsymbol{s}';\nu)\Big] \\
&= -\Big\{I(S':A|S=\boldsymbol{s};\psi^{(i)},\varphi,\theta) - I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi,\theta) \\
&+ \mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s};\varphi)}\mathbb{E}_{\boldsymbol{s}'\sim P(\cdot|\boldsymbol{a},\boldsymbol{s};\theta)}\Big[\nu(\boldsymbol{s}';\nu)\Big]\Big\} \qquad (9)
\end{aligned}$$

Notice that the expected devaluation progress becomes the difference between conditional mutual information I before ($\psi^{(i)}$) and after devaluation ($\psi^{(i+1)}$).

Assume, for the moment, that the agent is equipped with devaluation capacity only. In other words, we replace the devaluation progress and fall back on devaluation objective, $R := \mathcal{L}_{mm}(\psi)$ (cf. Equation 5). The agent is now interestingness-seeking with homeostatic regulation. We further suppose that the dynamics of ψ and φ evolve in tandem, which gives

$$\begin{aligned}
I(S':A|S=\boldsymbol{s};\psi^{(i)},\varphi^{(k)}) &\to I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k)}) \\
&\to I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k+1)}) \\
&\to I(S':A|S=\boldsymbol{s};\psi^{(i+2)},\varphi^{(k+1)}) \to \dots \qquad (10)
\end{aligned}$$

In practice, the nature of devaluation and policy optimisation often depends on replaying agent's experience. Taking turn applying gradient updates to ψ and φ creates a self-reinforcing cycle that drives the policy to converge toward a point mass. For instance, if the policy is modeled by some Gaussian distribution, this updating scheme would result in infinite precision (zero spread).

For curiosity, however, such parameter dynamics should not be catastrophic if we subsume the homeostatic regulation and ensure the preservation of the relation given in Equation 11:

$$\begin{aligned}
I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k)}) &\le I(S':A|S=\boldsymbol{s};\psi^{(i)},\varphi^{(k)}) \\
&\le I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k+1)}) \\
\Rightarrow -I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k)}) &+ I(S':A|S=\boldsymbol{s};\psi^{(i)},\varphi^{(k)}) \\
&\le I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k+1)}) \qquad (11)
\end{aligned}$$

This equation holds because the devaluation process on average has a tendency to make A less informative about S', after which A is perturbed to encourage a new S' less predictable to Q. By rearranging the equation such that the left hand side remains positive, we have arrived at a lower bound on $I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k+1)})$ which recovers the expected devaluation progress.

Equation 12 summarizes the argument associated with Equations (10, 11).

$$\begin{aligned}
\varphi^{(k+1)} &= \arg\max_{\varphi^{(k)}}\Big[I(S':A|S=\boldsymbol{s};\psi^{(i)},\varphi^{(k)}) \\
&\quad - \min_{\bar{\psi}^{(i)}} I(S':A|S=\boldsymbol{s};\bar{\psi}^{(i)},\varphi^{(k)})\Big] \\
&\neq \arg\max_{\varphi^{(k)}}\Big[\min_{\psi^{(i)}} I(S':A|S=\boldsymbol{s};\psi^{(i)},\varphi^{(k)})\Big] \qquad (12)
\end{aligned}$$

Finally, we offer an intuition on how policy optimisation gives rise to heterostatic motivation. This is made clear from the optimized target $I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k+1)})$, found on the right hand side of Equation 11. It is instructive to re-introduce the true marginalization $P(S'|S=\boldsymbol{s};\theta,\varphi)$ from Equation 3; write:

$$\begin{aligned}
&I(S':A|S=\boldsymbol{s};\psi^{(i+1)},\varphi^{(k+1)}) = \sum_{\boldsymbol{a}}\pi(\boldsymbol{a}|\boldsymbol{s};\varphi^{(k+1)}) \\
&\sum_{\boldsymbol{s}'}P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a};\theta)\log\frac{P(\boldsymbol{s}'|\boldsymbol{a},\boldsymbol{s};\theta)}{Q(\boldsymbol{s}'|\boldsymbol{s};\psi^{(i+1)})} = \sum_{\boldsymbol{a}}\pi(\boldsymbol{a}|\boldsymbol{s};\varphi^{(k+1)}) \\
&\sum_{\boldsymbol{s}'}P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a};\theta)\log\frac{P(\boldsymbol{s}'|\boldsymbol{a},\boldsymbol{s};\theta)}{P(\boldsymbol{s}'|\boldsymbol{s};\theta,\varphi^{(k+1)})}\frac{P(\boldsymbol{s}'|\boldsymbol{s};\theta,\varphi^{(k+1)})}{Q(\boldsymbol{s}'|\boldsymbol{s};\psi^{(i+1)})} \\
&= I(S':A|S=\boldsymbol{s};\varphi^{(k+1)}) + D_{KL}\big[P(\boldsymbol{s}'|\boldsymbol{s};\theta,\varphi^{(k+1)})\|Q(\boldsymbol{s}'|\boldsymbol{s};\psi^{(i+1)})\big] \qquad (13)
\end{aligned}$$

Simply, the optimized policy is such that the agent increases the conditional mutual information and is pushed away (via increasing the KL-divergence) from its homeostatic state Q.

4. IMPLEMENTATION CONSIDERATIONS

This section presents practical considerations when motivating the aforementioned agent using neural networks. These considerations were mainly for the ease of calculating KL-divergence analytically.

4.1. Forward Model

We assumed that the state follows some Gaussian distribution with mean $\boldsymbol{s}$ and covariance Σ. The future state is described by its mean $\boldsymbol{s}'$ according to the deterministic mapping $\boldsymbol{s}' = f(\boldsymbol{a},\boldsymbol{s};\theta)$, where $\boldsymbol{a}$ is the action sampled from policy. f represents a neural network with trainable parameter θ:

$$f(\boldsymbol{a},\boldsymbol{s};\theta) = \boldsymbol{A}\boldsymbol{s} + \left(\sum_\iota a_\iota \boldsymbol{B}^\iota\right)\boldsymbol{s} + \boldsymbol{C}\boldsymbol{a} + o \qquad (14)$$

$\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are approximations of Jacobian matrices and o a constant, all depending on θ. $\boldsymbol{B}$ is a three-way tensor indexed by ι along the first axis. This treatment is similar to Watter et al. (2015) (also cf. Karl et al., 2016), except that we considered a bilinear approximation and that, in the following sections, we used only the mean states in a deterministic environment.

The above formalism follows that $\boldsymbol{s}'$ has covariance matrix $\mathbb{E}[\boldsymbol{s}'\boldsymbol{s}'^{\intercal}] = \boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^{\intercal}$, where $\boldsymbol{J} = \left(\boldsymbol{A} + \sum_\iota a_\iota \boldsymbol{B}^\iota\right)$. The transition probability is then given by

$$P(S'|A = \boldsymbol{a}, S = \boldsymbol{s}; \theta) = \mathcal{N}\left(f(\boldsymbol{a}, \boldsymbol{s}; \theta), \boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^{\intercal}\right) \tag{15}$$

The model parameter θ represented four fully connected layers of width 512; the four layers were complemented by a residual connection, which was a single fully connected layer. We used rectified linear units (ReLU) as output nonlinearities. Next, four fully connected, linear layers each mapped the 512-dimensional output into vectors of dimension 16, 32, 8, and 1. These vectors were then reshaped into tensors and used as $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and o.

4.2. Meta Model

Our meta model was defined as a Gaussian distribution $Q(S'|S = \boldsymbol{s}; \psi) = \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}'; \psi)$, where the mean $\boldsymbol{\mu}'$ and covariance matrix $\boldsymbol{\Sigma}'$ are outputs of a neural network parametrized by ψ. Specifically, to construct the covariance matrix, we used the fact that the eigendecomposition of a positive semi-definite matrix always exists. This then means we can use neural networks to specify an orthogonal matrix $\boldsymbol{H}$ and a diagonal matrix $\boldsymbol{D}$, such that the covariance matrix is equivalent to:

$$\begin{aligned} \boldsymbol{\Sigma}' &= \boldsymbol{H}\boldsymbol{D}\boldsymbol{H}^{\intercal}, \quad \boldsymbol{D} = \mathrm{diag}(\boldsymbol{d}) \\ \boldsymbol{H} &= \boldsymbol{I} - 2\frac{\boldsymbol{u}\boldsymbol{u}^{\intercal}}{\|\boldsymbol{u}\|^2}, \end{aligned} \tag{16}$$

where $\boldsymbol{d}$ is a positive-valued vector that specifies the diagonal elements of $\boldsymbol{D}$. The second line of Equation 16 shows how an orthogonal matrix can be built from a real-valued vector $\boldsymbol{u}$, called Householder vector (Tomczak and Welling, 2016). $\boldsymbol{I}$ is an identity matrix.

The network architecture used to compute μ', $\boldsymbol{d}$, and $\boldsymbol{u}$ consisted of three trainable layers, each of which was identically structured. Three fully connected layers with ReLU activation functions, complemented by a residual connection, were followed by a linear, fully connected output layer. The output layer for $\boldsymbol{d}$ used a Softplus nonlinearity to ensure positive values.

We can, of course, let the neural network output a full matrix $\boldsymbol{X}$ and have $\boldsymbol{\Sigma}' = \boldsymbol{X}\boldsymbol{X}^{\intercal}$. However, our method is less costly when scaling up the problem dimension.

4.3. Policy and Value Functions

Both the policy and value functions were identically structured in terms of network architecture. They consisted of four fully connected layers with ReLU activation functions, complemented by a residual connection. This was then followed by a linear output layer. The outputs for the policy network were treated as logits of a categorical distribution over action space.

5. EXPERIMENT

One testable hypothesis that emerges from our previous remark—that boredom gives rise to novelty seeking policy (cf. KL-divergence term in Equation 13)—is that *boredom helps improve agent's forward model learning*. This is because novelty seeking essentially implies diversity in agent's experience. In other words, a boredom-driven curious agent must exhibit a tendency toward *exploration* and against *perseveration*. This tendency is critical when the agent was not given a training set (on which it based its forward model learning) but has to self-assist in accumulating one from scratch.

Briefly, an agent that tends to explore would appear to accumulate experience that reflects a more complete picture of the environment and, therefore, leads to a more accurate forward model. By contrast, if an agent perseverates, it can only afford to occupy a limited set of states, leaving its forward model an inadequate representation of the environment.

The primary goal and purpose of the ensuing experiments is thus to illustrate, with and without boredom, (1) the extent to which an agent explores and perseverates, and (2) the forward model performance.

To this end, we motivated a model pruning hierarchy on which the comparisons above were based. The model pruning hierarchy, as summarized in **Table 1** and section 5.3, provides a principled way to assess agent's behavior by progressive degrading model components. As a result, the difference between a boredom agent and a boredom-free curious agent or non-curious agent can be highlighted.

Exploration and perseveration were assessed qualitatively using Coverage Rate (CR) and Coverage Entropy (CE), reported in section 6. CR simply counts the number of states an agent has visited amongst all possible states. CE focuses on weighing the number of time steps a state was being occupied. CR thus indicates the proportion of the environment explored by the agent. Whereas, a CE curve declining over time indicates the agent tends to perseverate around a limited state space.

Forward model performance was assessed based on validation error. The validation set was sampled from the oracle dataset (see section 5.2). Contrary to self-assisted data accumulation, the oracle dataset was acquired by uniformly sampling the state-action grid. This dataset is therefore an idealized case to learn the best possible forward model.

TABLE 1 | Model pruning hierarchy that helps highlight the contribution of boredom and curiosity in regulating agent's exploration and perseveration.

	Oracle	P/RW	PG/GR	PG/IRS	C/PE	C/B
FM	✓	✓	✓	✓	✓	✓
AP		○	✓	✓	✓	✓
IR				○	✓	✓
VF					✓	✓
MM						✓

Ticks mark the existence or dependence of trainable network components; circles indicate independent intervention. Top row: P/RW, random-walk policy; PG/GR, policy gradients with rewards drawn from a Gaussian distribution; PG/IRS, policy gradients with intrinsic reward samples; C/PE, curiosity using forward model error; C/B, curiosity from boredom. First column: FM, forward model; AP, action policy; IR, intrinsic rewards; VF, value function approximator; MM, meta-model.

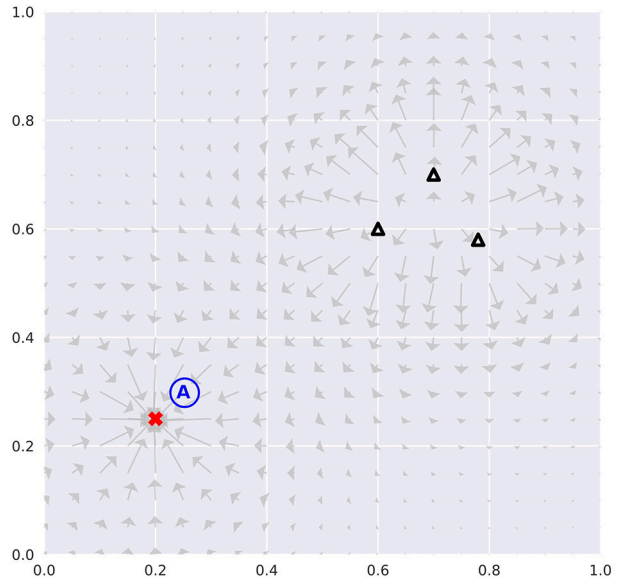

FIGURE 2 | Environmental configuration. The red cross represents an attractor, whilst black triangles repellers. Vector plots indicate the forces exerted if the agent assumed the positions with zero velocities. The initial position is set at the blue letter "A." This configuration remains identical cross all model variants and test runs.

Overall, we set the following constraints on training and environment conditions: (1) agent is responsible for assembling its own training set from scratch; (2) the probability of visiting different states is not uniformly distributed if the agent will commit to random walk; (3) the amount of time to accumulate training data points is limited.

5.1. Training Environment

Our agents were tested in a physics simulator, free of stochasticity, built to expand the classical Mountain Car environment (e.g., "MountainCar-v0" included in Brockman et al., 2016) into two-dimensional state space. The environment is analogous to the Mountain Car in ways that it has attractors and repellers that resemble hill- and valley-like landscapes (**Figure 2**). The presence of both structures serves as acceleration modifier to the agent. This makes state visitation biased toward attractors. Therefore, the acquisition of an accurate forward model necessitates planning visits to the vicinity of repellers.

The states an agent can occupy were defined as the tuple $(x, y, \dot{x}, \dot{y})$ in continuous real space. Positions $(x, y) \in [0, 1]^2$ were bounded in a unit square, whereas velocities $(\dot{x}, \dot{y})$ were not. Boundary condition resets x and y to zero velocities. However, it is possible for the agent to slide along the boundaries if its action goes in the direction parallel to the nearby boundary. We note that being trapped in the corners is possible; though an agent could potentially get itself unstuck if appropriate actions were carried out.

Agent's action policy was represented by a categorical distribution over accelerations in x and y directions. The distribution was defined on the interval $[-2.0, 2.0]^2$, evenly divided into a 11×11 grid. When an action is selected, the corresponding acceleration is modified according to forces exerted by the attractors and repellers.

Unlike the classical Mountain Car, our environment does not express external rewards, nor does it possess any states that are indicative of termination. Agents were allowed a pre-defined time limit ($T = 30,000$ steps; *Data Accumulation Phase* or DAP) to act without interruption. Agent's experiences in terms of state transitions were collected in a database, which was sampled from for training at each step. During DAP, learning rates for model parameters remained constant. After DAP (or *post*-DAP), agent entered an action-free stage lasted for $T = 30,000$, during which only sampling from own experience pool for forward model training was performed. Learning rate scheduling scheme was implemented at post-DAP.

An implementation of our training environment is available online [1].

5.2. Oracle Dataset

To contrast with self-assisted data accumulation, we constructed an oracle dataset. This dataset assumed unbiased state occupancy and action selection. We acquired the dataset by evenly dividing the state-action space into a $49 \times 49 \times 11 \times 11 \times 11 \times 11$ grid. Each state-action pair was passed to the physics simulator to evaluate the next state. The resultant tuple $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$ then represents one entry in the dataset. The training, testing, and validation sets were prepared by re-sampling the resulting dataset without replacement according to the ratio 0.8, 0.16, and 0.04.

A class of model referred to as Oracle, which consists of a forward model only (**Table 1**), was trained on this dataset. The Oracle model does not need to learn an action policy, as actions are already specified in the oracle dataset. The Oracle model was trained for 60,000 epochs. During training, the learning rate was scheduled according to test error. Benchmarking was performed on the validation set as part of model comparisons (see section 5.4).

The oracle dataset differs from the ones that are populated by an agent as it explores. For instance, some locations in the state space are essentially inaccessible to our agent due to the force exerted by the repellers. These locations greatly inform forward model learning, however, but are only present in the oracle dataset and available to the Oracle model.

5.3. Model Pruning

We defined five variants of our boredom-driven curious agent. With each variation, the agent receives cumulative reductions in network components. Theses reductions are summarized as model pruning hierarchy in **Table 1**.

The reason that we motivated model comparisons based on model pruning is to emphasize the contribution of boredom and curiosity in regulating agent's exploratiiveness and perseveration. Overall, as model pruning progresses the agent was deprived of functional constructs like devaluation progress, intrinsic motivation, and planning. Eventually, the agent lost the ability to contextualize action selection and became a random-walk object. This corresponds to an ϵ-greedy policy with $\epsilon = 1$. A random-walk agent is explorative but it cannot be considered curious

[1] https://github.com/arayabrain/MountainCar2D

in the sense that no principled means are applied to regulate explorative behaviors. With the model variants detailed below we intended to demonstrate the impact boredom and intrinsic motivation have on regulating exploration and, as a consequence, on forward model learning.

5.3.1. Boredom-Driven Curiosity (C/B)

The first agent variant retained all distinctive components introduced in Section 3. The meta-model provides the devaluation progress as intrinsic rewards, whilst the value function enables the agent to plan actions that are intrinsically rewarding in the long run.

5.3.2. Predictive Error-Driven Curiosity (C/PE)

The C/PE variant tests whether the induction of boredom is a constructive form of intrinsic motivation. This is achieved by removing the meta-model, thereby requiring an alternative definition of intrinsic reward. We replaced the devaluation progress with *learning progress* defined by mean squared errors of the forward model:

$$\begin{aligned} R_\theta^{(\ell+1)} &:= \mathcal{L}_{fm}(\theta^{(\ell)}) - \mathcal{L}_{fm}(\theta^{(\ell+1)}) \\ \mathcal{L}_{fm}(\theta) &:= \mathcal{L}(\boldsymbol{s}', \boldsymbol{a}, \boldsymbol{s}; \theta) \\ &:= \|\boldsymbol{s}' - f(\boldsymbol{a}, \boldsymbol{s}; \theta)\|^2 \end{aligned} \tag{17}$$

The construction of learning progress is one typical approach to intrinsic motivation and curiosity (Schmidhuber, 1991; Pathak et al., 2017).

5.3.3. Policy Gradients, Intrinsic Reward Samples (PG/IRS), Gaussian Rewards (PG/GR)

Next, we examined how reward statistics alone influences policy update and, as a consequence, model learning. The value function was removed at this stage to dissociate policy learning from any downstream effects of value learning.

One distinctive feature of devaluation progress is that it entails time-varying rewards — depending on the amount of time over which an agent has evolved in the environment. We hypothesized that the emergence of curious policy is associated with reward dynamics over time. That is to say, if one perturbs the magnitudes and directions of the policy gradients with reward statistics appropriate for the ongoing time frame, the agent should exhibit similar curious behaviors. Nevertheless, we argue that such treatment is only sensible given virtually identical initial conditions. Specifically, all agent variants shared the same, environmental configuration, initial position, and network initialization.

To this end, we prepared a database for intrinsic reward samples. During C/B performance, all reward samples were collected and labeled with the corresponding time step. Afterwards, the PG/IRS agents randomly sampled from the database in a temporally synchronized manner and applied standard policy gradients.

The PG/IRS was contrasted with the PG/GR variant. Their difference lies in that a surrogate reward was used in place of the database. We defined the surrogate reward as a Gaussian distribution with time-invariant parameters, in which the mean $\mu = 0$ is under the assumption of equilibrium devaluation progress and the standard deviation $\sigma = 0.01$, as derived from the entire database.

5.3.4. Random-Walk Policy (P/RW)

Finally, we constructed a random-walk agent. All network components, apart from the forward model, were removed. This agent variant represents the case without intrinsic motivation and is agnostic to curiosity. Broadly speaking, the agent was still explorative due to its maximum entropy action policy. We regarded this version as the worse case scenario to contrast with the rest of the variants.

5.4. Model Comparisons

All model variants were compared on the basis of validation error given the oracle dataset. We performed 128 runs for each of the six variants (Oracle, C/B, C/PE, PG/IRS, PG/GR, and P/RW). All variants, across all runs, were assigned to identical environmental configuration (e.g., initial position, attractor/repeller placements). Network components, whenever applicable, shared identical architecture and were trained with consistent batch size and learning rate. Model parameters followed the Xavier initialization (Glorot and Bengio, 2010). During post-DAP, learning rate scheduling was implemented such that a factor 0.1 reduction was applied upon a 3000-epoch loss plateau.

6. RESULTS

In this section, we offered qualitative and quantitative assessment of agent's behavioral pattern and performance across different agent variants. As established previously, an agent's performance in modeling its own environment necessarily depends on both explorative and non-perseverative behaviors. The overall picture being delivered here is that the boredom-driven curious agent (abbrev. C/B) exhibited stronger tendency toward exploration (**Figures 3A,B**) and against perseveration (**Figures 3C,D**). In accordance with our prediction, the forward model performance was significantly better for the boredom agent, as compared with other curious or non-curious variants (**Figure 4** and **Tables 2**, **3**).

We first characterized individual agent variants' qualities of being i) explorative and ii) perseverative. Active exploration is one defining attribute of curiosity (Gottlieb et al., 2013), simply because it differentiates between uncertain and known situations, thus giving rise to effective information acquisition. This, however, should be complemented with suppressed perseveration; namely, to prevent oneself from being permanently or dynamically captured—i.e., by the corners or the attractor.

The two qualities can be distinguished, as shown in **Figure 3**, by respective measures of Coverage Rate (CR) and Coverage Entropy (CE). The two measures were computed by first turning the state space into a 50×50 grid, ignoring velocities. CR keeps track of whether or not a grid cell has been visited and, at each time step, corresponds to the proportion of visited grid cells. A CR curve increasing over time indicates that an agent would be exploring new grid cells.

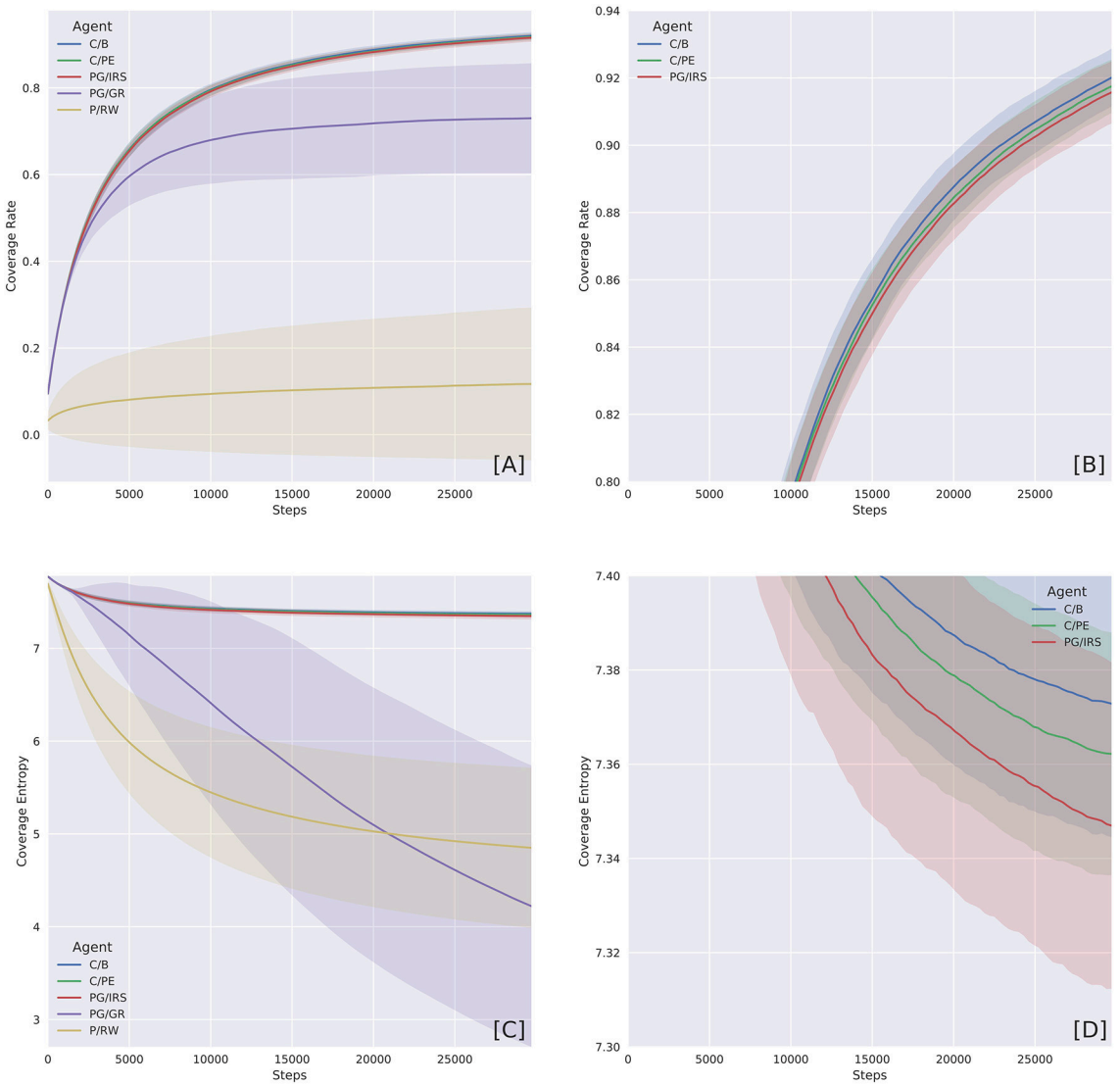

FIGURE 3 | Coverage Rate (CR) and Coverage Entropy (CE) by agent variants. The two measures were computed by first turning the state space into a 50 × 50 grid, ignoring velocities. CR then marks over time whether or not a cell has been visited. Whereas, CE treats the grid as a probability distribution. Starting with maximum entropy, CR cumulatively counts the number of times a position is being visited. Entropy was calculated at each time step using the normalized counter. **(A)** Overview of CR shows the distinction between curious and non-curious agents. Curiosity caused the agents to explore faster. **(B)** Close-up on the curious agent variants, which were equally explorative. **(C)** Overview of CE shows agents with different levels of perseverance. The P/RW variants were captured by the attractor, whilst the PG/GR variants were prone to blockage. **(D)** Close-up on curious agents, which were characterized by higher CE due to attractor avoidance and more frequent repeller visitation attempts. Shaded regions represent one standard deviation.

CE, on the other hand, accounts for the number of time steps an agent revisited one grid cell. This then gives an empirical probability distribution at each time step that reports the likelihood of finding an agent occupying a grid cell. A concentrated probability distribution means an agent only paid visit to a small set of grid cells and, as a result, the probability distribution has low entropy.

Because (state) visitation bias was inherent in our training environment, naturally, agents occupying a subset of states would cause CE to reduce faster than those who attempted to escape. The C/B, C/PE, and PG/IRS variants were regarded as curious and intrinsically motivated. Our results showed that these variants were predominantly explorative and non-perseverative. By contrast, the P/RW agent, albeit explorative, had no principled means to escape the potential well. However, if $t \rightarrow \infty$ the P/RW should be able to explore further by chance. The PG/GR variant, on the other hand, exhibited, intermediate exploratiVeness and extreme perseverance with disproportionately high variance. We attributed this behavior to the detrimental effects of inappropriately informative reward statistics.

Next, we benchmarked forward model performance of individual variants by their validation loss and error percentage. We reported DAP and post-DAP performances separately as a

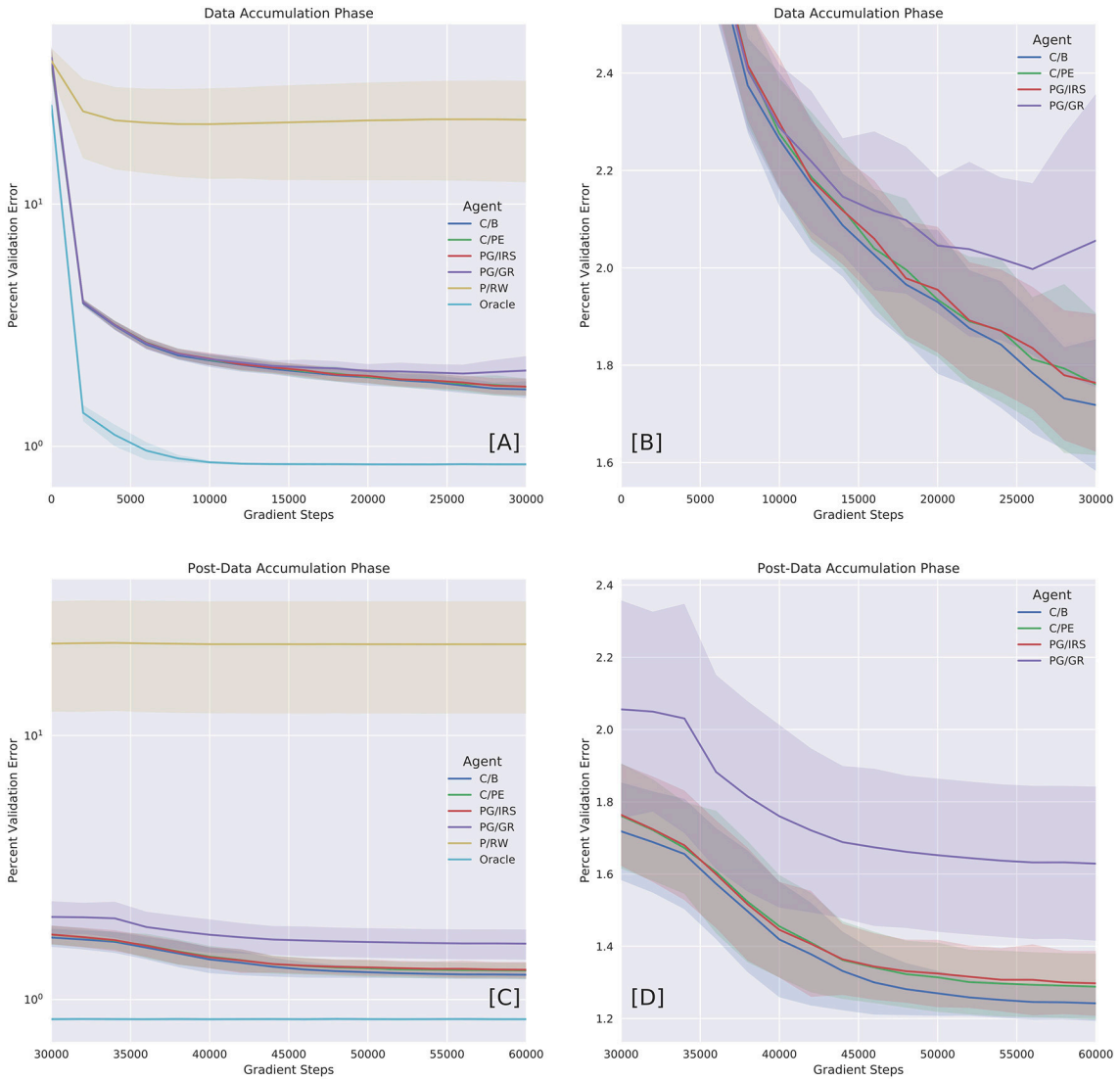

FIGURE 4 | Benchmarking model variants with oracle dataset. Performances were reported in error percentage **(**also, see **Table 2)**. **(A)** Performance as a function of time during Data Accumulation Phase (DAP). **(B)** Close-up on curious variants (C/B, C/PE, and PG/IRS), as well as policy gradients (PG/GR) informed by surrogate reward statistics. The C/PE and PG/IRS variants performed similarly, but differed significantly from C/B **(Table 2)**. **(C)** Performance over time during post-DAP. **(D)** Close-up on post-DAP performances for curious variants and PG/GR.

function of time in **Figure 4**. Error percentage was calculated as the percent ratio between root mean squared loss and the maximum pair-wise Euclidean distance in the validation set. This ratio can be summarized by $\|\boldsymbol{s}'_k - f(\boldsymbol{a}_k, \boldsymbol{s}_k; \theta)\| / \max_{i,j} \|\mathcal{D}_i - \mathcal{D}_j\|$, where $\mathcal{D}$ is the validation set and $(\boldsymbol{s}'_k, \boldsymbol{a}_k, \boldsymbol{s}_k) \in \mathcal{D}$.

The Oracle model, trained under the supervision of oracle training set, reached an error percentage of 0.84% for both DAP and post-DAP, amounting to approximately 30% improvement over the terminal performance of the C/B variant. All variants considered curious (C/B, C/PE, and PG/IRS) had similar performances during DAP. In particular, the PG/IRS, which received independent intervention from the 'true' reward distributions achieved marginally lower performance but indistinguishable from the C/PE variant. This outcome was observed for both DAP and post-DAP, suggesting intrinsic reward samples derived from C/B contributed favorably even to the standard policy gradients algorithm.

Though without the ability to approximate value function, the PG/IRS variant underperformed in benchmarking, as compared with the value-enabled, C/B variant. Using non-parametric test, the difference was detected for DAP ($p = 0.0006$) and post-DAP ($p = 6.4\text{E-}8$), respectively. Similar observations were also made for comparisons between C/B and C/PE, at $p = 0.0029$ (DAP) and $p = 5.9\text{E-}5$ (post-DAP). Overall, this suggested significant differences in the experiences accumulated across agent variants. The aforementioned statistics were reported in **Tables 2**, **3**.

TABLE 2 | Summary statistics on validation loss and error percentage as benchmarking scores.

Agent	DAP		Post-DAP	
	MSE loss (SD)	Mean Percent Error (SD)	MSE loss (SD)	Mean Percent Error (SD)
Oracle	0.0008 (2.3E-5)	0.8430 (0.0123)	0.0008 (2.2E-5)	0.8428 (0.0114)
C/B	0.0033 (0.0006)	1.7181 (0.1357)	0.0017 (0.0001)	1.2420 (0.0488)
C/PE	0.0035 (0.0006)	1.7611 (0.1464)	0.0019 (0.0003)	1.2882 (0.0916)
PG/IRS	0.0035 (0.0006)	1.7637 (0.1418)	0.0020 (0.0003)	1.2976 (0.0902)
PG/GR	0.0048 (0.0017)	2.0559 (0.3026)	0.0030 (0.0008)	1.6288 (0.2140)
P/RW	0.6663 (0.3904)	22.2734 (10.0085)	0.6615 (0.3864)	22.1453 (10.0775)

Apart from the Oracle model, a trend of declining scores can be observed as the agent degraded from C/B to P/RW, indicating the contribution of boredom and curiosity in model learning. Key: DAP, Data Accumulation Phase; SD, standard deviation. For agent codes, see ***Table 1****.*

TABLE 3 | Non-parametric statistical tests comparing terminal performance at DAP and post-DAP for curious model variants.

Mann-Whitney U-Test ($n = 128, \alpha = 0.025$, Bonferroni corrected)			
Validation loss		DAP ($T = 30,000$)	Post-DAP ($T = 60,000$)
C/B < C/PE	Statistics	6558.0	5911.0
	p-value	0.0029	5.9E-5
C/B < PG/IRS	Statistics	6275.0	5062.0
	p-value	0.0006	6.4E-8

Following ***Table 2****, even though the boredom score came close to other curious variants (C/PE and PG/IRS), the boredom variant still outperformed the other two on statistical grounds.*

7. LIMITATION

One obvious limitation of the proposed method is scalability. We imposed Gaussian assumption on the forward model and meta-model because this lends the KL-divergence between the two to have a closed form solution. However, this solution depends on both matrix inversion and log-determinant, whose computational complexity normally falls around an order of 3 when using Cholesky decomposition. To circumvent this limitation, the intrinsic reward (devaluation progress) may be replaced with one based on (forward model) prediction error at the expense of lesser curiosity.

The Gaussian assumption also puts limitations on the expressiveness of the models. This can be slightly relaxed to admit Gaussian mixture models. KL-divergence between Gaussian mixture models is not tractable but can nonetheless be approximated (e.g., Hershey and Olsen, 2007). Alternatively, employing normalizing flows (Rezende and Mohamed, 2015) also allows expressive models. Calculating KL-divergence in this case is typically resorted to Monte Carlo approximation. These are potential extensions that can be applied to the current work in the future.

8. CONCLUSION

We have provided a formal account on the emergence of boredom from an information-seeking perspective and addressed its constructive role in enabling curious behaviors. Boredom thus motivates an instrumental view of action selection, in which an action serves to disclose outcomes that have intrinsic meaningfulness to an agent itself. This is, a bored agent must seek out information worth assimilating into itself. This led to the central claim of this study—pertaining to the superior data-gathering efficiency and hence effective curiosity. We supported this claim with empirical evidence, showing that boredom-enabled agents consistently outperformed other curious agents in self-assisted forward model learning. Our results solicited the interpretation that the relationship between homeostatic and heterostatic intrinsic motivations can in fact be complementary; therefore, we have offered one unifying perspective for the intrinsic motivation landscape.

Our proposed method is general in formalization and sits comfortably with existing MDP problems. Our future work is then to apply the method to more complex problems, such as embedding into a robot for real-world scenarios.

AUTHOR CONTRIBUTIONS

YY conceived of this study, performed the experiments, and wrote the first draft of the manuscript. AC programmed the physics simulator, wrote part of Introduction, and created **Figure 1**. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This study was funded by the Japan Science and Technology Agency (JST) under CREST grant number JPMJCR15E2.

ACKNOWLEDGMENTS

YY would like to thank Martin Biehl and Ildefons Magrans de Abril for insightful discussions.

REFERENCES

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Q. J. Exp. Psychol.* 34, 77–98. doi: 10.1080/14640748208400878

Adams, C. D., and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. B* 33, 109–121. doi: 10.1080/14640748108400816

Balleine, B. W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates.

Neuropharmacology 37, 407–419. doi: 10.1016/S0028-3908(98)00033-1

Barmack, J. E. (1939). A definition of boredom: a reply to Mr. Berman. *Am. J. Psychol.* 52, 467–471.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). "Unifying count-based exploration and intrinsic motivation," in *Advances in Neural Information Processing Systems*, eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (New York, NY: Curran Associates, Inc.), 1471–1479. doi: 10.3390/bs3030459

Bench, S. W., and Lench, H. C. (2013). On the function of boredom. *Behav. Sci.* 3, 459–472.

Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity*. New York, NY: McGraw-Hill Book Company. doi: 10.1037/11164-000

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym. *arXiv [Preprint]. arXiv:1606.01540*.

Dahlen, E. R., Martin, R. C., Ragan, K., and Kuhlman, M. M. (2005). Driving anger, sensation seeking, impulsiveness, and boredom proneness in the prediction of unsafe driving. *Accid. Anal. Prev.* 37, 341–348. doi: 10.1016/j.aap.2004.10.006

Danckert, J., Hammerschmidt, T., Marty-Dugas, J., and Smilek, D. (2018). Boredom: Under-aroused and restless. *Consciousness Cogn.* 61, 24–37. doi: 10.1016/j.concog.2018.03.014

Eastwood, J. D., Frischen, A., Fenske, M. J., and Smilek, D. (2012). The unengaged mind: defining boredom in terms of attention. *Perspect. Psychol. Sci.* 7, 482–495. doi: 10.1177/1745691612456044

Elpidorou, A. (2014). The bright side of boredom. *Front. Psychol.* 5:1245. doi: 10.3389/fpsyg.2014.01245

Elpidorou, A. (2017). The bored mind is a guiding mind: toward a regulatory theory of boredom. *Phenomenol. Cogn. Sci.* 35, 17. doi: 10.1007/s11097-017-9515-1

Fahlman, S. A., Mercer, K. B., Gaskovski, P., Eastwood, A. E., and Eastwood, J. D. (2009). Does a lack of life meaning cause boredom? results from psychometric, longitudinal, and experimental analyses. *J. Soc. Clin. Psychol.* 28, 307–340. doi: 10.1521/jscp.2009.28.3.307

Fahlman, S. A., Mercer-Lynn, K. B., Flora, D. B., and Eastwood, J. D. (2013). Development and validation of the multidimensional state boredom scale. *Assessment* 20, 68–85. doi: 10.1177/1073191111421303

Fairbank, M., and Alonso, E. (2012). "Value-gradient learning," in *Neural Networks (IJCNN), The 2012 International Joint Conference on IEEE* (Brisbane, QLD), 1–8.

Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Front. Psychol.* 3:130. doi: 10.3389/fpsyg.2012.00130

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Geiwitz, P. J. (1966). Structure of boredom. *Jo. Personal. Soc. Psychol.* 3, 592. doi: 10.1037/h0023202

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Sardinia), 249–256.

Gomez-Ramirez, J., and Costa, T. (2017). Boredom begets creativity: a solution to the exploitation exploration trade-off in predictive coding. *BioSystems* 162, 168–176. doi: 10.1016/j.biosystems.2017.04.006

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends Cogn. Sci.* 17, 585–593. doi: 10.1016/j.tics.2013.09.001

Harris, M. B. (2000). Correlates and characteristics of boredom proneness and boredom. *J. Appl. Soc. Psychol.* 30, 576–598. doi: 10.1111/j.1559-1816.2000.tb02497.x

Heess, N., Wayne, G., Silver, D., Lillicrap, T., Tassa, Y., and Erez, T. (2015). Learning continuous control policies by stochastic value gradients. *arXiv [Preprint]. arXiv:1510.09142*.

Hershey, J. R., and Olsen, P. A. (2007). "Approximating the kullback leibler divergence between gaussian mixture models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on IEEE*, Vol. 4 (Honolulu, HI), IV–317.

Karl, M., Soelch, M., Bayer, J., and van der Smagt, P. (2016). Deep variational bayes filters: unsupervised learning of state space models from raw data. *arXiv [Preprint]. arXiv:1605.06432*.

Kass, S. J., and Vodanovich, S. J. (1990). Boredom proneness: its relationship to type a behavior pattern and sensation seeking. *Psychology* 27, 7–16.

Kurzban, R., Duckworth, A., Kable, J. W., and Myers, J. (2013). Cost-benefit models as the next, best option for understanding subjective effort. *Behav. Brain Sci.* 36, 707–726. doi: 10.1017/S0140525X13001532

Larson, R. W. (1990). Emotions and the creative process; anxiety, boredom, and enjoyment as predictors of creative writing. *Imagination Cogn. Personal.* 9, 275–292. doi: 10.2190/XT9G-WXRF-BK4M-36AK

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2015). Continuous control with deep reinforcement learning. *arXiv [Preprint]. arXiv:1509.02971*.

London, H., Schubert, D. S., and Washburn, D. (1972). Increase of autonomic arousal by boredom. *J. Abnorm. Psychol.* 80, 29. doi: 10.1037/h0033311

Mannella, F., Mirolli, M., and Baldassarre, G. (2016). Goal-directed behavior and instrumental devaluation: a neural system-level computational model. *Front. Behav. Neurosci.* 10:181. doi: 10.3389/fnbeh.2016.00181

Merrifield, C., and Danckert, J. (2014). Characterizing the psychophysiological signature of boredom. *Exp. Brain Res.* 232, 481–491. doi: 10.1007/s00221-013-3755-2

Mikulas, W. L., and Vodanovich, S. J. (1993). The essence of boredom. *Psychol. Rec.* 43, 3.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., et al. (2013). Playing atari with deep reinforcement learning. *arXiv [Preprint]. arXiv:1312.5602*.

Mugon, J., Struk, A., and Danckert, J. (2018). A failure to launch: regulatory modes and boredom proneness. *Front. Psychol.* 9:1126. doi: 10.3389/fpsyg.2018.01126

Ostrovski, G., Bellemare, M. G., Oord, A. v. d., and Munos, R. (2017). Count-based exploration with neural density models. *arXiv [Preprint]. arXiv:1703.01310*.

Oudeyer, P.-Y., and Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Front. Neurorob.* 1:6. doi: 10.3389/neuro.12.006.2007

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). "Curiosity-driven exploration by self-supervised prediction," in *ICML* (Sydney, NSW).

Pattyn, N., Neyt, X., Henderickx, D., and Soetens, E. (2008). Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? *Physiol. Behav.* 93, 369–378. doi: 10.1016/j.physbeh.2007.09.016

Perkins, R. E., and Hill, A. (1985). Cognitive and affective aspects of boredom. *Br. J. Psychol.* 76, 221–234. doi: 10.1111/j.2044-8295.1985.tb01946.x

Rezende, D. J., and Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv [Preprint]. arXiv:1505.05770*.

Schmidhuber, J. (1991). "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats* (Paris), 222–227.

Schmidhuber, J. (2008). "Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes," in *Workshop on Anticipatory Behavior in Adaptive Learning Systems* (Munich: Springer), 48–76.

Schubert, D. S. (1977). Boredom as an antagonist of creativity. *J. Creat. Behav.* 11, 233–240. doi: 10.1002/j.2162-6057.1977.tb00631.x

Schubert, D. S. (1978). Creativity and coping with boredom. *Psychiatr. Ann.* 8, 46–54.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*, Vol. 1. Cambridge, MA: MIT Press.

Tomczak, J. M., and Welling, M. (2016). Improving variational auto-encoders using householder flow. *arXiv [Preprint]. arXiv:1611.09630*.

van Tilburg, W. A., and Igou, E. R. (2012). On boredom: Lack of challenge and meaning as distinct boredom experiences. *Motiv. Emotion* 36, 181–194. doi: 10.1007/s11031-011-9234-9

Vodanovich, S. J., and Kass, S. J. (1990). A factor analytic study of the boredom proneness scale. *J. Personal. Asses.* 55, 115–123. doi: 10.1080/00223891.1990.9674051

Vodanovich, S. J., Verner, K. M., and Gilbride, T. V. (1991). Boredom proneness: Its relationship to positive and negative affect. *Psychol. Rep.* 69, 1139–1146. doi: 10.2466/pr0.1991.69.3f.1139

Vogel-Walcutt, J. J., Fiorella, L., Carper, T., and Schatz, S. (2012). The definition, assessment, and mitigation of state boredom within educational settings: a comprehensive review. *Educ. Psychol. Rev.* 24, 89–111. doi: 10.1007/s10648-011-9182-7

Watter, M., Springenberg, J. T., Boedecker, J., and Riedmiller, M. (2015). Embed to control: a locally linear latent dynamics model for control from raw images. *arXiv [Preprint]. arXiv:1506.07365*.

Zuckerman, M. (1971). Dimensions of sensation seeking. *J. Consult. Clin. Psychol.* 36, 45–52. doi: 10.1037/h0030478

Zuckerman, M. (2008). "Sensation seeking," in *The International Encyclopedia of Communication, 1st Edn*, ed W. Donsbach (John Wiley & Sons, Ltd), 1–3. doi: 10.1002/9781405186407.wbiecs029

Zuckerman, M., Eysenck, S. B., and Eysenck, H. J. (1978). Sensation seeking in england and america: cross-cultural, age, and sex comparisons. *J. Consult. Clin. Psychol.* 46, 139. doi: 10.1037/0022-006X.46.1.139

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.