The Optimal Explorer Hypothesis and Its Formulation as a Combinatorial Optimization Problem

Mikel Malagón *
University of the Basque Country UPV/EHU
Donostia-San Sebastián
mikel.malagon@ehu.eus

Jon Vadillo
University of the Basque Country UPV/EHU
Donostia-San Sebastián
jon.vadillo@ehu.eus

Josu Ceberio
University of the Basque Country UPV/EHU
Donostia-San Sebastián
josu.ceberio@ehu.eus

Jose A. Lozano
Basque Center for Applied Mathematics
University of the Basque Country UPV/EHU
Donostia-San Sebastián
ja.lozano@ehu.eus

Abstract

This research project explores the hypothesis that, given a bounded number of steps in an environment, agents that most efficiently optimize their model of the environment are more likely to induce emergent intelligent behavior in a reward-free scenario. We refer to this as the optimal explorer hypothesis. The project aims to formalize and analyze this hypothesis, investigating its theoretical implications and connections to related areas such as open-ended learning and active inference. Building on this foundation, we will develop a practical implementation of an approximate "optimal explorer" agent by formulating it as a combinatorial optimization problem and leveraging established methods from the field. Finally, we will conduct extensive experiments to evaluate whether the proposed agent induces emergent behaviors in diverse and challenging environments.

1 Introduction

Agents—understood as systems acting by themselves according to certain goals or norms in an environment [Barandiaran et al., 2009]—are the substrates of intelligent life as we know it. However, not all agents can be classified as intelligent. For instance, a thermostat fits into most definitions of agency, while not exhibiting intelligence as found in biological agents. This raises the question: *what are intelligent agents doing?* In other words, what are the objectives that intelligent agents pursue that give rise to the incredibly complex emergent behaviors that we broadly observe in natural life?

On the Artificial Intelligence (AI) field side, Reinforcement Learning (RL) has been the area of research that has most prominently focused on the subject of intelligent agents [Kaelbling et al., 1996, Sutton and Barto, 2018]. Although RL has led to many breakthroughs in the last decades [Silver et al., 2016, Abramson et al., 2024], most RL literature has focused on developing agents to pursue a single, well-defined objective. In fact, Sutton's *reward hypothesis* states that all goals can be framed as cumulative reward maximization [Sutton, 2004, Bowling et al., 2023]. This led Silver et al. [2021] to hypothesize that optimizing reward can lead to the emergence of general intelligence and complex behavior in a sufficiently rich environment Silver et al. [2021].

On the other hand, as discussed by Stanley and Lehman [2015] and Soros et al. [2017], following an explicit objective can lead to dead ends. These works challenge the effectiveness of explicit

*Corresponding author.

XVI XVI Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (maeb 2025).

objectives, arguing that direct goal optimization often fails to discover necessary stepping stones. It emphasizes open-ended exploration over direct optimization, suggesting that breakthroughs arise from serendipity and novelty search rather than predefined goals [Lehman and Stanley, 2011, Kumar et al., 2024]. The emergence of intelligent behavior via open-ended novelty search has inspired a growing number of works in recent years [Bauer et al., 2023, Bruce et al., 2024, Matthews et al., 2025], even characterizing it as essential for superhuman-level intelligence [Hughes et al., 2024]. Many of these works focus on automating an open-ended environment (task) generation process [Bruce et al., 2024, Faldor et al., 2025] and learn a robust policy that will generalize to unseen tasks, known as Unsupervised Environment Design (UED) [Parker-Holder et al., 2022, Bauer et al., 2023, Rigter et al., 2024, Beukman et al., 2024]. However, most open-ended literature assumes that the learning method has access to and control of the environment to generate vast amounts of tasks (e.g., UED) or high-level control (e.g., Voyager by Wang et al. [2024]).

Instead, in natural life, agents interact with a (single) complex environment only through perception and (low-level) action. Since Helmholtz [1867], most prominent theories of cognition of today agree that the brain maintains and updates a model of its environment (i.e., the real world) [Doya, 2002, Friston, 2009]. Based on these ideas and recent work on lifelong learning and open-endedness theory [Abel et al., 2023, Hughes et al., 2024] this work hypothesizes that (informally):

> *The agents that most efficiently learn an internal model of the environment are more likely to produce emergent intelligent behavior in reward-free scenarios over a bounded time scope.*

In this context, the agent's model of the environment—referred to as the *world model*—is trained on agent-generated trajectories (i.e., sequences of interactions with the environment). Efficiency is measured as the expected sum across timesteps of the world model's prediction error with respect to the environment over all the possible trajectories. We refer to this as the *optimal explorer hypothesis*. Note that this hypothesis does not state that the agents that most efficiently learn their world model are the only or the most likely ones to induce emergent behaviors, just that they are more likely to cause them by doing so.

In the search for emergent behavior, this hypothesis directly introduces the intrinsic objective of acting to generate the most informative trajectories for the world model in the long run. Based on this hypothesis and literature on active inference [Friston, 2009] and model-based RL [Chua et al., 2018], the next part of this project will propose a practical implementation of an agent to optimize this long-term intrinsic objective. Equipped with a deep Neural Network (NN) ensemble-based world model [Lakshminarayanan et al., 2017], we aim to introduce an agent that plans and selects the sequences of actions that maximize the world model's epistemic uncertainty, in the long run, using the Cross-Entropy Method (CEM) [Rubinstein, 1999]. This way, the action selection policy and the world model (constantly updated with the trajectories sampled by the latter) play a minimax game that explores in face of the unknown while otherwise exploiting to explore. Finally, we will conduct an extensive empirical evaluation on challenging environments to analyze the behavior of the proposed agent. We expect the agent to solve complex games (in episodic setups) even without having a reward signal, and to improve the sample efficiency of reward-based RL methods model-based (e.g., DreamerV3 [Hafner et al., 2023]) and non-model-based methods (e.g., proximal policy optimization [Schulman et al., 2017]).

In summary, the main objectives of this project are the following:

1. **Formalization of the optimal explorer hypothesis.** Define and analyze the hypothesis, establishing its theoretical foundations and connections to related research areas such as open-ended learning and active inference.

2. **Combinatorial optimization formulation.** Frame the problem of optimal exploration as a Combinatorial Optimization (CO) task, identifying suitable problem representations and constraints.

3. **Algorithm development.** Design and implement an approximate optimal explorer agent by leveraging techniques from model-based RL and combinatorial optimization, such as Estimation of Distribution Algorithms [Larrañaga and Lozano, 2002] (employed in the CEM).

4. **Empirical evaluation.** Conduct experiments in diverse and challenging environments to assess the effectiveness of the proposed agent in inducing emergent behaviors in reward-free scenarios.

2 Previous work

The following lines provide a brief overview of the fields and work upon which this work is mainly based.

Lifelong and open-ended learning. Lifelong and open-ended learning focus on agents that continuously acquire and refine knowledge over time, adapting to novel scenarios by leveraging past experiences. However, learning continuously introduces many challenges as catastrophic forgetting and interference, loss of plasticity, or computational cost [Hadsell et al., 2020]. Addressing these issues is an active area of research [Khetarpal et al., 2022, Wolczyk et al., 2024, Malagon et al., 2024]. Other works depart from sequential tasks and focus on meta-learning a robust policy on a distribution of environments [Parker-Holder et al., 2022, Beukman et al., 2024]. Although these deeply connected fields have gained increasing attention in recent years, they are still in the phase of formally defining themselves [Abel et al., 2023, Hughes et al., 2024].

Exploration strategies. Although many goals can be framed as a reward maximization problem [Sutton, 2004], learning a policy can be extremely difficult in the absence of a dense informative reward signal. Thus, the field of RL has come with a vast body of work on intrinsic reward: an auxiliary reward function to guide exploration to promising trajectories [Pathak et al., 2017, Burda et al., 2019, Nikulin et al., 2023]. Even with intrinsic motivation, RL agents greatly suffer from sample efficiency. In this realm, model-based methods learn (or directly employ when available) a model of the environment which is used to plan the actions [Kaiser et al., 2020, Hafner et al., 2023]. However, model-based RL incorporates additional complexity and agents can exploit biases in the model that lead to substantial degradation of performance in these types of methods [Janner et al., 2019].

Active inference. Active inference is based on Friston's Free Energy Principle (FEP) [Friston, 2009]. According to the FEP, living beings minimize expected free energy, maximizing the probability of being in desirable states (maintaining homoeostatic equilibrium) while maximizing information gain (minimizing epistemic uncertainty) in the long run [Friston et al., 2015]. Despite the appeal of biologically plausible active inference agents [Friston, 2010] and recent efforts to incorporate deep neural networks [Fountas et al., 2020], scaling beyond toy environments remains a challenge for these methods [Sajid et al., 2021].

3 The optimal explorer hypotheis

As described in the introduction, we focus on agents that maintain and update an internal model (i.e., world model) of their environment.[2] Moreover, the environment is only composed of a transition function and without a reward function (i.e., reward-free environments). Every timestep the agent interacts with the environment by generating a new transition, and the world model is updated accordingly.

In this setup, we hypothesize that the agents that generate the trajectories (sequences of interactions) that most efficiently update their world model are more likely to induce emergent intelligent behavior in a finite scope of time. In this context, we define the efficiency of an agent as the expected sum of the global world model error at each timestep by following the agent's policy.[3] In turn, we refer to global error as the world model's error modeling of the environment given all the possible trajectories. Thus, if the global error is zero, the world model and the environment define the same probability distribution.

[2]The world model can be naturally defined as the distribution over all the possible states given the current state and action, $p_\phi(s_{t+1}|s_t, a_t)$.

[3]We refer to an agent's policy in the classic RL sense, that is, the probability distribution over actions given the current state $p(a|s)$.

Intuitively, those agents that most efficiently optimize their world models will be those that find (often by exploiting *shorcuts* in complex environments) the best trajectories to explore their environments. Note that this substantially differs from random exploration (e.g., ϵ-greedy exploration), as the most efficient agents will be those that exploit to explore. For instance, in an episodic environment such as an Atari game [Machado et al., 2018] (and most games) an efficiently exploring agent (in terms of our hypothesis) would have to solve the game as fast as possible to update its model with interactions from advanced stages of the game.

4 Proposing a practical implementation based on CO

In this part of the project, we aim to explore the implications of the hypothesis proposed in the previous section. Specifically, we leverage the ideas from the optimal explorer hypothesis to propose an agent that efficiently explores its environment in the absence of a reward function (i.e., without explicit objectives). Note that many possible implementations of such an agent exist and that the one from this part of the project is just a proposal to analyze the experimental implications of the hypothesis.

Specifically, we aim to leverage previous work on uncertainty quantification for deep NN models [Gal and Ghahramani, 2016, Lakshminarayanan et al., 2017] to select those trajectories that are more informative to the world model (a deep NN). Finding the action that will cause the most efficient world model update (in terms of the hypothesis) at each timestep can be framed as searching for the action sequence that will lead the agent toward the most informative interactions—of highest epistemic uncertainty—for the world model.

Note that this problem can be framed as a **combinatorial optimization** problem where the space of **possible solutions** Ω are the sequences of actions, $a \in \mathcal{A}$, $\mathbf{x} = (a_1, a_2, \ldots, a_n)$ of a given length n (that corresponds to the planning horizon).[4] Accordingly, the **objective function** $f(x)$ is the cumulative epistemic uncertainty of the world model at each interaction, where the interactions are autoregressively sampled from the world model itself. Formally, considering the world model a probability distribution parametrized by ϕ over the next states of the environment conditioned on the current action and state, the objective function f can be written as,

$$f(\mathbf{x}, i, s) = EU(p_\phi(\cdot|s, \mathbf{x}_i)) + \mathbb{E}_{s' \sim p_\phi(\cdot|s,\mathbf{x}_i)}[f(\mathbf{x}, i+1, s')]. \tag{1}$$

Where $EU(p_\phi(\cdot|s, \mathbf{x}_i))$ is the epistemic uncertainty of the state s and action $\mathbf{x}$ in the world model p_ϕ. Note that the fitness of a solution $\mathbf{x}$ is always given with respect to a specific state s, as the utility of a given action sequence is completely dependent on the initial state in which it is taken. Thus, our agent proposal, being in a state s, would select the action a such that,

$$a = \arg\max_{a_1 \in \mathcal{A}} f((a_1, \ldots), 1, s). \tag{2}$$

Intuitively, at each state, the agent would choose the action that maximizes the expected long-term epistemic uncertainty of its world model.

5 Conclusion

This project outlines a novel approach to emergent intelligent behavior in the absence of explicit objectives (i.e., without reward function). We first (informally) introduce the optimal explorer hypothesis, which connects the efficiency of learning a model of the environment and the likelihood of inducing emergent behaviors. From this hypothesis, we structure the project into four objectives: (1) formalizing the hypothesis, (2) formulating it as a combinatorial optimization problem, (3) developing an agent based on the combinatorial optimization problem formulation, and (4) extensive empirical analysis of the agent. The project aims to advance our comprehension of exploration strategies and learning dynamics in artificial agents, paving the way for more adaptable and intelligent systems and their formal understanding.

[4] Where $\mathcal{A}$ is finite and its elements discrete.

References

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado P van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *Proceedings of the 2023 Advances in Neural Information Processing Systems (NeurIPS)*, 36:50377–50407, 2023.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pages 1–3, 2024.

Xabier E Barandiaran, Ezequiel Di Paolo, and Marieke Rohde. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386, 2009.

Jakob Bauer, Kate Baumli, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, et al. Human-timescale adaptation in an open-ended task space. In *Proceedings of the 2023 International Conference on Machine Learning (ICML)*, pages 1887–1935, 2023.

Michael Beukman, Samuel Coward, Michael Matthews, Mattie Fellows, Minqi Jiang, Michael D Dennis, and Jakob Nicolaus Foerster. Refining minimax regret for unsupervised environment design. In *Proceedings of the 2024 International Conference on Machine Learning (ICML)*, 2024.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *Proceedings of the 2023 International Conference on Machine Learning (ICML)*, pages 3003–3020, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Proceedings of the 2024 International Conference on Machine Learning (ICML)*, 2024.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Proceedings of the 2018 Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Kenji Doya. *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press, 2002.

Maxence Faldor, Jenny Zhang, Antoine Cully, and Jeff Clune. OMNI-EPIC: Open-endedness via models of human notions of interestingness with environments programmed in code. In *Proceedings of the 2025 International Conference on Learning Representations (ICLR)*, 2025.

Zafeirios Fountas, Noor Sajid, Pedro Mediano, and Karl Friston. Deep active inference agents using monte-carlo methods. *Proceedings of the 2020 Advances in Neural Information Processing Systems*, 33:11662–11675, 2020.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13 (7):293–301, 2009.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138, 2010.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 2016 International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hermann v. Helmholtz. *Handbuch der physiologischen Optik*. L. Voss, 1867.

Edward Hughes, Michael D Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktäschel. Position: Open-endedness is essential for artificial superhuman intelligence. In *Proceedings of the 2024 International Conference on Machine Learning (ICML)*, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Proceedings of the 2019 Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for atari. In *Proceedings of the 2020 International Conference on Learning Representations (ICLR)*, 2020.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research (JAIR)*, 75: 1401–1476, 2022.

Akarsh Kumar, Chris Lu, Louis Kirsch, Yujin Tang, Kenneth O Stanley, Phillip Isola, and David Ha. Automating the search for artificial life with foundation models. *arXiv preprint arXiv:2412.17799*, 2024.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *proceedings of the 2017 Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Pedro Larrañaga and Jose A Lozano. *Estimation of distribution algorithms: A new tool for evolutionary computation*, volume 2. Springer Science & Business Media, 2002.

Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 2011.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research (JAIR)*, 61:523–562, 2018.

Mikel Malagon, Josu Ceberio, and Jose A Lozano. Self-composing policies for scalable continual reinforcement learning. In *Proceedings of the 2024 International Conference on Machine Learning (ICML)*, 2024.

Michael Matthews, Michael Beukman, Chris Lu, and Jakob Nicolaus Foerster. Kinetix: Investigating the training of general agents through open-ended physics-based control tasks. In *Proceedings of the 2025 International Conference on Learning Representations (ICLR)*, 2025.

Alexander Nikulin, Vladislav Kurenkov, Denis Tarasov, and Sergey Kolesnikov. Anti-exploration by random network distillation. In *Proceedings of the 2023 International Conference on Machine Learning (ICML)*, pages 26228–26244, 2023.

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *Proceedings of the 2022 International Conference on Machine Learning (ICML)*, pages 17473–17498, 2022.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 2017 International Conference on Machine Learning (ICML)*, 2017.

Marc Rigter, Minqi Jiang, and Ingmar Posner. Reward-free curricula for training robust world models. In *Proceedings of the 2024 International Conference on Learning Representations (ICLR)*, 2024.

Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1:127–190, 1999.

Noor Sajid, Philip J Ball, Thomas Parr, and Karl J Friston. Active inference: demystified and compared. *Neural Computation*, 33(3):674–712, 2021.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

Lisa B Soros, Joel Lehman, and Kenneth O Stanley. Open-endedness: The last grand challenge you've never heard of, 2017. URL `https://www.oreilly.com/radar/open-endedness-the-last-grand-challenge-youve-never-heard-of/`.

Kenneth O Stanley and Joel Lehman. *Why greatness cannot be planned: The myth of the objective*. Springer, 2015.

Richard Sutton. The reward hypothesis. `http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html`, 2004. Accessed: 2025-02-28.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.

Maciej Wolczyk, Bartłomiej Cupiał, Mateusz Ostaszewski, Michał Bortkiewicz, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Fine-tuning reinforcement learning models is secretly a forgetting mitigation problem. In *Proceedings of the 2024 International Conference on Machine Learning (ICML)*, 2024.