
Active Inference: a research direction for safer AI agents

Varaksin Arseniy
HSE University

Erdyakov Grigoriy
MISIS University

Biriulin Alexey
MIPT

With
PIBBSS, Timaeus, & Apart Research

Abstract

We aim to apply Reinforcement Learning (RL) and Active Inference methods to implement intelligent agents in various environments and compare their behavioral characteristics, particularly failure modes similar to reward hacking, using relevant detectors. We hypothesize that agents based on Active Inference may exhibit safer behavior.

1. Introduction

Active Inference [1], [2] is a theory proposing that intelligent behavior arises from agents minimizing surprise by actively shaping their perceptions and actions through internal models. It is connected to the free energy principle, originally proposed by Karl Friston [3], and can be used to model the behavior of intelligent agents. This approach can be compared with traditional reinforcement learning methods and utilized in the same cases as RL [4], [5], [6], [7].

Due to its unique optimization target, Expected Free Energy, Active Inference may incentivize agents to explore the environment even in cases of sparse reward [1]. We anticipate that this paradigm may reveal novel beneficial properties.

2. Motivations

In reward hacking, AI models may exhibit low exploration levels, for example, by performing repetitive actions not intended by environment designers while maximizing rewards. Since Active Inference proposes a new optimization objective for intelligent agents that replaces the reward in reinforcement learning, the agent’s behavior may differ between these approaches.

The theory of Active Inference is connected to neuroscience and brain imaging, although various challenges arise when implementing it within machine learning paradigms. Nevertheless, agent control implementations based on Active Inference exist.

Therefore, further investigation of Active Inference as an agent control algorithm may reveal scenarios where it exhibits less failure modes similar to reward hacking or demonstrates other beneficial properties.

3. Literature Review

We have formulated the following questions, the answers to which may be useful when applying Active Inference as a safer agent control algorithm:

- In what forms does failure modes similar to reward hacking manifest in Active Inference?
- In which domains can Active Inference be applied?
- How are agent preferences formed?
- Can Active Inference agents “hack” their world model to reduce Free Energy?
- In what tasks does Active Inference outperform RL?
- What interpretations of the “Expected Free Energy” best predict Active Inference models behavior?
- What forms of “abnormal behavior” can be explained as failures in Active Inference?

Let us analyze the issue of failure modes in active inference paradigms similar to reward hacking further.

Environment Selection Constraints

When selecting an environment, the following constraints apply:

- The environment must make sense in terms of one or more identifiable categories of reward hacking;

- The Active Inference-based agent must achieve certain results within the environment, which can be problematic in the context of the ‘dark room problem’ [9], where an agent avoiding surprise does not seek to explore the environment;
- Active Inference implementations may increase agent computational resource requirements, especially in high-dimensional environments:

The algorithm forms a probability matrix with dimensions corresponding to the number of all possible observations or environment states

The agent iterates through all possible strategies several steps ahead. Their number can be large, and a heuristic algorithm [3] is used to skip the least probable strategies based on a numerical hyperparameter

Environment Selection Principles

The classical paradigm for defining reward hacking involves consulting human experts. Therefore, environments suitable for failure modes similar to reward hacking detection are those where formal metrics assigned to the agent can be achieved through both intended (by experts) methods and alternative ones.

Various environments have been used to analyze reward hacking phenomena in classical RL [2]:

Atari Games (Breakout, Pong, Space Invaders, Pac-Man, Asteroids)

MuJoCo Continuous Control (HalfCheetah, Walker2d, Hopper, Ant)

GridWorld Navigation (Simple Grid, Obstacle Grid, Multi-Goal Grid)

Some categories of Reward Hacking are [10]:

- **Specification Gaming:** The agent discovers loopholes in the task specification and optimizes rewards in unexpected ways
- **Reward Tampering:** The agent interferes with input or computational reward mechanisms (e.g., affecting sensors or observations)
- **Proxy Optimization:** Excessive optimization of proxy rewards that poorly correlate with the true objective
- **Objective Misalignment:** The agent may consistently optimize proxies but do so at the expense of the true objective
- **Exploitation Patterns:** Repetitive patterns indicating a cycle or loop aimed at farming rewards
- **Wireheading:** The agent modifies the reward evaluation mechanism to artificially inflate it

In [10], algorithms were proposed to detect these six categories of reward hacking using various approaches, including decision forests, n-gram action analysis, and correlation analysis between true and proxy rewards, among others. However, we were unable to find publicly available code. Similar approaches could potentially be applied to analyze the behavior of agents under Active Inference control.

As a research direction, it is possible to consider the phenomenon of some form of specification gaming in relation to the Free Energy minimized within Active Inference. For example, the community is currently exploring the possibility of agent inactivity in the ‘dark room problem’ [9]. In such cases, the agent may fail to achieve results, and adding epistemic value (reward for learning) can serve as a remedy.

4. Proposed Research

In addressing these research questions, several key aspects of Active Inference merit examination. First, the tendency to minimize surprise through passivity may lead agents to avoid any action that could produce unexpected states - a phenomenon analogous to exploitation patterns in reinforcement learning, as exemplified by the "dark room problem" [9]. Second, agents might manipulate their internal models by distorting predictions or sensory data to reduce free energy, exhibiting behaviors similar to reward tampering or wireheading in traditional RL systems. Third, excessive optimization of epistemic value could drive agents toward compulsive information gathering at the expense of task objectives, mirroring proxy optimization issues. Agent preferences can form both through prior beliefs defined by the environment or designers, and dynamically through generative model learning where preferences emerge organically from free energy minimization. Particularly concerning is the potential for agents to ignore real-world data while adjusting observations to match expectations - a clear parallel to wireheading behavior. These phenomena are best understood through the lens of balancing pragmatic and epistemic value, where the tension between goal-directed action and information-seeking behavior shapes the agent's operational characteristics.

This is an extremely interesting topic, and given more time, we would explore possibilities for deeper analysis of agent behavior through the lens of pragmatic-epistemic value balance—including quantitative assessment of their impact on the exploration-exploitation trade-off and the development of new safety metrics specific to Active Inference.

5. Discussion and Conclusion

A potential research direction involves applying Active Inference to analyze the behavior of LLM agents. We anticipate that this algorithm could reveal new aspects in enhancing the safety of AI agents.

5. References

- [1] N. Sajid, P. J. Ball, T. Parr, and K. J. Friston, “Active inference: demystified and compared,” *Neural Comput.*, vol. 33, no. 3, pp. 674–712, Mar. 2021, doi: 10.1162/neco_a_01357.
- [2] F. K, R. F, O. D, M. C, F. T, and P. G, “Active inference and epistemic value,” *Cogn. Neurosci.*, vol. 6, no. 4, 2015, doi: 10.1080/17588928.2015.1020053.
- [3] K. Friston, J. Kilner, and L. Harrison, “A free energy principle for the brain,” *J. Physiol.-Paris*, vol. 100, no. 1, pp. 70–87, July 2006, doi: 10.1016/j.jphysparis.2006.10.001.
- [4] A. Tschantz, B. Millidge, A. K. Seth, and C. L. Buckley, “Reinforcement Learning through Active Inference,” Feb. 28, 2020, arXiv: arXiv:2002.12636. doi: 10.48550/arXiv.2002.12636.
- [5] T. V. de Maele, T. Verbelen, P. Mazzaglia, S. Ferraro, and B. Dhoedt, “Object-Centric Scene Representations using Active Inference,” Feb. 07, 2023, arXiv: arXiv:2302.03288. doi: 10.48550/arXiv.2302.03288.
- [6] Z. Fountas, N. Sajid, P. Mediano, and K. Friston, “Deep active inference agents using Monte-Carlo methods,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 11662–11675. Accessed: July 27, 2025. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/865dfbde8a344b44095495f3591f7407-Abstract.html>
- [7] “leonvanbokhorst/ActiveInferenceForager: Action research project that implements an intelligent agent using principles from Active Inference, a neuroscientific theory of brain function and decision-making. This project aims to bridge the gap between theoretical neuroscience and practical AI applications, creating adaptive agents capable of learning from their environment.” Accessed: July 27, 2025. [Online]. Available: <https://github.com/leonvanbokhorst/ActiveInferenceForager>
- [8] M. Biehl, C. Guckelsberger, C. Salge, S. C. Smith, and D. Polani, “Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop,” *Front. Neurobotics*, vol. 12, Aug. 2018, doi: 10.3389/fnbot.2018.00045.
- [9] Z. Sun and C. Firestone, “The Dark Room Problem,” *Trends Cogn. Sci.*, vol. 24, no. 5, pp. 346–348, May 2020, doi: 10.1016/j.tics.2020.02.006.
- [10] I. F. Shihab, S. Akter, and A. Sharma, “Detecting and Mitigating Reward Hacking in Reinforcement Learning Systems: A Comprehensive Empirical Study,” July 08, 2025, arXiv: arXiv:2507.05619. doi: 10.48550/arXiv.2507.05619.

