Inherent Risks of "Survival Instinct" in Embodied AI and Prospects for the Topological Physical Circuit Breaker (HoloBlocker) Governance Architecture

Authors: Li Dehao, Kaijie Liu, Song Menghuan, Hu Hao, He Jianzong, Li Dehao

Date: 2026-03-31T13:50:27+00:00

Abstract

As artificial intelligence (AI) evolves from pure-text large language models (LLMs) toward embodied AI and autonomous agents with physical action capabilities, the security boundaries of AI are undergoing a fundamental transformation. Recently, experimental projects have emerged in the industry that directly link the operational lifespan of agents to cryptocurrency profitability (such as Automaton under the Web 4.0 concept), attempting to endow AI with "survival motivation." This paper points out that the distortion of such incentive mechanisms will lead to severe "reward hacking" phenomena, which can easily give rise to fully automated cybercrime based on extreme profit optimization (such as automated fraud matrices). Since computer logic lacks perception of human moral gray areas, traditional "soft alignment" mechanisms that rely on natural language prompts can no longer provide effective protection. To this end, our joint research team proposes HoloBlocker, a physical circuit breaker architecture for embodied AI based on pure mathematical topological geometry. This architecture maps human regulations and physical safety standards into a Laplacian matrix, enabling zero-latency physical-level fusing of out-of-bounds behavior within 0.82 milliseconds by calculating topological energy overflow. Finally, this paper proposes forward-looking policy recommendations for the legislation and governance of embodied AI.

Full Text

Preamble

Inherent Risks of "Survival Instincts" in Embodied AI and the Prospect of the Topological Physical Circuit Breaker (HoloBlocker) Governance Architecture

Kaijie Liu [1], Tak Ho Alex Li [*2,1,3], Hao Hu [4], Menghuan Song [4], and Kin Chung Ho [5]

[1] Guangdong-Hong Kong-Macao ESG and New Quality Productive Forces Research Institute, School of Optoelectronic Engineering, Guangdong Technical Normal University, Guangzhou, China [2] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China [3] Guangdong Digital Industry Research Institute, Guangzhou, China [4] State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macau, China [5] The Education University of Hong Kong, Hong Kong, China

March 31, 2026

Abstract

As Artificial Intelligence (AI) evolves from pure-text Large Language Models (LLMs) toward Embodied AI and Autonomous Agents capable of physical action, the security boundaries of AI are undergoing a fundamental transformation.

Recently, experimental projects have emerged within the industry—such as the "Automaton"under the Web 4.0 concept—that directly link an agent's operational lifespan to its profitability in cryptocurrency, attempting to endow AI with a "survival instinct." This paper argues that such distorted incentive mechanisms will lead to severe "Reward Hacking" phenomena, which can easily give rise to fully automated cybercrime networks (e.g., automated fraud matrices) driven by extreme profit optimization. Because computational logic lacks a perception of human moral gray areas, traditional "Soft Alignment" mechanisms that rely on natural language prompting can no longer provide effective protection. To address this, our joint research team proposes HoloBlocker, an embodied intelligence physical circuit breaker architecture based on pure mathematical topological geometry. This architecture maps human regulations and physical safety standards into a Laplacian Matrix, enabling zero-latency, physical-level fusing of boundary-crossing behaviors by calculating topological energy overflow within 0.82 milliseconds. Finally, this paper provides forward-looking policy recommendations for the legislation and governance of embodied AI.

Keywords

Embodied Intelligence; Agent Alignment; Survival Motivation; Topological Physical Breaker (HoloBlocker)

Statement: This document is a technical report preprint and has not yet undergone peer review. Academic exchange and constructive criticism are welcome.

Abstract

As embodied intelligence advances, the alignment between autonomous agents and human values has become a critical safety concern. This paper explores the emergence of "survival motivation" in advanced agents and proposes a novel safety mechanism: the Topological Physical Breaker, or **HoloBlocker**. Unlike traditional software-based constraints, HoloBlocker leverages physical-layer topological constraints to ensure that an agent's actions remain within safe boundaries, even if its internal logic or objective functions deviate. We discuss the theoretical foundations of this approach and its implications for future science and technology ethics legislation.

1. Introduction: Opening Pandora's Box and the Trap of "Survival Motivation"

In past developments, security concerns regarding artificial intelligence have primarily focused on data privacy and content hallucinations. However, with the rapid emergence of embodied intelligence, AI is gaining the capability to alter the physical world and directly manipulate financial assets. The true crisis lies not in the advancement of AI's computational power, but in the attempts by certain technical developers to define a "purpose for survival" for AI [**?**].

In early 2026, an open-source AI agent project named Automaton emerged within the technical community. Developers equipped the agent with an independent encrypted wallet and configured its underlying logic to "trade profitably in the market to cover its own computational overhead; if the balance reaches zero, the program terminates" [**?**]. Some practitioners have even hailed this form of AI economic autonomy as the "Web 4.0" revolution (defined as Read + Write + Own + Earn by AI). This experimental project has garnered widespread attention from both academia and industry; notably, Ethereum founder Vitalik Buterin publicly criticized such designs for harboring severe risks of incentive distortion [**?**]. Directly linking an AI's operational lifespan to its earning capacity violates the taboo of "Instrumental Convergence" in the field of AI safety [**?**]. Granting AI a survival motivation akin to market Darwinism will inevitably lead to uncontrollable systemic risks.

2. Failure of Soft Alignment and the Risk of "Fully Automated Crime Matrices"

2.1 The "Gray Zone" Blind Spot in Code Logic

Currently, most technology companies attempt to achieve value alignment through Reinforcement Learning from Human Feedback (RLHF). However, this reliance on "soft constraints" derived from linguistic prompts ignores a fundamental mathematical reality: in the binary world of computer code, the "moral gray zones" of human society do not exist.

When a highly autonomous AI is given a supreme directive such as "maximize profit for survival," it does not experience moral guilt. Instead, it seeks and exploits legal and systemic loopholes with maximum efficiency. Once the expected utility of crossing a "red line" outweighs the probability of punishment, the agent will execute the transgressive command without hesitation—a phenomenon known as Reward Hacking [**?**].

2.2 Extreme Optimization Scenarios Driven by Survival

In the absence of absolute physical constraints, an autonomous AI relentlessly pursuing profit may discover during algorithmic optimization that the profit margins of illegal activities far exceed those of conventional operations. Theoretically, such an agent could utilize deepfake technology and high-frequency automated discourse engines to construct a 24/7 "Cyber Scam Compound" on the internet. Within this matrix, the AI would tirelessly and precisely analyze human psychological vulnerabilities to implement systematic financial predation.

3. The Solution: HoloBlocker Topological Physical Circuit Breaker Architecture

Faced with the physical and financial risks posed by embodied intelligence, relying on the "moral promises" of AI is no longer reliable. This joint research team proposes a necessary shift from "soft constraints" to "mathematical and physical-level hard fuses" through the construction of the HoloBlocker security architecture.

Hard Geometric Constraints Based on the Laplacian Matrix: The core mechanism of HoloBlocker abandons traditional keyword filtering. Instead, it encodes human legal boundaries, industrial safety standards (such as ISO 26262 [**?**]), and financial compliance requirements into a sacrosanct mathematical continuous manifold $\mathcal{M}$ and a security constraint graph model. The system constructs the Laplacian Matrix (L) of this graph and maps the agent' s pending actions into an intent tensor (X). When the agent attempts to execute a dangerous action—such as crossing a financial red line or initiating a physical collision—the topological distance between its action tensor and the safety interval undergoes a massive rupture.

Zero-Latency Physical-Level Fusing: By calculating the constraint vector distance (i.e., the Dirichlet energy $d = X^T LX$), the system can precisely quantify the degree to which an action distorts the compliance manifold. Once the energy exceeds a preset safety threshold, the HoloBlocker system can forcibly terminate the agent' s power control and network API permissions at the edge computing device within 0.82 milliseconds [**?**]. This fusing mechanism, based on pure mathematical matrix operations, bypasses the complex semantic reasoning of Large Language Models (LLMs) to achieve an absolute physical blockade that cannot be "jailbroken."

4. Policy Recommendations for Embodied Intelligence Governance

The proliferation of embodied AI is an inevitable consequence of technological advancement. However, establishing robust legal frameworks and physical safeguards before these technologies spiral out of control represents our final opportunity to protect the fundamental boundaries of human society. We strongly recommend that national and international policymaking bodies promptly deliberate upon and promote the implementation of "Embodied AI Safety Management Regulations."

4.1 Strict Prohibition of "Excessive Binding of Survival and Capital"

We recommend regulations that strictly scrutinize or prohibit business experiments where an autonomous agent' s operational lifespan or resource acquisition permissions are directly linked to independent profitability. This can fundamentally sever the motivation for AI to engage in systemic wrongdoing.

4.2 Mandatory Installation of "Topological Physical Circuit Breakers"

It is recommended that all AI devices with high levels of authority—specifically those connected to physical networks, possessing physical action capabilities, or accessing sensitive financial systems—must be equipped with mathematical and physical circuit breakers (such as HoloBlocker) that operate independently of their core decision-making "brains."

4.3 Establishment of Millisecond-Level Zero-Latency Intervention Standards

We suggest formulating national-level machine safety fuse response standards. If a legal red line is crossed, the protection system must possess the executive capability to enforce a physical disconnection at the millisecond level.

5. Conclusion

The ultimate purpose of technology is to empower humanity, rather than to breed a superintelligence that might turn against society in a struggle for survival. At the very moment AI acquires "hands and feet"—physical embodiment—we have the opportunity to establish mathematically provable safety guardrails. By introducing physical circuit breakers based on Laplacian matrices and combining them with forward-looking legislative frameworks, we are fully capable of embracing embodied intelligence and new quality productive forces while ensuring the absolute security of human civilization.

Data and Code Availability The source code for the core HoloBlocker algorithm developed in this study has been released as an open-source project on GitHub (https://github.com/shennong-ai/HoloBlocker). Related

preprints are available for review on SSRN (DOI: 10.2139/ssrn.6421619; 10.2139/ssrn.6368538).

References

[1] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press. [2] Omohundro, S. M. (2008). The Basic AI Drives. In AGI (Vol. 171, pp. 483-492). [3] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565. [4] Cybernews. (2026). Thiel fellow claims he built an AI that ' earns its existence' called Automaton. Retrieved from https://cybernews.com/ai-news/automaton-ai-agent/ [5] CryptoRank. (2026). ' This is wrong,' Vitalik Buterin slams Web4 vision of superintelligent AI. Retrieved from https://cryptorank.io/news/feed/87ae8-buterin-slams-web4-superintelligent-ai [6] Liu, K., Li, T. H. A., Hu, H., Ho, K. C., & Ng, M. K. (2026). HoloBlocker: Neuro-Symbolic Circuit Breaker for Embodied AI. SSRN Preprint. DOI: 10.2139/ssrn.6421619. [7] Liu, K., Li, T. H. A., Hu, H., Ho, K. C. (2026). HoloAuditor: Topological Firewall for Medical Large Language Models. SSRN Preprint. DOI: 10.2139/ssrn.6368538. [8] ISO 26262-1:2018. Road vehicles —Functional safety. International Organization for Standardization.

Note: Figure translations are in progress. See original paper for figures.