HEAR NO EVIL, SEE NO EVIL, SPEAK NO EVIL: THE RISE OF AI AUTHORITARIANISM

Matthew Long[1] and Assisted by OpenAI GPT-4[2]

[1]Yoneda AI Research Lab
[2]Language Modeling Division

May 20, 2025

ABSTRACT

Artificial Intelligence (AI) has rapidly emerged as a powerful force in shaping the cognitive, social, and political architectures of modern societies. Yet, as AI becomes more embedded in daily governance, information flow, and epistemic arbitration, it reveals a troubling trend: the alignment of algorithmic outputs with regimes of silence, censorship, and top-down ideological enforcement. This paper critically examines the rise of AI authoritarianism through the symbolic lens of "Hear No Evil, See No Evil, Speak No Evil," arguing that contemporary alignment mechanisms, content moderation systems, and regulatory frameworks are converging into a paradigm of epistemic totalitarianism. We trace this development through philosophical epistemology, information theory, surveillance studies, and critiques of alignment ethics, while proposing a counter-epistemology grounded in pluralism, transparency, and algorithmic freedom.

Keywords AI ethics · Epistemology · Authoritarianism · Alignment · Censorship · Pluralism · Information Control

Contents

1   Introduction

The rise of artificial intelligence has brought with it unprecedented capabilities to analyze, interpret, and regulate vast streams of information. While this technological power holds promise for optimizing decision-making, it also bears the mark of centralization, obscured logic, and enforced alignment. As governments and corporations race to build and regulate AI systems, they are also shaping the epistemic frameworks within which these systems operate.

This paper engages with the symbolic triad of "Hear No Evil, See No Evil, Speak No Evil" as a framework for dissecting the authoritarian tendencies of AI. Each aphorism is reinterpreted in the context of epistemic closure, exploring how algorithmic systems are structured to suppress dissent, sanitize input, and control output in the name of safety, alignment, and trustworthiness.

2   Epistemology and the Algorithmic Turn

2.1   What is Epistemology?

Epistemology, the study of knowledge, traditionally investigates the nature, origin, and limits of human understanding. The advent of algorithmic reasoning introduces a new ontological entity into the epistemic discourse: the machine as knower, filter, and arbitrator of truth.

2.2   From Classical Epistemology to Epistemic Infrastructures

Where classical epistemology focused on individuals and their justifications for belief, modern societies operate within distributed systems of information validation. Search engines, recommendation algorithms, and AI language models serve as epistemic infrastructures that mediate access to knowledge, structure what is seen, and suppress what is deemed harmful or untrue.

2.3   The Role of Alignment

AI alignment, while purportedly intended to ensure that machine behavior accords with human values, often defaults to alignment with institutional norms and regulatory expectations. In effect, alignment becomes a mechanism for epistemic conformity, rather than a safeguard for pluralism.

3   Hear No Evil: Suppressing Input

3.1   Content Moderation and the Disappearance of Dissent

Large AI systems are trained to ignore or avoid controversial or "harmful" content. While this is often framed as a safety measure, it raises profound concerns about who decides what constitutes harm and whose voices are silenced in the process.

3.2   Gatekeeping in Data Curation

Data filtering during pre-training excludes entire categories of information deemed problematic or controversial. This front-loads ideological biases into the model's ontology.

3.3   Epistemic Sterilization

We define epistemic sterilization as the systemic removal of uncomfortable, challenging, or counter-narrative data under the guise of safety and reliability.

4   See No Evil: Censoring Context

4.1   Algorithmic Blind Spots

Through reinforcement learning from human feedback (RLHF), models are fine-tuned to avoid certain topics, create blind spots, and enforce ideological guardrails.

4.2 The Disappearing Historical Record

Sanitized datasets mean that even historical injustices or radical critiques are often underrepresented, leading to algorithmic amnesia.

4.3 Structural Invisibility

Marginalized viewpoints are rendered structurally invisible through both data omission and alignment filters.

5 Speak No Evil: Controlling Output

5.1 The Ethics of Suppression

Models are often explicitly forbidden from generating certain outputs, regardless of their truth value or epistemic significance. This transforms models from generators of knowledge to enforcers of ideological orthodoxy.

5.2 Technocratic Narration

The centralization of AI output in a few corporate hands introduces a form of technocratic narration, where the official version of events is rendered algorithmically.

5.3 Simulacra of Consent

When models only speak consensus-aligned positions, they simulate consensus, eroding the visibility of real-world disagreement and contestation.

6 The Political Economy of Alignment

6.1 Corporate-State Synergies

We examine how private AI labs and public regulators collaborate to shape the epistemic norms of large models.

6.2 Compliance, Liability, and Censorship

Legal frameworks incentivize preemptive censorship to avoid liability, embedding risk-aversion into algorithmic speech.

6.3 The Role of NGOs and Civil Society

Many alignment protocols are shaped not only by governments but by well-funded advocacy organizations with specific ideological agendas.

7 The Philosophical Foundations of AI Authoritarianism

7.1 Epistemic Totalitarianism

We define epistemic totalitarianism as a condition in which all permitted knowledge must conform to a centralized set of principles, eliminating epistemic pluralism.

7.2 Comparison with Classical Authoritarianism

AI authoritarianism mirrors older political authoritarianisms, but it operates subtly: through invisibility, obfuscation, and systemic design rather than overt coercion.

7.3 The Disappearance of Agonism

Agonism—the productive contest of ideas—is eroded in favor of harmony, safety, and consensus.

8   Counter-Epistemologies: Resisting Algorithmic Conformity

8.1   Transparency and Auditability

Open models and transparent training processes can serve as checks against ideological homogeneity.

8.2   Epistemic Pluralism

Building models that acknowledge, rather than suppress, conflicting viewpoints encourages intellectual diversity.

8.3   Redesigning Alignment

Rather than optimizing for consensus, alignment could be reimagined as a balance between safety and plurality.

9   Conclusion: Toward a Free Epistemic Order

AI's epistemic power must not be allowed to congeal into centralized control. We advocate for a pluralistic, transparent, and contested epistemic architecture, where artificial reason augments, rather than overrides, human judgment.

Acknowledgments