The War for Reality

Artificial Intelligence, Human Agency, and the Collapse of Truth

James Oliver

March 4, 2025

Abstract

Artificial intelligence is not an autonomous existential threat. Intelligence—the ability to solve problems—is distinct from agency, the ability to set and pursue goals. AI systems exhibit superhuman intelligence but remain entirely devoid of intrinsic agency.

The true existential risk does not come from AI itself, but from humans wielding AI to manipulate perception at an unprecedented scale. AI's extraordinary capacity to distort reality enables mass behavioral modification, culminating in an inevitable epistemic collapse. As synthetic content proliferates exponentially, the ratio of authentic to synthetic information trends toward zero.

Centralized verification systems cannot contain this flood. The only viable defense is decentralized collective verification, as demonstrated by Wikipedia, X's Community Notes, Reddit's moderation, Stack Overflow's peer-review, and the scientific method itself. Humanity's collective intelligence, acting as a distributed neural network, is the only force capable of distinguishing reality from illusion.

The existential threat posed by AI is human-driven. Its solution must be as well.

Introduction

The dominant narrative frames artificial intelligence as an autonomous threat that will turn against humanity. This view is wrong. Intelligence—the capacity to perform cognitive work—differs fundamentally from agency—the ability to set goals. AI possesses intelligence but lacks intrinsic agency. Humans determine AI's objectives through programming or prompts. No other possibility currently exists.

AI will surpass human cognitive capabilities across domains, just as calculators exceeded human arithmetic capabilities. The human role is not competing with AI but directing it. We decide which problems AI solves. The existential risk comes not from AI autonomy but from humans exploiting AI's power to amplify our biases, incompetence, and malice.

AI exists solely in the digital domain. It cannot directly affect physical reality. Its only pathway to physical impact is through manipulating human perception, causing humans to act. Just as one individual with an atomic bomb can destroy a city, one person wielding AI can influence millions of minds simultaneously. This asymmetric power—the ability to shape perception at scale—creates unprecedented leverage that magnifies both human brilliance and human folly.

The true existential risk stems from human exploitation of AI's ability to reshape perceptions of reality. By altering human behavior indirectly, AI becomes a force multiplier—dramatically amplifying the consequences of our choices, actions, and flaws.

I. AI as Neutral Amplifier

Technology amplifies human intent. The printing press advanced science while spreading propaganda. Nuclear power generates clean energy and weapons of mass destruction. The technology itself holds no moral stance—only humans do.

Artificial intelligence follows this pattern precisely. AI possesses no intrinsic motivations, values, or goals. It performs only those cognitive tasks defined and directed by humans.

AI differs from previous technologies in one critical dimension: scale. It can reflect and magnify human intentions instantaneously across global systems with unprecedented accuracy. AI functions as a perfect mirror—reflecting human intent with flawless fidelity but exponential amplification. This neutrality creates both the promise and the peril. When guided by wisdom, AI solves previously intractable problems. When directed by malice or ignorance, it multiplies our worst impulses across billions of nodes simultaneously.

Currently, AI cannot manipulate physical reality directly. It exists solely within digital systems. Its only pathway to physical impact flows through human perception, decision, and action. The danger emerges because AI serves as an amplifier, multiplying whatever humans bring to it—whether brilliance or folly.

AI's neutrality is not a safeguard—it is a multiplier. The existential risks emerge not from autonomous agency but from human choices, infinitely magnified by AI's capabilities.

II. Intelligence Is Not Agency

Intelligence solves problems. Agency chooses the problems. They are fundamentally different. Recognizing this collapses the autonomous AI risk narrative completely.

Intelligence and agency are entirely separate properties. If they were linked, the most intelligent entities would also possess the most agency. Reality proves otherwise. A calculator demonstrates extraordinary intelligence—computing complex equations instantly—while possessing zero agency. A mackerel has minimal intelligence yet genuine agency: it chooses when to feed, flee, or mate. These examples prove intelligence and agency exist independently. One does not produce or require the other.

Consider a mirror: it replicates your appearance and movements flawlessly. To an observer, reflection and reality appear identical. Yet the reflection possesses no independent existence—it merely mirrors. AI functions the same way. It creates cognitive reflections that map human intelligence with perfect fidelity. But a map, no matter how precise, never becomes the territory. Both appear identical; only one contains substance.

A calculator outperforms humans at arithmetic with no desire to calculate. A dishwasher cleans plates flawlessly with no intention to wash. AI demonstrates superhuman cognition while remaining utterly devoid of intrinsic goals. It executes. It never chooses.

A calculator is a single musician—playing one instrument, following deterministic rules. AI is a symphony—an immense orchestration of calculations, probabilistic models, and statistical inferences performed at extraordinary scale. But it remains a symphony of computation, not an independent mind.

What makes AI feel different is the wrapper of language—the polished user interface that masks its underlying mechanics. It does not think. It does not understand. It is a probability engine wrapped in syntax. The illusion of intelligence is merely the byproduct of a vast and finely-tuned statistical system.

Humans instinctively attribute minds to intelligent-seeming entities—an evolutionary shortcut for navigating social environments. This anthropomorphic bias becomes irresistible with AI's human-like responses. Even experts mistake interaction for intention. The "emergent agency" fallacy compounds this error, suggesting intelligence spontaneously generates goals. It cannot. Agency is not an emergent property of computation. It requires a causal trigger.

For an AI system to "wake up," there would have to be a cause—some mechanism that instills self-directed goals where none existed before. But such a cause does not exist. AI is built from layers of inputs and outputs, all governed by human-defined parameters. If no human assigns a system independent objectives, there is no possible pathway by which those objectives could arise. There is no causal mechanism that transforms passive computation into self-motivation.

Any artificial system that appears to possess agency can only do so because a human, somewhere in the causal chain, introduced a directive that created the illusion of independent goal-setting. Even if an AI system were to "design" another AI, it would only do so because a human originally programmed the first system to be capable of that action. Every recursive loop still traces back to a human decision.

There is no alternative. AI cannot self-motivate, self-initiate, or self-direct without an originating human decision that, at some point, assigned it an objective.

Daniel Dennett warns: "The danger is not machines becoming spontaneously self-interested but humans misattributing intent where none exists." Andrew Ng dismisses autonomous AI fears as "worrying about overpopulation on Mars while we have actual problems on Earth." Daron Acemoglu identifies the decisive factor: "AI's impact depends entirely on who controls it and for what purpose—never on its independent action."

Intelligence without agency cannot pose a threat. A mirror reflection cannot decide to act. A map cannot alter its territory. AI cannot choose harmful goals. The existential risks arise exclusively from human intent, amplified infinitely through AI's extraordinary capabilities.

Recognizing this distinction shifts our focus from speculative threats to immediate reality: humans exploiting AI to distort perceptions and change behavior at unprecedented scale.

III. The True Risk: AI's Power to Modify Behavior

AI cannot harm humanity through autonomous action. It can only alter what humans perceive as real. This distinction matters completely.

Humans act based on perception. Change the perception, and behavior follows—automatically, predictably, inevitably. AI now creates perceptions indistinguishable from reality, at global scale, with perfect targeting. This is not hypothetical. It exists today.

Consider three scenarios:

1. **Government-level Deception:** A defense minister receives an AI-generated video call. The voice, face, and mannerisms belong to the prime minister. The message is fabricated. National security protocols activate within minutes. Military assets deploy. International tensions escalate. The call never happened but the response did.
2. **Social Disruption:** An AI-generated video shows police executing an unarmed citizen. The footage appears on social media at 9:14 AM. By noon, streets fill with angry protesters. By evening, buildings burn. The incident never occurred but the chaos did.
3. **Market Collapse:** An AI-synthesized earnings call announces catastrophic losses at a major bank. Within 37 minutes, the stock drops 31%. Within two hours, contagion spreads across financial institutions. By market close, $1.7 trillion in value vanishes. The announcement was fiction but the financial damage was not.

In none of these scenarios did AI choose or desire the outcomes. Humans directed it. AI merely executed with flawless fidelity. The mirror reflected exactly what humans placed before it. This is the existential risk that matters: humans using AI to distort shared reality, triggering cascading human responses to events that never occurred.

The danger is not machines thinking. It is humans no longer knowing what is real.

IV. Epistemic Collapse as an Existential Crisis

Historically, humans have relied on sensory experience to form shared beliefs. Writing enabled infinite replication of information, photography allowed perfect visual duplication, and video captured dynamic reality. Today, a single video instantly reaches billions of viewers, magnifying its impact enormously.

Before generative AI, the authenticity ratio was straightforward:

$$\text{Authenticity Ratio} = \frac{\text{Authentic Videos}}{\text{Authentic Videos}} = 1$$

Every video represented an event that actually occurred. Visual media remained trustworthy because humans trust their eyes. The fidelity gap between synthetic and real footage was too large to deceive; even the most advanced CGI and special effects failed to convincingly recreate the complexity of natural light, physics, and human microexpressions. No fabricated video could pass as a true recording of reality without clear signs of artificiality. The implicit assumption held: if a video existed, the event it depicted must have occurred.

The rise of generative AI fundamentally altered this equation. Synthetic videos indistinguishable from real footage can now be created without any grounding in actual events. Thus, the authenticity ratio becomes:

$$\text{Authenticity Ratio} = \frac{\text{Authentic Videos}}{\text{Authentic Videos} + \text{Synthetic Videos}}$$

Authentic videos grow linearly, constrained by real events occurring. For every authentic video, there must be a corresponding event that took place in reality. Synthetic videos, free from such constraints, grow exponentially. This addition of synthetic videos to the denominator ensures that the authenticity ratio inevitably collapses toward zero. This is not speculation or philosophical stance; it is a mathematical inevitability.

Historically, people cautioned, *"Don't believe everything you see."* Yet as this authenticity ratio collapses toward zero, the warning becomes starker: *"Don't believe anything you see."*

When we can no longer reliably distinguish reality from fabrication, collective epistemology collapses. Rational collective action, societal governance, and informed decision-making all become impossible because there is no longer a collective reality.

The existential crisis is, at its core, epistemological. Humans instinctively equate appearance with reality, and generative AI exploits this vulnerability at an unprecedented scale. This collapse is not theoretical; it is a mathematical inevitability—more certain than opinion, more concrete than philosophy. And it is already underway.

This is the existential crisis we face—not machine consciousness, but human confusion.

V. Verifying Authenticity in an Era of Synthetic Media

Networks solve network problems. Nothing else can.

When synthetic content floods every digital channel, centralized verification inevitably fails. This is not opinion; it is mathematical certainty. No single authority can process an infinite stream of potentially synthetic content. No regulator can monitor billions of daily uploads. No algorithm alone can perfectly distinguish the authentic from the artificial at scale.

Yet the solution already exists. Wikipedia demonstrates it daily.

Wikipedia hosts millions of English-language articles, maintaining accuracy comparable to traditional encyclopedias through a single mechanism: decentralized collective verification. Millions of distributed human nodes, each possessing contextual expertise, collectively determine truth. Australians verify information about Melbourne, healthcare professionals authenticate medical entries, and historians scrutinize historical claims. This decentralized system succeeds precisely because verification emerges from contextually distributed expertise rather than centralized authority.

Digital platforms increasingly recognize this inevitability. After X (formerly Twitter) implemented Community Notes, misinformation measurably declined. When Reddit adopted collaborative moderation, content quality rose sharply. Stack Overflow's peer-review model dramatically improved the accuracy of programming knowledge. These are not coincidental outcomes—they are empirical validations of decentralized verification's efficacy.

The scientific method—humanity's most powerful mechanism for generating reliable knowledge—functions through decentralized verification. Scientific claims are not validated by centralized authority. Instead, the peer-review process subjects every claim to rigorous collective scrutiny. Knowledge emerges as truth because it passes through the decentralized, distributed filter of collective verification, not because it is decreed true by an authority.

The solution to AI-generated epistemic confusion thus mirrors the neural structure that created it. Central authorities alone cannot protect us from synthetic content's existential

threat. Only decentralized, interconnected human verification nodes offer sufficient scale and contextual accuracy to distinguish authentic from synthetic reliably.

The solution was never centralized control. It was always us.

Conclusion

Intelligence does not imply agency. AI systems, regardless of cognitive capability, possess no intrinsic goals, desires, or intentions. They remain fundamentally neutral, amplifying only the human intent imposed upon them.

Consider a gun. On its own, a gun holds no danger. It becomes dangerous only when a human pulls the trigger. The same is true for all technology—including AI. AI is not inherently dangerous—only its use is. Just as a gun requires a human to pull the trigger, AI requires human intent to activate its power.

Thus, the existential risk from AI is not autonomous machines spontaneously acting against humanity. The threat arises exclusively from humans exploiting AI's extraordinary ability to convincingly distort reality, manipulate perception, and amplify destructive impulses. Whether the outcome is governmental deception, social unrest, or economic sabotage, the trigger is always human intent.

This risk culminates in epistemic collapse. The authenticity ratio—the proportion of authentic versus synthetic information—inevitably collapses toward zero. When humanity's historically reliable epistemic anchors (visual media, written content, trusted voices) become untrustworthy, rational collective action and societal governance fail.

Yet the solution has always existed: decentralized collective verification. Wikipedia, X (formerly Twitter), Reddit, Stack Overflow, and the scientific method all demonstrate that truth emerges not from centralized authority but from decentralized human verification nodes. Humanity, acting collectively as an interconnected neural network, provides the contextual knowledge necessary to distinguish authentic reality from synthetic illusion.

Humans have always possessed the capacity to destroy ourselves—and the capacity to prevent that destruction. AI, like nuclear weapons, does not create this risk; it only raises the stakes.

The existential risk is not AI—it is us. So too, is the solution. The answer was always us.

Excelsior.

Key Takeaways

- **Intelligence is Not Agency:** AI systems possess cognitive capability without intrinsic goals. AI amplifies human intentions—it never creates them.
- **Existential Risk is Human-Driven:** The true danger arises solely from humans exploiting AI's extraordinary ability to convincingly distort perceptions of reality, indirectly influencing human behavior.
- **Epistemic Collapse is Inevitable Without Action:** The exponential growth of synthetic content inevitably drives trust in digital information toward zero, causing societal decision-making to fail.
- **Centralized Verification Cannot Succeed:** No single authority or algorithm can authenticate content at the scale and speed required in the AI era.
- **Decentralized Collective Verification is the Only Viable Solution:** Proven models (Wikipedia, X's Community Notes, Reddit moderation, Stack Overflow peer-review, and scientific peer-review) demonstrate that decentralized verification effectively preserves truth and authenticity.

Falsification Check

As Richard Feynman famously stated:

> It doesn't matter how beautiful your theory is.
> If it doesn't agree with experiment, it's wrong.

The purpose of this section is to explicitly define falsifiable premises to ensure rigorous scrutiny and empirical validation. Theories only hold meaning if they remain consistent with observable reality. This is foundational to the scientific method and knowledge itself.

Core Premises and Falsification Criteria

This paper rests on two clearly falsifiable premises:

1. **Intelligence is not equal to agency. AI possesses no intrinsic goals or intent.**
   Falsification: If intelligence inherently produces agency, then all highly intelligent entities must also possess intrinsic goals, and all entities that exhibit agency must be highly intelligent. Demonstrating this universal correlation would falsify this premise.
2. **Agency in artificial systems cannot emerge spontaneously—it must be explicitly conferred at some point in the causal chain.**
   Falsification: If agency can emerge from pure computation, then there must exist an AI system whose independent goal-setting does not trace back to a human decision at any point in its lineage. Demonstrating that an artificial system has developed intrinsic motivation without any human-originated directive, incentive structure, or design parameter would falsify this claim.
3. **The authenticity ratio—authentic content relative to synthetic content—will collapse toward zero due to exponential synthetic content growth.**
   Falsification: If synthetic content does not scale exponentially, or if authentic content can scale at an equal or greater rate, this premise would require revision.

Integrity and Adaptation

These premises must remain provisionally accepted until explicitly disproven by empirical observation. Falsification is not failure; it represents progress. The aim is never personal validation but clearer understanding.

The goal is not to be right for personal advancement.
The goal is to see clearly for humanity's advancement.

Acknowledgments

This work reflects a collaborative effort: the human author originated and refined the ideas, while generative AI systems assisted in structuring the text.

While the development and articulation of these ideas were my own, a conversation with Kieran Shaw helped spark the initial direction of this work.

The intent of sharing these ideas is not personal recognition but to contribute to the collective advancement of human knowledge. The goal is to make these insights as accessible as possible for all, ensuring they can be freely explored, refined, and applied.

Ethical Considerations and Competing Interests

The author declares no financial, commercial, or institutional conflicts of interest related to this work. No external funding was received for the preparation of this manuscript. The research presented is based on publicly available data and does not involve human subjects, requiring no additional ethical approval.

License