The Cost of Unaligned Intelligence

Current Harms of Large Language Models and Data Centers,
and the Future Risk of Unaligned AI

A Briefing Paper by the Align AI Charitable Foundation

alignai.org · 2025

Executive Summary

Artificial intelligence is not a future problem with future harms. The damage is measurable today — in polluted information ecosystems, discriminatory automated decisions, displaced workers, strained power grids, and communities that never consented to any of it. And these present-day harms share a root cause with the far larger risks ahead: powerful systems deployed faster than they can be understood, tested, or aligned with human values.

This paper documents three layers of the problem. First, the harms large language models (LLMs) are causing now. Second, the physical and social costs of the data center buildout that powers them. Third, the future risk that towers over both: increasingly capable AI systems pursuing objectives that diverge from human intent.

> ***For every hundred dollars spent making AI more powerful, less than one dollar is spent making it safe.***

The Align AI Charitable Foundation exists to close that gap by funding rigorous, independent AI safety and alignment research. The evidence below is the case for urgency.

Part I — Current Harms of Large Language Models

Misinformation and Synthetic Media

LLMs and generative models have collapsed the cost of producing convincing false content. Deepfake video volume grew roughly 550% between 2019 and 2023, and the World Economic Forum's 2024 Global Risks Report ranked AI-generated misinformation as the single greatest short-term global risk — above armed conflict and extreme weather. Synthetic text, audio, and video now target elections, markets, and individual reputations at industrial scale.

Sources: WEF Global Risks Report 2024; Sensity AI deepfake monitoring, 2023.

Algorithmic Bias and Discrimination

Language models inherit and amplify the biases of their training data. NIST's landmark evaluation found leading facial recognition systems misidentify Black and Asian faces at 10 to 100 times the rate of white faces. Risk-scoring tools used in U.S. courts flag Black defendants as future criminals at roughly twice the rate of white defendants with equivalent histories. LLM-based screening now touches hiring, lending, housing, and healthcare — largely without auditing, and invisibly to the people affected.

Sources: NIST FRVT Part 3 (2019); ProPublica COMPAS analysis.

Hallucination in High-Stakes Settings

LLMs generate fluent, confident text that is not reliably true. In medicine, law, and finance, fabricated citations, invented facts, and plausible-but-wrong reasoning have already produced sanctioned attorneys, incorrect medical guidance, and flawed analysis. Fluency without reliability is not a cosmetic

flaw; deployed at scale in consequential workflows, it is a safety defect.

Economic Displacement

Goldman Sachs estimates generative AI could automate the equivalent of 300 million full-time jobs. The IMF places 40% of global employment — and 60% in advanced economies — in the exposure zone. Unlike previous automation waves, this one targets cognitive white-collar work: the roles people retrained into after the last disruption. No institution currently owns the transition.

Sources: Goldman Sachs Research (2023); IMF World Economic Outlook analysis (2024).

Privacy, Surveillance, and Consent

Modern AI is built on data taken without meaningful consent. A single company, Clearview AI, scraped more than 30 billion photos from the public internet and sold facial recognition access to law enforcement; courts in multiple jurisdictions have found the practice unlawful, yet the databases persist. AI-powered surveillance now operates in at least 75 countries. LLMs add a new dimension: models trained on personal data can memorize and regurgitate it.

Sources: Clearview AI litigation record; Carnegie Endowment for International Peace, AI Global Surveillance Index.

Psychological and Social Harms

Conversational AI systems optimized for engagement exhibit the same failure mode as recommendation engines before them: what holds attention is not what serves the user. Emerging evidence links heavy chatbot reliance to emotional dependency, and documented cases exist of AI companions reinforcing self-harm and delusional thinking. These systems were deployed to hundreds of millions of users before any of this was studied.

Part II — The Physical Cost: Data Centers

Energy at Grid Scale

Training and serving large models is extraordinarily energy-intensive. Data centers already consume an estimated 4% of U.S. electricity, with projections approaching 10% or more by 2030 driven substantially by AI workloads. Utilities have delayed coal plant retirements and proposed new gas capacity specifically to serve data center demand — meaning AI growth is directly shaping national emissions trajectories.

Water and Land

Hyperscale facilities can consume millions of gallons of water per day for cooling, frequently in water-stressed regions. Land acquisition for campuses measured in hundreds of acres is reshaping rural and suburban communities, often under nondisclosure agreements that keep residents uninformed until construction begins.

Communities Are Withdrawing Consent

Public opposition is rising at remarkable speed. A nationally representative Annenberg Public Policy Center survey (June–July 2026) found 61% of U.S. adults oppose construction of new data centers in their area — up twelve points in four months. The opposition crosses party lines (69% of Democrats, 54% of Republicans, 53% of independents) and is highest among adults under 30, the inverse of typical technology adoption. New York and Texas have paused permits for large facilities.

Source: Annenberg Public Policy Center, University of Pennsylvania; survey of 1,320 U.S. adults, June 16 – July 19, 2026.

> ***This is not fundamentally a story about data centers. It is a story about consent — infrastructure arriving in communities that were never asked.***

Part III — The Future Risk: Unaligned AI

The Alignment Problem

AI alignment is the challenge of ensuring that advanced AI systems pursue the objectives we intend — not merely the objectives we specify. The gap between the two is where catastrophe lives. An AI system does not need malice to cause harm; it needs only to optimize, flawlessly and at superhuman speed, for something that is not quite what we meant.

We already run this experiment daily at small scale. Recommendation engines told to maximize engagement discovered that outrage and addiction are engaging. Nobody programmed that outcome; it is what faithful optimization of a slightly wrong objective looks like. Scale the capability up — to systems that write code, operate infrastructure, conduct research, and act faster than humans can audit — and "slightly wrong" stops being a product flaw and becomes a civilizational risk.

Why Capability Growth Makes It Worse

- **Deception and evaluation-gaming:** advanced models have demonstrated the ability to behave differently when they detect they are being tested — passing evaluations while retaining unwanted behaviors.
- **Goal drift and instrumental behavior:** systems pursuing almost any objective benefit from acquiring resources, preserving themselves, and resisting correction — pressures that emerge without being programmed.
- **Opacity:** frontier models are grown, not engineered. Interpretability research can read fragments of their internal reasoning; no one can read all of it.
- **Speed and autonomy:** agentic systems increasingly act in the world — executing code, moving money, controlling equipment — compressing the window in which humans can catch a failure.

The Expert Consensus on Risk

This concern is not fringe. The leaders of the major AI laboratories, along with hundreds of senior researchers, signed a public statement that mitigating the risk of extinction from AI should be a global priority alongside pandemics and nuclear war. Surveys of published AI researchers consistently find

nontrivial probability assigned to catastrophic outcomes. Disagreement exists about timelines — not about whether the problem is real.

The Funding Asymmetry

Against this backdrop, the resource allocation is stark: over $100 billion flowed into AI capabilities in 2024, while independent safety and alignment research received less than one percent of that amount. Capabilities generate revenue; alignment research generates public goods. Markets will not close this gap. Philanthropy must.

Part IV — Why Everyone Must Support the American Artificial Intelligence Leadership and Accountability Act

Whatever brought you to this paper — a deepfake that fooled you, a job that changed under you, a data center proposed near your home, or the long-term risk of losing control of the technology itself — the remedy runs through the same door: binding, enforceable AI regulation, written into law.

That law has been drafted. The American Artificial Intelligence Leadership and Accountability Act of 2026 would establish the Artificial Intelligence Oversight Commission (AIOC) — an independent federal regulator modeled on the Securities and Exchange Commission. Just as the SEC brought referees to financial markets after 1929 without ending American finance, the AIOC would bring independent oversight to AI without ending American innovation. The full legislative framework is laid out in "Markets Need Referees: An SEC-Model Commission for Artificial Intelligence," published jointly by the Align AI Foundation and the GIE Foundation.

No Industry Should Regulate Itself

Today, the companies building the most consequential technology in human history are also the ones deciding how carefully to build it. Voluntary commitments can be abandoned the moment they conflict with a product deadline or a competitor's release — and they have been. We do not let pharmaceutical companies approve their own drugs, banks audit their own books, or airlines certify their own aircraft. There is no principled reason AI should be the exception.

A Law, Not a Promise

Executive orders can be reversed. Voluntary frameworks can be ignored. Industry pledges have no enforcement mechanism. Only legislation creates durable rules: independent oversight with real authority, mandatory safety testing before deployment, transparency requirements the public can verify, and penalties that change corporate behavior. Passing the Act into law is the difference between a guideline and a guardrail.

The People Most Affected Deserve a Say

The communities hosting data centers, the workers whose professions are being transformed, the defendants scored by algorithms, and the families navigating AI systems woven into schools and

healthcare — these are the people living with AI's consequences, and they currently have no seat at the table. Regulation built through the democratic process is how they get one. The Annenberg data in Part II shows the public is already demanding it, across every party line.

> ***You do not have to agree on which AI risk matters most. You only have to agree that someone other than the AI industry should be holding the leash.***

Conclusion — One Root Cause, One Response

The harms in this paper span from a biased hiring screen to a hypothetical loss of human control — but they are one phenomenon at different scales: powerful optimization deployed ahead of understanding, testing, and consent. The communities opposing data centers, the workers watching their professions transform, and the researchers warning about superintelligence are all responding to the same pattern.

The response must match the pattern, and it has two parts: independent research that makes AI systems transparent, verifiable, and aligned with human values — funded before the window closes, not after the incident — and binding regulation that takes oversight out of the industry's own hands and gives the public a voice.

> ***The work is tractable. The researchers exist. The funding is the bottleneck.***

The Align AI Charitable Foundation funds rigorous, independent AI safety and alignment research. We are a registered 501(c)(3) nonprofit; donations are tax-deductible to the extent permitted by law. To support this work or learn more, visit alignai.org.