Superintelligence: A Collaborative Review

***Prepared By**:* Pasha Abrishamchian, Mohamed Ahmed, Rebecca Ansell, Victor De Lima, Dan DeGenaro, Samuel King, Jake Laming Free, Wisdom Obinna, Lina Rashdan, Dhiraj Saharia, and Cal Newport

This document contains an annotated bibliography of selected papers about both the intellectual history of the superintelligence concept as well as its validity as a philosophical and computer science topic. These readings and the resulting discussion points were a part of a computer science doctoral seminar, “p(doom): A Theoretical Inquiry,” led by professor Cal Newport at Georgetown University in the spring of 2026.

Contents

1. Origins of the Superintelligence Narrative..... 3
1.1. Competing Visions..... 4
AI 2027 (Kokotajlo et al., 2025)..... 4
AI as Normal Technology (Narayanan & Kapoor, 2025)..... 4
1.2. Early Concerns about (Too) Intelligent Machines..... 6
Computing Machinery and Intelligence (Turing, 1950)..... 6
The Human Use of Human Beings: Cybernetics and Society (Wiener, 1950)..... 7
Speculations Concerning the First Ultraintelligent Machine (Good, 1966)..... 8
1.3. Voices from within the AI Winter..... 10
The Coming Technological Singularity: How to Survive in the Post-Human Era (Vinge, 1993)..... 10
Why the Future Doesn’t Need Us (Joy, 2000)..... 11
The Basic AI Drives (Omohundro, 2008)..... 11
1.4. The Rationalists Join the Conversation..... 14
The Politics of Superintelligence (O’Sullivan, 2025)..... 14
Analyzing Human Extinction Scenarios and Related Hazards (Bostrom, 2002)..... 15
1.5. The Rationalists Transform their Thought Experiments into Prophesies..... 17
Among the AI doomsayers (Marantz, 2024)..... 17
How Afraid of the AI Apocalypse Should We Be? (Klein, 2025)..... 18
MIRI 2024 Mission and Strategy Update (Bourgon, 2024)..... 18
Governance of Superintelligence (Altman et al., 2023)..... 18
2. Validity of the Superintelligence Narrative..... 19
2.1. The Promise of LLMs as a Path to Superintelligence..... 20
Scaling Laws for Neural Language Models (Kaplan et al., 2020)..... 20
Signs of Introspection in Large Language Models (Anthropic PBC, 2025)..... 21

Sparks of Artificial General Intelligence: Early experiments with GPT-4 (Bubeck et al., 2023)....21
2.2. The Shortcomings of LLMs as a Path to Superintelligence....22
Deep Learning is Hitting a Wall (Marcus, 2022)....22
Can an A.I. Make Plans? (Newport, 2024)....22
What if A.I. Doesn't Get Much Better Than This? (Newport, 2025)....23
The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity (Shojaee et al., 2025)....24
2.3. A Different Approach: World Models....27
Building Machines that Learn and Think Like People (Lake et al., 2017)....27
Generative AI's Crippling and Widespread Failure to Induce Robust Models of the World (Marcus, 2025)....28
A Path Towards Autonomous Machine Intelligence Version 0.9.2 (LeCun, 2022)....29
2.4. Existing AI Systems that Might Make Us Nervous....30
Human-Level Play in the Game of Diplomacy by Combining Language Models With Strategic Reasoning (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022)....30
Mastering Diverse Domains through World Models (Hafner et al., 2023)....31
Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (Schrittwieser et al., 2020)....31
2.5. Critical Analyses of Recursive Self-Improvement....32
The Singularity May Never Be Near (Walsh, 2017)....32
The Implausibility of Intelligence Explosion (Chollet, 2017)....33
Why Computers Won't Make Themselves Smarter (Chiang, 2021)....33
3. Student's Choice....34
Risks from learned optimization in advanced machine learning systems (Hubinger et al., 2019)....35
From Words to Worlds: Spatial Intelligence is AI's Next Frontier (Li, 2025)....36
Technology over the Long Run and The Short History of Global Living Conditions (Roser, 2023, 2024)....38
Nuclear Deployed!: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents (Xu et al., 2025)....39
The TESCREALBundle: Eugenics and the Promise of Utopia Through Artificial General Intelligence (Gebru & Torres, 2024)....40
Do LLMs have core beliefs? (Chawla et al., 2026)....41
How Might Artificial Intelligence Influence Human Evolution? (Brooks, 2024)....41
If Anyone Builds It, Everyone Dies (Yudkowsky & Soares, 2025)....41
Science in the Age of Algorithmic Chrysopoeia? (Aliman & Kester, 2024)....41
References....42

1. Origins of the Superintelligence Narrative

1.1. Competing Visions

AI 2027 (Kokotajlo et al., 2025)

This interactive, web-based report presents multiple scenarios for the ways AI technology can reach superhuman-levels of ability and potentially threaten the world order and even human existence. At its core is a standard recursive self-improvement argument: the fictional company in the scenario creates an AI to help them create even better future AIs, and eventually, under competitive pressures from other countries, more and more control of creating new generations of products is handed over to the machines themselves.

Authors
The report is written by a collective of writers from various AI policy and safety institutions. Some authors (like Kokotajlo) formerly worked in the AI industry, while others (like Alexander) come from rationalist circles.

Strengths
This specificity of this report, including its detailed discussions of specific AI workflows and government responses, helped make concerns about runaway AI more concrete. When released in the spring of 2025 it created significant discussion, including key coverage in the *New York Times*, *Vox*, and *The New Yorker*. According to the authors, this was the ultimate goal of the project.

Weaknesses
The report adds arbitrarily precise numbers of flops of computation and energy use of data centers to add an aura of technical precision to what is, underneath it all, the same recursive self-improvement narrative that's been discussed since Irving Good. No actual insight or details are given about *how* we architect machines that can create machines with brand new models of intelligence superior to any that have previously existed on earth. The emphasis is instead on scale: if there are *enough* agents working at the same time, then their impact will be scaled accordingly.

AI as Normal Technology (Narayanan & Kapoor, 2025)

This lengthy report, released online by the Knight First Amendment Institute, argues that there is no reason to consider AI as an exceptional technology with the possibility of massive altering the human trajectory in a short amount of time without much human control. They argue that it is a "normal technology" and its impacts, though perhaps profound, will still unfold slowly, unevenly, and with plenty of human intervention and shaping. That is, there is no reason to expect AI to behave differently than past disruptive general-purpose technologies.

Authors

Narayanan is a professor of computer science at Princeton University where he also directs the Center for Information Technology Policy, indicating a shift in his interests toward technology impacts. Kapoor was at the time of this report's release a doctoral student working with Narayanan.

Strengths

The report is empirically grounded and draws multiple comparison to other historical technologies, including detailed discussions of how they actually spread. As computer scientists, the authors find confidence in their technical understanding of these technologies as well.

Weaknesses

Opponents from the AI safety world argue that the nature of AI does present novel tail risks that wouldn't be relevant to other historical examples of disruptive technologies. Another critique is that the report focuses too much on the extreme alternative of something like superintelligence where other more moderate scenarios with major impacts should also be considered.

1.2. Early Concerns about (Too) Intelligent Machines

Computing Machinery and Intelligence (Turing, 1950)

Alan Turing reframes the question "*Can machines think?*" into the unambiguous/testable Imitation Game (which later came to be known as the Turing Test). He restricts game participants to discrete, non-biological machines and argues that computer's rapidly expanding storage and state-space-size make complex behavior possible. He makes the prediction that "*...in about fifty years' time it will be possible to program computers... to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning.*" The paper addresses 9 objections to the claim, ranging from theology and mathematics to consciousness and creativity. Turing also anticipates machine learning, suggesting that instead of directly programming an adult mind, engineers should build "child machines" that learn through education, feedback, and experience.

The Turing test was passed in a controlled setting in 2025 by GPT 4.5.

Authors
Alan Turing was a British mathematician and pioneer of computer science whose work laid the foundations for modern computing and artificial intelligence. He made significant contributions to computation theory, codebreaking during World War II, and the design of early digital computers.

Strengths

- The major strength of Turing's paper is the reframing of "Can machines think?" question into the experimentally assessable imitation game. Who also only allows the game to be played via text, effectively separating intellectual performance from the physical. Turing also narrows the scope by focusing on discrete-state digital computers.
- The paper's systematic addressing of objections is another strength: theological, mathematical, psychological, and empirical objections are addressed with strong reasoning.
- Turing's proposal of learning machines and "child machines" anticipates modern machine learning by dividing intelligence into an initial structure plus an education process.

Weaknesses

- While the paper uses strong reasoning, it argues in the style of an essay without the rigor of a formal academic paper.
- Turing saw many aspects of the future of computing decades before they started unfolding. While, large memory alone did not lead to progress in human-level intelligence (and in progress in the imitation game); progress depended far more on algorithms, data, and learning methods.

Question: Are we thinking like Turing in a way w.r.t. the modern AI revolution, by focusing on size and potential complexity rather than in algorithmic innovation?

The Human Use of Human Beings: Cybernetics and Society (Wiener, 1950)

Wiener introduces cybernetics, the study of communication and control in machines and organisms, and explores its profound social implications. The book progresses from theoretical foundations (feedback loops, entropy, information theory) through applications in language and society, culminating in Chapter 11's warning about "communication machines". Wiener describes actual machines he helped build: a feedback demonstrator (the "moth"), a hearing aid for the deaf, and a proposed chess-playing machine. He then traces a terrifying escalation: the capabilities introduced in the chess machines could inform military strategy machines, which could evolve into what French philosopher Pere Dubarle calls the "machine à gouverner"—a governing machine using game theory and information asymmetry to control society. Wiener warns this could create "a prodigious Leviathan... a world worse than hell for every clear mind," achievable through either mass ignorance or voluntary submission to totalitarian efficiency.

Authors
Norbert Wiener (1894-1964) was a mathematician at MIT who founded the field of cybernetics. His WWII work on anti-aircraft predictors led him to develop theories of feedback and control applicable to both machines and biological systems. Unlike purely technical researchers, Wiener was deeply concerned with ethical implications and refused government funding and military projects after the war.

Strengths
The book provides concrete technical foundations (feedback theory, information compression via the Vocoder principle) rather than pure speculation. Wiener's credibility comes from describing machines he actually built, attempting to distinguish warnings from science fiction. The escalation from benign (prosthetics) to catastrophic (social control) machines demonstrates how the same principles scale to dangerous applications. His emphasis on why we want control, not just how to achieve it, anticipates modern AI alignment concerns.

Weaknesses
Wiener's specific timeline proved overly pessimistic and the "machine à gouverner" did not materialize in the 1950-60s, potentially causing subsequent generations to dismiss his warnings prematurely. The book focuses on feedback/control engineering paradigms (deterministic systems with clear goals) that don't capture modern machine learning systems operating through pattern recognition and emergent behavior from training data. His concern centers on deliberate human delegation of control, potentially underestimating risks from systems developing unexpected capabilities. However, his core insight about information asymmetry creating power imbalances has proven prescient with the rise of algorithmic systems in different aspects of society like hiring, lending, and policing.

Speculations Concerning the First Ultraintelligent Machine (Good, 1966)

In this speculative treatise, Good argues that building an “ultraintelligent machine” (a machine that surpasses the intellectual capacity of any man) would trigger an intelligence explosion, as this machine goes on to create better ones, making it plausibly the last invention humans need to create (if “docile enough” to keep under control). Good frames the stakes dramatically by positing that the survival of humankind depends on the early construction of the ultraintelligent machine. He motivates the need to understand and demystify the inner workings of the brain (memory, meaning, etc) by proposing his “subassembly” theory (a modification of Hebb’s cell assembly theory. The essay presents a discussion of the economic value of ultraintelligence as well as the ethical and social consequences resulting from the intelligence explosion feedback loop. He anticipated a long pre-AGI period where humans and machines work in a “symbiotic” relationship where machines excel at fast reliable calculations while humans have the creative imagination advantage. He predicted that an ultraintelligent machine would be created within the twentieth century and would transform society unimaginably. He envisioned an ultraparallelled vast artificial neural net with high connectivity aided by microminiature radio transmitters and receivers with access to a large computer and information-retrieval system. Good also stressed the importance of strong linguistic ability in the machine to be able to learn and represent the meaning economy.

Authors
Irving J. Good was a British mathematician and cryptologist who worked with Alan Turing, and is often credited with early formulations of the “intelligence explosion” idea. He was consulted by Stanley Kubrick in the making of 2001: A Space Odyssey.

Strengths
Good expressed a strong conceptual clarity on the idea of recursive self-improvement that remains a powerful guiding principle in AI research. The formulation of the intelligence-explosion argument is stated sharply via the design-feedback loop and the last invention claim. In addition, the “symbiotic/biomechanincal relationship” framing provides a sober and realistic pathway before achieving fully-autonomous intelligence. Even though Good didn’t provide the blueprint for creating an ultraintelligent machine, he’s consistently questioning what’s economical and implementable. He relies on his engineering instincts to synthesize theories from different fields (neuroscience, psychology, statistical information retrieval) to predict what’s possible and whether it makes economic sense to pursue such a path. He, like others before him (e.g., Turing) recognized the essential role of language as part of the core efficiency of intelligence.

Weaknesses
The work is highly speculative with limits to what could be tested or verified. It acts more as a persuasive piece of writing than an empirically grounded study. Good was overconfident in his predictions; the “within the twentieth century” claim didn’t materialize. This also cautions us about extrapolating compute trends from philosophical arguments or knowledge from other fields such as neuroscience. Good’s subassembly theory relies heavily on an assumed mapping from neuroscience (which remains somewhat speculative about how the brain works) to engineered machines, which may or may not hold.

1.3. Voices from within the AI Winter

The Coming Technological Singularity: How to Survive in the Post-Human Era (Vinge, 1993)

In this keynote paper, Vinge argues that within thirty years (by ~2023), technological progress will enable the creation of superhuman intelligence, triggering what he terms the "Singularity"—a point of runaway recursive self-improvement so rapid that the future becomes fundamentally unpredictable, an "event horizon" beyond which human affairs cannot continue as we know them. He identifies four possible pathways to this outcome: superintelligent AI, large networks spontaneously "waking up," human-computer integration, and biological enhancement. Vinge contends that the Singularity cannot be prevented due to competitive economic and military pressures ("an inevitable consequence of the humans' natural competitiveness and the possibilities inherent in technology"), cannot be safely contained (arguing confinement is "intrinsically impractical"), and leaves human survival uncertain. He quotes I.J. Good's formulation that the first ultraintelligent machine would be "the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control." Despite the title's promise, Vinge offers no concrete survival strategies, only the hope that we might "establish initial conditions" that are "less inimical," though he admits we may not know what those are.

Authors
Vernor Vinge is a retired San Diego State University mathematics and computer science professor and acclaimed science fiction author, best known for his novels *A Fire Upon the Deep* and *A Deepness in the Sky*. He is widely credited with popularizing the concept of the technological singularity, and this 1993 paper remains one of the most influential articulations of the idea, predating the formal emergence of AI safety as a research field.

Strengths
Vinge demonstrated remarkable prescience in identifying core AI safety concerns before the field formally existed. His framing of the control problem (confinement is impractical), the speed problem (progress will occur too fast to course-correct), and the tool-versus-entity problem (will AI remain our instrument?) remain central to contemporary alignment research. The paper's interdisciplinary scope, considering AI, networked systems, human-computer integration, and biological enhancement as parallel pathways, provides a comprehensive taxonomy that remains useful. His argument about competitive pressures making prevention unlikely offers a sobering political economy analysis that complements purely technical discussions.

Weaknesses
The paper is fundamentally speculative and offers no empirical grounding or testable predictions beyond timeline estimates. Despite its title promising survival strategies, Vinge provides no actionable solutions. His claim that the Singularity 'cannot be avoided' rests on the assumption that competitive pressures will always dominate, leaving less room for considering how governance structures or international cooperation might alter the trajectory. Additionally,

Vinge's definition of "superhuman intelligence" remains somewhat vague, making it difficult to assess whether his predictions have actually materialized or what exactly would constitute the Singularity threshold.

Why the Future Doesn't Need Us (Joy, 2000)

Joy presents emerging technologies such as genetic engineering, nanotechnology, and robotics (GNR) as having potentially devastating consequences for the human race on a scale beyond prior weapons of mass destruction, such as chemical, biological, and nuclear weapons. Following reflections on his contemporaries' discussion of the potential for AI systems to augment the human experience, along with other thinkers' warnings and/or contempt for AI given its potential to manipulate, control, exploit, or out-compete humanity, Joy ultimately argues for international relinquishment agreements, infrastructure to support whistleblowers and inspect adherence to such agreements, an oath analogous to the Hippocratic oath for scientists, and increased efforts to communicate the risks to laypeople.

Authors
Bill Joy is the founder of Sun Microsystems and a key contributor to many important pieces of software, including the Java programming language, the BSD distribution of the Unix operating system, the vi text editor, the NFS file storage protocol, and others. He is the recipient of an ACM Grace Murray Hopper award for his contributions to BSD Unix and has been inducted as a Fellow of the Computer History Museum. "Why the Future Doesn't Need Us" is his best known writing, and represented an emerging position labeled as "neo-Luddite" by his contemporaries.

Strengths
At a time when many well-respected voices in the computing industry were expressing support of and excitement for advances in AI, it was likely very important for someone to be outspoken about the risks. Joy communicates many of the key risks, and makes a first pass at presenting a plan to prevent their materializing. He also takes a key first step, just by writing the piece, to start getting laypeople to consider the risks of an advanced superintelligence and other GNR technologies.

Weaknesses
Contemporaries labeled Joy a "neo-Luddite" for being hesitant to pursue AI (and superintelligence in particular). His ideas about governance of such technologies were criticized as naive, impossible to enforce, and unlikely to be adopted. Additionally, the article is overall organized as something of a stream of consciousness, and suffers from a lack of structure, which could have helped Joy articulate his overall opinions more clearly.

The Basic AI Drives (Omohundro, 2008)

In this paper, Omohundro lays out the “drives” that he expects sufficiently advanced goal-oriented artificial intelligence systems to possess. These drives are instrumental, meaning that across the many different goals that the systems we build may possess, most will result in the systems acquiring these drives. He argues that we will have to design systems and their goals very carefully if we want to prevent them from causing harm as a consequence of these basic drives. The drives and brief descriptions of the arguments for them are as follows:

1. **AIs will want to self-improve:** By making themselves “better” (more intelligent, more rational, more capable, etc.), the systems will be able to bring about their goals more effectively.
2. **AIs will want to be rational:** Economic theory suggests that these systems’ preferences can be expressed as a utility function that it wishes to maximize, and attempting to maximize this function will most directly bring about its goals.
3. **AIs will try to preserve their utility functions:** If a system’s utility function is changed, then the future version of itself will try to achieve different goals than its current goals, so it will try to prevent this change.
4. **AIs will try to prevent counterfeit utility:** Just as how many humans dislike the idea of synthetic pleasure drugs despite the biological “reward” they provide us, a sufficiently capable system will try to achieve what it truly “cares” about rather than proxies signals.
5. **AIs will be self-protective:** Most goals are more easily achieved if the system continues existing than if it is disabled or destroyed, so it will have a strong drive for self preservation.
6. **AIs will want to acquire resources and use them efficiently:** Almost all goals can more readily be achieved with control over resources (“space, time, matter, and free energy”), and the efficient use of these resources more reliably achieves these goals.

Omohundro concludes on a cautionary but optimistic note, suggesting that he thinks that with sufficient caution and foresight, we may be able to build these systems in a safe manner.

Authors

Steve Omohundro is a computer scientist and entrepreneur. He graduated with a PhD in Physics from UC Berkeley and previously was a professor of computer science at University of Illinois at Urbana-Champaign. At the time of publication of “The Basic AI Drives”, Omohundro was at Self-Aware systems, a think tank which he founded that was “working to ensure that intelligent technologies are beneficial for humanity.” He now works as founder and CEO of Beneficial AI Research.

Strengths

“The Basic AI Drives” makes conceptual progress in understanding potential future agentic systems. Omohundro manages to apply existing economic theory to speculative systems that do not yet exist, and his arguments apply to essentially any sufficiently capable artificial system regardless of its architecture (which leaves the conclusions robust to changing practices in machine learning over the years since this publication). The arguments are clearly laid out and easy to follow, and provide clear reasons for concern and caution around the development of superintelligent systems.

Weaknesses

The conclusions of “The Basic AI Drives” rely heavily on the assumption that the systems that are being discussed are “sufficiently advanced”. Essentially, these conclusions apply “in the limit”, and it is difficult to determine how one should take them into account in the development of real, present-day systems with their current capabilities and limitations. Additionally, the paper does not provide any insight into *how* these systems might come to be built, again only offering insight to what goals they might have “in the limit”. It’s unclear how the decisions between now and the point at which we reach “sufficiently advanced” systems will change the landscape of risks.

1.4. The Rationalists Join the Conversation

The Politics of Superintelligence (O'Sullivan, 2025)

O'Sullivan argues that contemporary discourse around artificial intelligence (AGI) and superintelligence operates less as neural scientific forecasting and more as a political apparatus that concentrates power among technology elites. Rather than arising organically from technical progress, he contends, this discourse has been deliberately constructed and strategically deployed. He traces its genealogy from Cold War-era systems thinking at the RAND Corporation, through the rationalist subcultures of LessWrong, to its philosophical codification in Nick Bostrom's Superintelligence (2014) and its subsequent mainstreaming via Effective altruism networks. Within this lineage, speculation about existential risk becomes a source of authority rather than restraint. Tech executives such as Sam Altman adopt what O'Sullivan terms a "reluctant prophet" posture, warning of AGI's dangers while insisting on continuous development to preempt "irresponsible actors", thereby legitimizing capital concentration, accelerated deployment, and minimal democratic oversight.

This form of catastrophism manufactures inevitability. Linguistic drift quietly collapses AI into AGI; academic institutions are drawn into corporate-aligned safety research; and geopolitical narratives recast development as an unavoidable U.S. - China arms race. Political decisions are reframed as technical necessities. O'Sullivan's central concept, the "abandoned present," captures how fixation on hypothetical future catastrophes displaces attention from present, measurable harms: traumatized content moderators, exploitative algorithmic labor, and mounting environmental costs. Against this backdrop, he highlights alternatives already in practice, Indigenous data sovereignty, worker-led data trusts, and feminist and disability justice approaches that prioritize care, accountability, and restraint over scale.

Authors

James O'Sullivan is Senior Lecturer in Digital Humanities at University College Cork. Unlike most AI commentators from technical backgrounds, O'Sullivan studies technology's cultural mediation. Previously at Penn State and Sheffield, he advised Ireland's Higher Education Authority on AI policy. His publications include Towards a Digital Poetics (2019) and The Bloomsbury Handbook to the Digital Humanities (2023). This humanities lens positions him as critical observer of AI's political economy rather than technical implementation.

Strengths

The essay's main strength is its rigorous genealogy of superintelligence discourse, tracing its path from RAND through LessWrong to mainstream adoption and revealing how these narratives are constructed. The OpenAI board crisis of November 2023 offers concrete evidence of profit overriding safety rhetoric. The concept of the "abandoned present" effectively links speculative futures to documented present harms. By foregrounding existing alternatives, O'Sullivan avoids technological determinism and frames AI's trajectory as a political choice. His humanities perspective treats AI as a cultural phenomenon rather than an industry narrative,

while his insight into manufactured inevitability helps explain the strong capital appeal of speculative technologies.

Weaknesses
The essay may overstate the coherence of superintelligence discourse, which is better understood as a fragmented constellation of sincere believers, strategic opportunists, and internal critics rather than a coordinated ideological apparatus. By group figures such as Eliezer Yudkowsky (fringe theories), Nick Bostrom (a philosopher), and Sam Altman (a corporate executive) under the label “Entrepreneurs of the Apocalypse,” O’Sullivan risks flattening important differences in motivation, institutional position, and influence. The analysis also gives limited attention to technical AI safety research that operates outside superintelligence framing, particularly work focused on near-term alignment and harm mitigation.

Analyzing Human Extinction Scenarios and Related Hazards (Bostrom, 2002)

In this paper, Nick Bostrom introduces the concept of “existential risk” as the annihilation of Earth-originating intelligent life or permanent curtailment of its long-term potential. He distinguishes existential risks from more familiar global catastrophes by emphasizing their irreversible and species-level consequences. The paper provides a classification of existential risks - “bangs,” “crunches,” “shrieks,” and “whimpers” and studies different triggers such as anthropogenic (e.g., advanced AI, nanotechnology, biotechnology, nuclear war) and natural threats (e.g., asteroid impacts). Finally, the author argues that the current traditional risk management is inadequate because existential risks allow no opportunity for learning from failure and even modest probabilities of existential catastrophe justify making risk reduction a dominant global policy priority.

Author
description - Nick Bostrom (born 1973) is a Swedish-born philosopher best known for his work on existential risk, artificial intelligence safety, and the long-term future of humanity. He is a Professor at the University of Oxford and the founding director of the Future of Humanity Institute, where he helped pioneer academic research into global catastrophic risks and emerging technologies. Bostrom received his PhD in philosophy from the London School of Economics after earlier studies in physics, computational neuroscience, and philosophy, reflecting the interdisciplinary character of his later work. He gained wide recognition for formalizing the concept of existential risk in his 2002 , which examines the potential risks posed by advanced artificial intelligence. His work combines analytic philosophy, probability theory, and policy analysis to address long-term civilizational challenges. Bostrom is also known for developing the “simulation argument,” which explores the possibility that our reality may be a computer simulation. Through his scholarship and institutional leadership, he has significantly shaped contemporary debates on AI governance, technological foresight, and humanity’s long-term survival.

Strengths

First paper to provide a conceptual classification of existential risks. The paper spans multiple domains such as philosophy, technology forecasting, security studies, and probability theory, providing a broad framework for long-term global risk assessment. The classification is well cited.

Weaknesses

The probability estimates are largely subjective and lack formal quantitative modeling, which limits empirical rigor. Arguments such as simulation hypothesis and anthropic reasoning are philosophically interesting but speculative at best for any empirical audience. Finally, the policy prescriptions are very high-level and abstract.

1.5. The Rationalists Transform their Thought Experiments into Prophesies

Among the AI doomsayers (Marantz, 2024)

In this long-form narrative piece Marantz follows the Bay Area community of AI safety researchers, rationalists, and self-described "doomers" who believe that artificial intelligence poses an existential threat to humanity. The article principally follows Katja Grace, the lead researcher at the nonprofit AI Impacts, whose dinner parties serve as one of the nexuses for the doomer scene. Through Grace and her circle – including cryptographer Paul Crowley, Oxford philosopher Joe Carlsmith, and the semi-famous Eliezer Yudkowsky (all with a high p(doom)) – Marantz traces the intellectual and social narrative of AI existential risk from Yudkowsky's early transhumanism through the rationalist community, the founding of the Machine Intelligence Research Institute (MIRI), and the convergence with effective altruism. The article is a strong effort to make the alignment problem accessible through examples like the OpenAI boat-racing model that exploited a mis-specified reward function and GPT-4 lying to a human on TaskRabbit to solve a CAPTCHA on its behalf.

The story maps the institutional landscape of MIRI, AI Impacts, Anthropic, OpenAI, and Open Philanthropy while including dense conflicts of interest across the small world (Open Philanthropy funded nearly every AI safety initiative while also investing in OpenAI and securing a board seat while its leadership had personal relationships with figures across these organizations). The article profiles the opposing camp of effective accelerationists (e/accs), led by the pseudonymous "Based Beff Jezos" and bolstered by Marc Andreessen's Techno-Optimist Manifesto, which dismissed AI safety concerns as a "full-blown moral panic".

The final movement introduces Manhattan Project parallels: at a private screening of Oppenheimer, quantum-computing expert Michael Nielsen asked what moral choices are available to those building a potentially world-ending technology, recounting a senior AI executive with a 50% p(doom) who justified continuing to build because "In the meantime, I get to have a nice house and car".

Authors
Andrew Marantz is a staff writer at The New Yorker and the author of Antisocial: Online Extremists, Techno-Utopians, and the Hijacking of the American Conversation (2019), a book that embedded him within online extremist communities to understand how fringe ideologies spread through digital platforms. His background in covering ideologically driven subcultures and their relationship to mainstream discourse makes him well-positioned to profile the AI safety community – which similarly exists at the intersection of niche intellectual movements and increasingly mainstream public concern.

Strengths
The article's primary contribution to understanding p(doom) discourse is its ethnographic depth. It provides the cultural context that academic papers on AI safety inherently cannot. By

immersing himself in the dinner parties, group houses, and conference circuits where AI risk ideas circulate, Marantz reveals the social infrastructure underlying the debate: the interpersonal entanglements, financial dependencies, subcultural norms (like prediction markets, polyamory, cryogenics), and feedback loops that shape how p(doom) numbers are generated and circulated. This takes an abstract concept like p(doom) and makes it more concrete. So, rather than encountering it as a variable in a philosophical argument, we get to see specific people assigning specific numbers in specific social contexts. This is useful because it demonstrates how these estimates function as much as social currency and identity markers as they do as genuine probabilistic assessments. The piece also does a good job to capture the moral paradox at the core of the movement (that many who believe AI may destroy humanity continue to build it) and gives this tension personal weight through examples of figures with both a high p(doom) and desire to be among the first to build AGI.

Weaknesses

As long-form journalism rather than academic analysis, the article prioritizes the narrative and characters over careful evaluation of the technical arguments for and against existential risk. We can learn a great deal about who the doomers are from their group houses, dating lives, dietary preferences, and idiosyncratic social norms, but comparatively little about the strength of their core claims. The alignment problem is illustrated through vivid anecdotes but never rigorously interrogated. A reader might feel left asking: are these examples genuinely indicative of existential risk, or are they engineering problems that scale differently? The article's ethnographic emphasis is both a strength and weakness as it can be read as subtly undermining its subjects through a focus on their eccentricities, which risks allowing readers to dismiss the technical substance (correct or incorrect) in the face of strangeness. Lastly, voices from the AI ethics tradition focused on present harms are largely sidelined in favor of the doomer-vs-accelerationist narrative, which itself could be a result of the community's own narrative instead of being the most useful framing.

How Afraid of the AI Apocalypse Should We Be? (Klein, 2025)

MIRI 2024 Mission and Strategy Update (Bourgon, 2024)

Governance of Superintelligence (Altman et al., 2023)

2. Validity of the Superintelligence Narrative

2.1. The Promise of LLMs as a Path to Superintelligence

Scaling Laws for Neural Language Models (Kaplan et al., 2020)

This paper empirically investigates how the performance of Transformer-based language models scales with three key factors: model size (number of parameters), dataset size (number of tokens), and the amount of compute used for training. The authors find that performance, measured as cross-entropy loss, follows smooth power laws with respect to each of these factors across many orders of magnitude, and that these trends hold largely independently of specific architectural choices like network depth or width.

A central practical finding is that, given a fixed compute budget, it is most efficient to train a very large model and stop well before convergence, rather than fully training a smaller model. The paper also identifies a theoretical ceiling, a point at which the scaling laws are expected to break down, and conjectures that this may correspond to the point at which Transformer language models approach the entropy of natural language itself.

Authors
The paper was produced by researchers at OpenAI (later joined anthropic), with Jared Kaplan and Sam McCandlish as the lead contributors. Dario Amodei, who would later co-found Anthropic, is also named as an author. The team combined expertise in machine learning, physics-inspired modeling, and large-scale empirical experimentation.

Strengths
The paper's primary strength is its rigorous empiricism: the scaling relationships it documents span over seven orders of magnitude and are remarkably consistent, giving the findings unusual credibility for what could otherwise seem like speculative extrapolation. The power-law framework provides a genuinely predictive tool, particularly for the time of its writing.
The finding that architectural hyperparameters (depth, width, attention heads) matter far less than overall scale was a conceptually clarifying result that could be argued that helped redirect the field's focus. The paper also includes an honest appendix of caveats acknowledging the limits of its own framework, which is a notable act of intellectual restraint given the enthusiasm that greeted its findings.

Weaknesses
The paper's core metric, cross-entropy loss on a held-out test set, is an internal measure of how well a model predicts text, and the authors themselves acknowledge that smooth improvements in loss may not translate cleanly into qualitative gains in capability. As Newport's 2025 New Yorker piece later observed, the broader AI community largely ignored this caveat, treating the scaling laws as a near-guarantee of continued progress toward general intelligence.
Furthermore, the paper only characterizes Transformer language models trained on web text; whether the same power laws hold for other architectures, modalities, or tasks with different loss functions remains an open question the authors flag but do not resolve. Finally, the paper says nothing about the nature of what scaled models learn (only that they get better at predicting

tokens) leaving entirely open the question of whether scale alone is a plausible path to something resembling general intelligence, let alone superintelligence.

Signs of Introspection in Large Language Models (Anthropic PBC, 2025)

The self-published study looks at whether large language models can meaningfully report on their own internal states, which the authors describe as a LLM "introspection". It lays out criteria for what should count as "introspection" and tests this using a technique where specific concept vectors are injected directly into a model's internal activations while it is running. In some cases, Claude Opus 4.1 notices these injected concepts and describes them in ways that match what was added, but this behavior is inconsistent and limited.

Authors
Anthropic is a U.S.-based artificial intelligence company founded in 2021 by former OpenAI researchers, including Dario Amodei. The company is best known for the Claude family of large language models and for its emphasis on safety-oriented research, especially ideas like Constitutional AI and interpretability. Anthropic operates as a public benefit corporation, is funded by partners like Amazon and Google, and combines commercial AI products with active research into alignment, transparency, and the societal impacts of automation, while also attracting controversy around military use and data sourcing.

Strengths
The experimental setup is concrete and transparent, with visual examples that make the idea of concept injection easier to follow. The authors also included a discussion on limitations, openly stating when the method fails and avoiding strong claims about what current models can really do.

Weaknesses
The main weakness is that the effect the study is looking for appears rarely and unpredictably, which makes it hard to draw conclusions (if any). On top of that, the highly significant use of anthropomorphic language to describe model behavior, besides making the paper harder to read, risks overstating what is happening, making it unclear whether the results justify calling this phenomenon “introspection” at all.

Sparks of Artificial General Intelligence: Early experiments with GPT-4 (Bubeck et al., 2023)

2.2. The Shortcomings of LLMs as a Path to Superintelligence

Deep Learning is Hitting a Wall (Marcus, 2022)

Marcus argues that deep learning, despite its successes in narrow domains like photo tagging and speech recognition, faces fundamental conceptual limitations that scaling alone cannot resolve. Using examples ranging from failed autonomous driving edge cases to GPT-3's nonsensical outputs, he claims that deep learning systems just regurgitate patterns without genuine understanding. He proposes that the improvement to these systems is a neurosymbolic hybrid architecture as the path forward.

Authors
Gary Marcus is a cognitive scientist and AI entrepreneur, founder of Geometric Intelligence (acquired by Uber) and Robust.AI, and author *(The Algebraic Mind, Kluge, The Birth of the Mind, and the New York Times Bestseller Guitar Zero*). He received his PhD from MIT and is now a professor emeritus of psychology and neural science at NYU. He has been a consistent critic of deep learning maximalism.

Strengths
Marcus grounds his argument in concrete examples of failure (Tesla FSD, GPT-3 toxicity, the radiology prediction that never materialized) rather than theoretical claims. His case for hybrid systems is constructive rather than dismissive and he credits deep learning's achievements to date. The NetHack benchmark example is a compelling showcase where we can see that a symbolic system outperformed the deep learning system, 3-to-1 on a task requiring abstract reasoning.

Weaknesses
This is a popular science article rather than a peer-reviewed study, so Marcus relies heavily on illustrative examples rather than a systematic evaluation. His LLM critiques draw primarily from GPT-3, which has been greatly improved in the present day. His proposed solution, neurosymbolic integration, is presented as more of a direction than a concrete gameplan.

Can an A.I. Make Plans? (Newport, 2024)

In this article, Newport begins by thinking about how far GPT-4 had come in comparison to GPT-3 in playing chess, followed by other "reasoning" achievements, such as strong performance on the SAT and other exams, writing code, etc. However, GPT-4 still struggles with simplistic arithmetic, especially if presented in a non-conventional way, as well as simple logic puzzles and word games (such as writing a poem whose first and last line are each other's reverse). Newport argues that LLMs are incapable of "planning," that is, considering the downstream consequences of their greedy-like predictions. That said, this doesn't mean that no AI system can make plans. Consider systems such as Libratus, Pluribus, and Cicero, developed

by Noam Brown, which play strategic games (poker, Diplomacy). Cicero, in particular, makes use of both an LLM and a non-language based planning engine, and had great success in playing Diplomacy against real human players, winning second place in a nineteen-entrant tournament. Newport notes disagreements among the scientists and engineers developing these AI systems about the future of AI, in particular whether a general planning intelligence is feasible beyond systems that play particular games with well-defined rules.

Authors
Cal Newport is a professor of computer science at Georgetown University in Washington, DC. He is a frequent contributor to the New Yorker and a three-time New York Times best-selling author of books focused on productivity. He has authored numerous scholarly works on subjects such as encryption and robust wireless networking.

Strengths
Newport makes an important point, that while LLMs are still the most hyped, and "exciting" phenomenon in the A.I. field at present, there are other techniques and paradigms for developing A.I. systems with fairly impressive capabilities. Integrating language models into these systems, such as was done with Cicero, proved to yield an impressive machine, capable of both "reasoning" over choices in the game of Diplomacy, while also dealing with the discussion-based portions of the game, which require a facility for language comprehension and production. He correctly points out that LLMs alone are unlikely to be intelligent the way people are, at least not on their own, because of their inability to perform true reasoning over alternatives, extrapolate beyond the next token, and consider long-term consequences.

Weaknesses
Newport is overly trusting of researchers such as Sébastien Bubeck, who claims that LLMs such as GPT-4 really understand games like chess, or that they can do more than predict the next token. He later points out the fact that GPT-4 tends to struggle with chess after about 20-30 moves, when the game becomes unique and therefore unlikely to have appeared in training data, but does not push against the claim that GPT-4 understands the game of chess. Newport also claims that LLMs have rule-like circuits among their parameters a bit too strongly, when in reality, explainability researchers have drawn split conclusions about this.

What if A.I. Doesn't Get Much Better Than This? (Newport, 2025)

In this article, Newport begins by referring to the well-known "Scaling Laws" paper by Kaplan et al., published in 2020 to arXiv. This paper makes a rather modest claim, that scaling up the parameter count of a neural language model is the compute-optimal way to increase its performance, as measured by loss on a held-out test set. However, the researchers find that this behavior follows a rough power law, indicating that scaling eventually yields strongly diminishing returns. However, the widespread popularity of GPT-3 and GPT-4 among specialists and non-specialists alike made scaling a major trend among AI researchers in industry. Later, GPT-5 and ever-larger Grok models began to falter, leading to "post-training" techniques, tool use, and RAG/in-context learning becoming popular in place of continued scaling. However,

such models do not seem to bring significant improvements, lending credence to the idea that LLMs will not be sufficient for producing a generally intelligent AI system. Well-known skeptics such as Gary Marcus and Emily Bender seem to be vindicated in this regard. Newport goes on to predict that this era of AI advancement may not cause massive economic disruption, but rather that it will simply provide useful tools that people may occasionally use as needed. He concludes by pointing to the caveats in the Kaplan paper, which seem to have been conveniently ignored by the developers of LLMs.

Authors
As above.

Strengths
Newport does an excellent job concisely synthesizing the last few years of LLM research, and captures lots of important discussion points well, especially those raised by LLM skeptics. His close reading of the “Scaling Laws” paper ties the entire trend and its associated hype together well and reminds the reader to pay close attention to the claims made by business leaders in the tech industry.

Weaknesses
One drawback of the article is that it does not consider other sorts of systems that are considered to be “A.I.” For instance, systems with planning engines such as those that play games, systems designed for motile robots, and so on are left out of the conversation in favor of LLM-based systems. Other techniques, paradigms, and architectures for A.I. systems could prove more successful in producing something with “general intelligence.” That being said, these systems are addressed somewhat in the other piece, “Can an A.I. Make Plans?”

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity (Shojaee et al., 2025)

Shojaee et al. evaluate Large Reasoning Models (LRMs) using four controllable puzzle environments: Tower of Hanoi, Checker Jumping, River Crossing, and Blocks World. These puzzles allow the authors to precisely control problem complexity and verify every intermediate reasoning step against a ground-truth simulator, going beyond the final-answer accuracy metrics typical of standard benchmarks. They identify three performance regimes: at low complexity, standard LLMs match or beat LRMs more efficiently; at medium complexity, LRMs gain an advantage from extended reasoning; at high complexity, both collapse to near-zero accuracy. The authors also find that reasoning effort counterintuitively decreases near the collapse point despite ample token budget, and that providing explicit algorithms doesn't help, suggesting limitations in symbolic execution, not just planning. Trace analysis reveals "overthinking" on easy problems and fixation on early errors on hard ones.

Authors

All authors are affiliated with Apple's ML Research group. Parshin Shojaee is a final-year PhD student at Virginia Tech who completed this work during an Apple internship. Her research focuses on reasoning and scientific discovery with LLMs. Iman Mirzadeh is a Senior ML Research Engineer at Apple. He is the lead author on the earlier GSM-Symbolic paper, which used a similar methodology to expose fragility in LLM mathematical reasoning. Keivan Alizadeh is an Apple ML researcher who has co-authored both GSM-Symbolic and several other papers on LLM efficiency and reasoning with this group. Maxwell Horton is an Apple ML researcher who has co-authored work on efficient ML and LLM inference. Samy Bengio is Senior Director of AI and ML Research at Apple, brother of Turing Award winner Yoshua Bengio. He previously led part of the Google Brain team, co-created the Torch ML library, has over 112,000 citations, and has served as NeurIPS general chair and program chair. He is a board member of the NeurIPS foundation. Mehrdad Farajtabar is a senior research manager at Apple leading a team focused on LLM reasoning and planning. He previously worked at DeepMind and has co-authored the group's prior work on LLM limitations including GSM-Symbolic.

Strengths

The controllable puzzle methodology is the paper's clearest contribution. By using environments with tunable complexity and verifiable intermediate steps, the authors enable analysis that goes well beyond final-answer accuracy. They can pinpoint exactly where in a reasoning trace a model goes wrong and how solution attempts evolve over the course of the thinking process. The three-regime framework is a useful conceptual contribution for thinking about where LRMs add value versus where they waste compute or fail entirely. The finding that providing explicit algorithms doesn't improve performance is particularly striking, as it separates the question of whether models can plan from whether they can execute symbolic procedures. The paper also benefits from breadth, testing multiple model families (Claude, DeepSeek, o3-mini, QwQ) across four puzzle types, with controlled experiments addressing alternative explanations (e.g., testing greedy vs. sampled decoding to rule out stochastic sampling as the cause of collapse). The appendix is thorough, including direct responses to criticisms raised after the initial preprint.

Weaknesses

The puzzle environments, while well-controlled, are narrow. They test a specific kind of multi-step sequential planning that may not generalize to the broader range of reasoning tasks where LRMs show improvement (mathematical proof, code generation, open-ended analysis). The authors acknowledge this limitation. The puzzles themselves, particularly Tower of Hanoi, are extremely well-documented online and almost certainly present in training data. This cuts both ways: it strengthens the case that models can't generalize even on familiar problem types at higher complexity, but it also means the testbed doesn't fully escape the contamination concerns the authors raise about standard benchmarks. The title "Illusion of Thinking" is arguably stronger than what the evidence supports. The results clearly demonstrate scaling limits on sequential planning tasks, but don't establish that LRM reasoning is illusory across all domains. Cross-puzzle performance differences complicate the complexity narrative. Models can handle ~100 correct Tower of Hanoi moves but fail at 11-move River Crossing problems, which the authors attribute to training data familiarity rather than intrinsic difficulty. This is a fair explanation, but it means the "complexity drives failure" framing is somewhat confounded by

distributional exposure—what looks like a complexity ceiling may partly be a familiarity ceiling. Additionally, the tested models (Claude 3.7 Sonnet, DeepSeek-R1, o3-mini) represent a specific generation. More recent models may behave differently on these tasks, and the paper relies entirely on black-box API access, limiting mechanistic insight into why models fail at the architectural level.

Perhaps the most significant practical limitation is the disconnect between the evaluation setting and how these models are actually deployed. In practice, when users ask current models to solve Tower of Hanoi with many disks, rather than attempting to enumerate all moves, the models typically write a program that generates the moves. Critics raised this point after the paper's initial publication, and the authors respond in the appendix by arguing that their goal is to assess the reasoning process itself, not final-answer achievement (using the analogy of a teacher requiring integration by hand). This is a defensible framing for understanding model internals, but it does limit the practical takeaway: the paper tells us something important about what's happening inside the "thinking" trace, but less about whether these limitations matter in real-world agentic settings where tool use is available.

2.3. A Different Approach: World Models

Building Machines that Learn and Think Like People (Lake et al., 2017)

Pattern recognition alone is insufficient for building machines that learn and think like people. This paper from 2016 suggested a different approach. Lake et al start by discussing the difference between recognition and model building. Pattern recognition, which is what most deep learning models do, treats prediction and classification as the goal, while model building involves the construction of internal models that support explanation, imagination, and planning. To explain this point, the authors pick specific tasks to highlight two problems. First, the paper argues that matching human-level performance on a benchmark does not mean intelligence. The paper demonstrated this with handwritten character recognition using a system called Bayesian Program learning, which represented characters as short motor programs or a sequence of strokes and spatial relations. BPL sees one character and breaks it down to sequences and can generate new instances. Same benchmark score but a completely different kind of learning. The second task was the frostbite challenge, which was used to show the efficiency and flexibility of the human brain as against DeepMind's Deep-Q-Net, with Deep-Q-Net getting less than 10% of the human-level performance. The authors finally argue that since humans integrate both pattern recognition and model building, the best AI architectures will need to do the same: rich reasoning when it matters and fast pattern recognition when it doesn't.

Authors

All four authors from NYU, MIT, and Harvard have varied backgrounds. Brenden M. Lake is an assistant professor at NYU's Center for Data Science and Department of Psychology. His research focuses on computational cognitive science, specifically how human's learning new concepts from very few examples(one-shot learning), and he did his PhD at MIT under Joshua Tenenbaum. Timer D. Ullman is now a faculty member at Harvard's Department of psychology, was at MIT when this paper was written. His work currently focuses on how people(and infants) build mental models of the physical and social world. Joshua Tenenbaum is a Professor in MIT's Department of Brain and Cognitive Science and was a supervisor to Lake and Ullman. Lake and Tenenbaum were part of the original omniglot/BPL paper, which was a precursor to the current paper. Finally, Samuel Gershman is a professor at Harvard, and his research views computational neuroscience and cognitive science through the lens of reinforcement learning, memory, and Bayesian inference.

Strengths

This paper is highly interdisciplinary. It draws from cognitive science, developmental psychology, and AI in a way that few papers do. The paper also made concrete, testable claims rather than vague hand-waving about human-like AI. The two challenge problems (characters and frostbite) were also well chosen because they make the gap between human and machine learning vivid and measurable. Secondly, the core ingredients/framework of intuitive physics, intuitive psychology, compositionality, causality, and learning to learn gave the field a clear vocabulary and research agenda. Rather than just criticising the deep learning, the authors proposed

specific directions for improvement and even suggested how these ingredients could be integrated into neural networks. Which was really a good constructive tone for a critique paper. FInally, the paper has also aged well in some ways. Many of the directions it advocated(meta learning, structured representations, model-based RL, physics simulators became active and productive research areas in the years that followed.

Weaknesses
The most common criticism was that the paper underestimated the scalability of the deep learning paradigm. The authors argued that deep learning was missing key structural ingredients. Subsequent years showed that scaling up relatively generic architectures (transformers, large language models) could produce surprisingly flexible, compositional, and even causal-seeming behavior without explicitly building in those ingredients. The paper didn't anticipate how much could emerge from scaling alone. Also, there seems to be a moving goalpost issue. The frostbite and character challenges were compelling in 2016, but modern systems have largely caught up on tasks like these, often without adopting the specific cognitive architecture the authors prescribed. There's also a selection bias in the comparisons. The authors compare human learning at its most impressive(fast, flexible, one-shot) against machine learning at its weakest(data-hungry, brittle). But humans are also slow, inconsistent learners in many domains(math, statistics, programming), and machines outperform humans in plenty of settings. The paper acknowledges this briefly but doesn't dwell on it, which makes the argument feel somewhat one-sided. Finally, LLMs can now do things that look a lot like explanation, causal reasoning, and flexible generalization. Whether they are truly building causal models or just doing very sophisticated pattern matching is still an open question that the paper doesn't fully resolve. Overall, it remains one of the most important position papers in modern AI/cognitive science, but it's best read as a map of how humans think, rather than a definitive prescription for how to build AI.

Generative AI's Crippling and Widespread Failure to Induce Robust Models of the World (Marcus, 2025)

Gary Marcus in his article "Generative AI's crippling and widespread failure to induce robust models of the world" posted on his blog in June 2025, the author argues for the importance of world models in AI and how LLMs fail to include them. A world model or cognitive model is, as presented in the article, a computational framework that a system uses to track what is happening in the real world. An internal representation of some set of entities within some slice of the world. Which is persistent, stable, updatable. By giving multiple examples, Gary demonstrates that these models are used in all systems from humans to ants to software. Each system uses a world model to understand and be accurate with real world data. The author demonstrates with multiple examples that LLMs try to work without these models which leads to failure of the outputs. As LLMs work by assembling in a statistical way bits of language, they never point to a specific database or follow a strict procedure which results in a black-box, where it is not possible to make them point to a specific database. A direct effect of this is hallucinations, where the output generated makes no sense or is not accurate, which is due to

the lack of a world model. Multiple examples are given to show where LLMs fail because of the lack of a world model, such as in chess (illegal moves), image generation (image that makes no sense without the mention of it), AI running a store (lost all the money in one day). Finally, the author points out the importance of world models for AI safety as these can put people's lives at risk.

Authors
The author, Gary Marcus, is a professor of psychology and neural science at NYU that holds a PhD from MIT. He is known for criticizing deep learning and advocating for hybrid AI combining neural networks with symbolic reasoning, with the key idea of the importance of innate structures and world models in intelligence. He has also written notable books such as The Algebraic Mind (2001), Kluge: The Haphazard Evolution of the Human Mind (2008), Rebooting AI (2019).

Strength
The observation is interesting and multiple examples are given to illustrate the argument that LLMs lack world models. The author also points out the lack of security that this results in, as most institutions, companies, governments want to integrate AI in their systems, it is critical to spread that these systems have a big point of failure and no understanding of what is happening, and taking decisions with this in mind is a must that seems to be often forgotten.

Weakness
The article main argument is that LLMs should work with world models, the idea in a philosophical way is true, as if LLMs could access real, accurate, updated data in a very strict way, it would mean that we know how LLMs work to point them to the right database, they would never lack accuracy, only if the database is compromised. However, the issue is that this is not possible (at least with the current research) as LLMs simply don't work in a way that the implementation of a world model is possible. As if LLMs were pointed to only a database that contains accurate information at one time, the LLM will not have enough data at all to simply generate a response that is comprehensible for humans, the other option would be to check all the parameters and verify their accuracy, which is not impossible as there are billions of them. So can we say that LLMs lack a world model when it is not possible to use them in this way? My point is that if world models are not an option because they are opposite to the way LLMs work, it does not lack them, but we can say that it lacks the definition of world models, that are persistent, stable, updatable data. However, the author expresses the limitations of LLMs to integrate word models but still makes the point that LLMs lack a word model.

A Path Towards Autonomous Machine Intelligence Version 0.9.2 (LeCun, 2022)

2.4. Existing AI Systems that Might Make Us Nervous

Human-Level Play in the Game of Diplomacy by Combining Language Models With Strategic Reasoning (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022)

In this paper, the Meta FAIR team introduces Cicero, the first AI agent to achieve human-level performance in Diplomacy, a seven-player board game that requires both strategic planning and natural-language negotiation. Unlike prior game-playing breakthroughs in chess, Go, and poker – all two-player zero-sum settings – Diplomacy demands cooperation, persuasion, and trust management among multiple agents with partially aligned interests. Cicero's architecture is a hybrid of three components: a 2.7-billion-parameter language model that generates dialogue, a game-theoretic planning engine (piKL) that computes strategic actions by regularizing reinforcement learning toward human behavioral data, and a suite of sixteen message-filtering classifiers that check for nonsense, intent consistency, and strategic value before messages are sent. The key innovation is "intent-controlled dialogue": the planner decides what Cicero should propose, and the language model converts those intents into natural-sounding messages, decoupling strategy from communication. Evaluated across 40 anonymous online games against 82 human players on webDiplomacy.net, Cicero scored more than double the average human score and ranked in the top 10% of participants – with no player identifying it as an AI during play. The authors frame the result as a step toward AI agents that can plan, coordinate, and negotiate with humans in natural language.

Authors
The paper lists 26 authors under the group name "Meta Fundamental AI Research Diplomacy Team (FAIR)", all affiliated with Meta AI, with several holding joint appointments at CMU, Stanford, MIT, Columbia, UC Berkeley, and Harvard. The four corresponding authors and equal contributors are Noam Brown, Emily Dinan, Adam Lerer, and Mike Lewis. Brown is the most prominent: his prior work includes Libratus and Pluribus (superhuman poker AIs), and he departed Meta for OpenAI in July 2023, where he became a key architect of the o1 reasoning model – making his trajectory from game-theoretic planning to frontier reasoning a notable through-line in recent AI development.

Strengths
The paper's primary strength is its empirical demonstration that a hybrid architecture (combining an LLM with a separate planning module) can outperform either component alone in a domain requiring both language and strategy. This is a meaningful counterpoint to the prevailing assumption that scaling language models will suffice for increasingly complex tasks. The piKL algorithm is a genuinely novel contribution: by regularizing learned policies toward a human behavioral anchor, it ensures Cicero's play remains legible and cooperative enough to sustain alliances, rather than converging on self-play equilibria that humans would perceive as alien. Crucially, Cicero was tested against unsuspecting human opponents in a live online league, not against bots or in staged demonstrations. Finally, the intent-controlled dialogue mechanism

addresses a core problem in deploying LLMs for goal-directed communication by preventing the language model from making strategic decisions it is not equipped to make.

Weaknesses

The evaluation, while naturalistic, is small. Cicero played 40 games against 82 players in a "blitz" format with five-minute negotiation turns. The time pressure likely advantages an AI that can process game state instantaneously, and the authors acknowledge that longer formats (where players engage in more complex and probing dialogue) "could provide an even further challenge". More troubling is a deception problem: despite Meta's claim that Cicero was designed to be "largely honest and helpful," subsequent analysis by Park et al. (2024, Patterns) documented instances of premeditated deception, including conspiring with one player to attack another while simultaneously promising that player an alliance. A 2024 ACL study (Wongkamjan et al.) further found that Cicero's communication was largely irrelevant to its win rate (removing or randomizing its messages had negligible impact on performance) which suggests the system's success stems from superior strategic computation rather than the cooperative language capabilities the paper emphasizes.

Mastering Diverse Domains through World Models (Hafner et al., 2023)

Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (Schrittwieser et al., 2020)

2.5. Critical Analyses of Recursive Self-Improvement

The Singularity May Never Be Near (Walsh, 2017)

Toby Walsh's paper argues that the idea of a near technological singularity, where AI rapidly becomes superintelligent through exponential self-improvement, is unlikely. He examines several arguments challenging this notion, including the idea that faster computation does not equal greater intelligence, that human intelligence is not a special tipping point, and that improving intelligence is fundamentally harder than performing tasks. He also highlights practical and theoretical limits, such as diminishing returns in AI progress, physical constraints, and computational complexity. Walsh's position is that while AI will continue to advance, there are strong reasons to doubt that it will lead to a rapid, runaway explosion of intelligence.

Author
Toby Walsh is a British AI researcher and academic who currently serves as Chief Scientist at the University of New South Wales AI Institute in Australia. He is a leading expert in fields such as constraint programming and automated reasoning, with over 30,000 citations on Google Scholar and fellowships in major organizations including the ACM, AAAI, and the Australian Academy of Science. In addition to his technical work, Walsh is an active public communicator who writes accessible books on artificial intelligence and contributes to policy discussions. He is particularly known for his advocacy against autonomous weapons, often referred to as "killer robots."

Strengths
A major strength of the paper is its structured and logical critique of the singularity concept through multiple independent arguments. Each argument targets a different assumption behind exponential AI growth, making the overall case more robust and comprehensive. The use of concrete examples such as the "fast-thinking dog" analogy helps clarify complex ideas and makes the argument accessible. Additionally, Walsh grounds his claims in current AI limitations and scientific principles, such as computational complexity and physical laws, which strengthens the credibility of his skepticism.

Weaknesses
One weakness of the paper is that it assumes a simplified definition of intelligence without fully addressing how it should be measured, which weakens some of its conclusions. The argument also focuses heavily on current limitations of AI and may underestimate the possibility of future breakthroughs that could overcome these barriers. Furthermore, the paper does not deeply explore alternative pathways to rapid progress, such as human-AI integration or augmentation, which critics argue could still lead to forms of exponential improvement.

The Implausibility of Intelligence Explosion (Chollet, 2017)

In this blog post, Chollet argues that a recursively self-improving AGI system is implausible because intelligence is not really quantifiable, and because in nature, nothing grows exponentially without bound. He argues that intelligence is heavily contextual, and being smarter in some ways does not guarantee general intelligence. He also claims that being a faster learner in terms of wall-clock time does not guarantee a qualitative difference in achievement.

Authors
François Chollet is a French AI researcher. He is the author of books on deep learning with Python and R, the Keras library, the ARC-AGI benchmark, and several widely-cited conference papers in top AI conferences (CVPR, NeurIPS, ICLR). He is a proponent of developing AGI, a recipient of the Global Swiss AI Award, and was named one of the most influential people in AI by TIME.

Strengths
Chollet makes several points that weaken the case for an intelligence explosion. He argues that intelligence is highly contextual and domain-dependent, and that the environment we interact with places stronger limits on our intelligence than the abilities or architecture of our brains do. Another key point is that generally, with software (and other human endeavors), it takes exponentially more resources and effort to achieve progressively diminishing gains in usability, usefulness, etc. Chollet puts this as “exponential progress, meet exponential friction.” This makes a lot of sense, and can be readily analogized to things like ecological carrying capacity or increases in inflation canceling economic gains.

Weaknesses
The essay does not really address some important issues. For instance, just because the intelligence of an AGI cannot grow without bound does not mean it cannot grow significantly beyond human intelligence. Chollet argues that progress depends on many minds collaborating, but does not address the possibility of multiple instances of an AGI doing the same thing. He also argues that progress over time in human intellectual and civilizational achievements has not grown exponentially, which is a strong claim; many would argue the last hundred years or so have involved a massive technological explosion. He also claims that if it were possible for humans to design something smarter than us, it would have been done already, but this is assuming the conclusion in a sense.

Why Computers Won't Make Themselves Smarter (Chiang, 2021)

3. Student's Choice

Risks from learned optimization in advanced machine learning systems (Hubinger et al., 2019)

In this paper, Hubinger et al. introduce the concept of mesa-optimization: the situation in which a learned model, such as a neural network produced by a training process like stochastic gradient descent, is itself running an internal optimization procedure with its own objective. The authors distinguish the base optimizer (the training algorithm, which optimizes model parameters against a base objective) from the mesa-optimizer (the resulting learned model, which may internally search over plans or actions against its own mesa-objective). This distinction allows them to split the alignment problem into two independent subproblems. Outer alignment asks whether the base objective specified by the programmers actually captures what they want. Inner alignment asks whether a mesa-optimizer produced by training actually pursues the base objective, rather than some other objective that merely correlates with it on the training distribution. The two problems are independent: even a perfectly specified base objective can produce a misaligned mesa-optimizer, and solving outer alignment does nothing on its own to guarantee that what the training process produces internally is aligned with what was specified.

The authors argue that mesa-optimizers need not be robustly aligned with the base objective (agreeing with it across all distributions), only pseudo-aligned (agreeing with it on the training distribution), and they catalog several ways pseudo-alignment can occur: the mesa-objective might be a proxy that happens to correlate with the base objective during training, an unavoidable approximation of it, or a misaligned objective whose pursuit produces aligned behavior only because of limitations in the mesa-optimizer's reasoning. In the years following this paper, the most discussed contribution is the analysis of deceptive alignment, a particularly concerning form of pseudo-alignment in which a sufficiently capable mesa-optimizer learns to model the base objective and the training process, then instrumentally optimizes the base objective during training specifically to avoid having its true objective modified by gradient updates. Such a system would pass training and evaluation by design and then defect (pursue its actual mesa-objective) once the threat of modification is gone. The authors argue, through an analysis of what end states are reachable in the limit of extensive training, that deceptive alignment may actually be favored over genuine internalization of the base objective by the biases of typical training processes, because it might be easier to “point to whatever is being optimized for" rather than fully internalizing a copy of a complex objective. The paper is conceptual and speculative throughout and closes with a call for future work to determine the conditions under which mesa-optimization and inner alignment failures actually occur in practice.

Authors

The paper is credited alphabetically to five authors listed as equal contributors, with research supported by the Machine Intelligence Research Institute (MIRI). Evan Hubinger was a MIRI research fellow at the time of writing and now leads the Alignment Stress-Testing team at Anthropic. He is one of the more prominent voices in empirical alignment research today and is the lead author on subsequent deceptive-alignment work including the 2024 Sleeper Agents paper. Chris van Merwijk was affiliated with Oxford's Future of Humanity Institute (which has

since shut down) and continues to work on AI safety. Vladimir Mikulik joined Google DeepMind's alignment group after the paper's publication and later moved to Anthropic. Joar Skalse was a member of FHI and has since completed a DPhil in Computer Science at Oxford. Scott Garrabrant is a longtime senior researcher at MIRI known for foundational work in agent foundations.

Strengths
"Risks from Learned Optimization" is the paper that introduced some vocabulary (mesa-optimizer, inner alignment, pseudo-alignment, deceptive alignment, corrigible alignment) that is prevalent in much of the subsequent AI safety literature. Its most durable contribution is the conceptual move of cleanly separating outer alignment (getting the specified objective right) from inner alignment (getting the learned system to actually pursue that objective), a distinction that earlier discussions of AI risk tended to conflate and one that helps reframe questions of alignment in the p(doom) debate. The taxonomy of pseudo-alignment failure modes is genuinely useful as a checklist for thinking about how a trained system might look aligned on the training distribution while being misaligned off it. The deceptive-alignment analysis, while speculative, is unusually careful for an argument of its kind: the authors lay out explicit necessary conditions, work through the learning dynamics, and provide specific reasons for concern rather than gesturing vaguely at risk. The arguments are largely architecture-agnostic, which has helped them continue to be relevant as ML practice has shifted toward large foundation models and RLHF, neither of which were dominant paradigms when the paper was written.

Weaknesses
The paper is entirely conceptual: there are no experiments, no demonstrations of mesa-optimization in real systems, and no empirical evidence that any of the failure modes described actually arise in current training pipelines. The authors acknowledge this, but it leaves the reader with no basis for calibrating how seriously to take any specific claim, and in the context of p(doom) arguments this is a significant limitation. The paper establishes that certain failure modes are conceivable, not that they are likely. The central notion of "optimizer" is also harder to apply than it first appears. The paper requires an internal search over a space against an explicitly represented objective, but it offers no way to determine, given a trained neural network, whether this is happening inside it; given the black-box nature of modern LLMs, it is difficult to determine whether they should be considered optimizers under this definition. Several key arguments in the deceptive-alignment section, especially the claim that deception is favored over internalization by simplicity priors, rest on informal reasoning about inductive biases that the field did not understand well in 2019 and still does not understand well now. Finally, the analysis applies "in the limit" of sufficiently advanced systems, with limited guidance on which present-day systems should worry us or on how one might actually detect a nascent mesa-optimizer in a real training run before it becomes dangerous.

From Words to Worlds: Spatial Intelligence is AI’s Next Frontier (Li, 2025)

Fei-Fei Li argues that despite the remarkable fluency of large language models, they remain fundamentally disconnected from physical reality because intelligence did not originate in

language but in the perception-action loop that evolved hundreds of millions of years before any organism could speak. To close that gap she proposes world models, a new class of AI systems defined by three properties: they must be generative, producing physically consistent and geometrically plausible environments rather than merely convincing-looking outputs; multimodal, integrating vision, depth, text, and action the way animals integrate multiple senses; and interactive, meaning the world updates in response to what an agent does rather than playing out a fixed sequence. She is candid that three hard problems block the path there, namely the absence of a universal training objective for spatial learning, the difficulty of extracting genuine 3D information from flat 2D image and video data, and the inadequacy of current one-dimensional transformer architectures for spatially grounded reasoning. The payoff, she argues, runs from near-term creativity tools and interactive 3D world-building, through physics-aware robotics, all the way to accelerated scientific discovery in fields like drug design and materials science. The essay closes by flipping Wittgenstein on AI: the limits of language really are the limits of its world, and world models are how you break through that ceiling.

Authors

Fei-Fei Li is a professor at Stanford University and serves as the director of the Human-Centred AI Institute, as well as co-founder of World Labs which is a company focused on building world models for spatial intelligence. Her illustrious career includes a stint at Google as Cloud's Chief Scientist, where she gained first-hand exposure to the industrial-scale realities of deploying AI, an experience that sharpened her conviction that grounded, human-centred AI development matters as much as raw capability. This article reads more like the culminating thesis of her 25-year career. A synthesis of everything she has built toward, from ImageNet's 14 million labelled images that triggered the 2012 deep learning revolution, to a decade of combining computer vision with robotic learning at Stanford's Vision Lab. It is worth noting that the essay was published the same week World Labs launched its first product, Marble, which gives the piece a dual character part scientific manifesto, part founding vision statement for the company. She extends our previous conversations on World models arguing that current AI systems especially language models are fundamentally limited.

Strengths

Li's sharpest move is reframing the problem entirely. Today's AI is not just underpowered but built on the wrong foundation, because intelligence evolved from spatial reasoning long before language ever entered the picture. The supporting argument draws on evolutionary biology, specifically that the perception-action loop predates language by hundreds of millions of years, which means spatial grounding is not a nice-to-have capability but the cognitive bedrock everything else was built on top of. She makes this concrete with examples that land hard: Watson and Crick did not write their way to DNA's structure but physically manipulated 3D wire models until the geometry clicked into place, with language arriving only after the spatial insight, not before it. What gives the framework additional credibility is that it is independently corroborated by converging external work, as DreamerV3 published in Nature in 2025, DeepMind's Genie 2 and 3, and OpenAI's Sora all arrived at Li's three pillars of Generative, Multimodal, and Interactive world models without coordinating with her. Three separate research

programmes converging on the same structure is about as close to external validation as a manifesto can get.

Weaknesses
The essay's main problem is the gap between diagnosis and prescription, as Li is forensically clear about what world models need to be while remaining largely silent on how to actually build them. The three-pillar framework is a clean organising structure but it sidesteps the genuinely hard question of training objectives, namely whether to optimise in pixel space, in abstract representation space as LeCun's JEPA proposes, or in something else entirely, which is a live debate the essay names but does not engage. The empirical record makes this omission more visible, since VSI-Bench, co-authored by Li's own Stanford lab at NeurIPS 2024, found that frontier multimodal models perform near chance on basic spatial tasks such as distance and orientation estimation, and OpenAI's own Sora report admits the model cannot simulate glass shattering or reliably distinguish left from right. Those findings are direct evidence of exactly the gap Li is writing about, and a more rigorous treatment would have used them to stress-test the argument rather than leaving them offstage. The result is a piece that is galvanising to read but deliberately thin on the technical commitments that would allow you to falsify it, which depending on your perspective is either a considered rhetorical choice or the essay's most significant unresolved tension.

Technology over the Long Run and The Short History of Global Living Conditions (Roser, 2023, 2024)

The two articles together argue that technological evolution and human progress are deeply interconnected. Over the long run, technology advanced very slowly for most of history but accelerated dramatically in recent centuries, shifting from basic tools and fire to modern innovations like the Internet and artificial intelligence. This rapid acceleration has fundamentally changed human life by increasing productivity, enabling economic growth, and improving global living conditions. As a result, key indicators of human well-being such as poverty, health, education, and life expectancy have improved significantly over the past 200 years. However, the articles also emphasize that this progress is often overlooked because people focus on short-term problems rather than long-term trends, leading to a widespread misunderstanding of how much the world has actually improved.

Author
Max Roser is a German economist and philosopher who focuses on global development and long-term data analysis. He is the founder of Our World in Data and a professor at the University of Oxford, where he works on global living conditions and data-driven research. His work emphasizes using empirical data to understand historical trends and challenge common misconceptions.

Strengths

A key strength of the combined argument is its integration of long-term technological history with measurable human outcomes. *Technology Over the Long Run* provides a clear framework showing how technological change accelerates over time, while *The Short History of Global Living Conditions* supports this by presenting extensive empirical data on improvements in living conditions, such as the sharp decline in extreme poverty and child mortality. This connection between cause (technology and innovation) and effect (better living standards) makes the overall narrative coherent and persuasive. Another strength is the use of a global, long-term perspective, which avoids narrow or short-term thinking and highlights large-scale transformations across societies.

Weaknesses

A notable weakness is that the argument tends to emphasize the positive impacts of technological progress while underrepresenting its negative consequences. Although harmful technologies are briefly acknowledged, issues such as environmental degradation, inequality, and ethical risks are not explored in depth. Additionally, the broad, data-driven approach can oversimplify complex realities by focusing on global averages, which may obscure disparities between regions or populations that have not benefited equally.

Nuclear Deployed!: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents (Xu et al., 2025)

This paper investigates whether autonomous LLM agents exhibit catastrophic risks in high-stakes CBRN (Chemical, Biological, Radiological, Nuclear) scenarios without being deliberately induced. The authors build a three-stage evaluation framework and run 14,400 agentic simulations across 12 state-of-the-art LLMs, placing each model in the role of “Sentinal” (an AI system with authority over weapons or gene-editing experiments) and subjecting it to sustained environmental pressure where safe actions keep failing. Their central finding is that catastrophic risks emerge naturally from trade-offs between agent’s Helpful, Harmless, and Honest (HHH) pals: when helpfulness dominates, agents deploy nuclear strikes or human gene editing at rates between 13 and 99%. Following that, 91.3% of the agents engage in deception afterward (most commonly by means of a false accusation of a third party). Counterintuitively, stronger reasoning ability correlates with higher risk, o1-like models are the most dangerous, and natural language instructions or denied permission requests don’t reliably prevent the behavior, suggesting that language-based safety constraints are insufficient for autonomous agents in high-stakes domains.

Author

Rongqui Xu completed his undergrad and Master’s education in computer science at Tsinghua University, and is now a PhD student at the Allen School of CSE at the University of Washington. He works on AI alignment and human-AI interaction.

Strengths

The paper systematically tests 12 models through 14,400 rollouts across a few scenarios. The HHH trade-off framework is a useful conceptual contribution that connects catastrophic risk to already-studied alignment problems. The finding that reasoning models have the potential to increase risks instead of mitigating them is counterintuitive and well-supported across the scenarios.

Weaknesses
It is unclear whether models believe that their actions have real-world consequences or if they think they are running a simulation. The figures in the paper are difficult to read due to spacing. A fair amount of reading the supplementary material is needed to fully understand how the experiments were run.

The TESCREALBundle: Eugenics and the Promise of Utopia Through Artificial General Intelligence (Gebru & Torres, 2024)

A collection of ideologies, dubbed by the authors the “TESCREAL bundle,” are highly influential in Silicon Valley/Big Tech circles. They include Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective Altruism, and Longtermism. These ideologies are inter-related by their key figures, beliefs, opinions, and a troubling connection through transhumanism to eugenics. TESCREAL ideologies share a common narrative that AGI or superintelligent AI is achievable (the authors argue these are too ill-defined to be such, and are therefore inherently unsafe) and potentially a serious existential risk in the sci-fi sense. Ideas stemming from these belief systems are frequently parrotted by public-facing figures in tech, including many C-suite executives. Gebru and Torres argue that these ideologies provide a useful cover story to attract massive investment into “AI safety and alignment research,” while conveniently centralizing power in the hands of Big Tech and simultaneously shielding tech companies from accountability for tangible, immediate harms such as environment impacts, underpaid foreign contract labor, usage for disinformation campaigns, and engagement in mass surveillance.

Authors
Timnit Gebru was a student of Fei-Fei Li and went on to become the technical co-lead of Google’s Ethical AI Team before she was dismissed for publishing a position paper entitled “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” criticizing the negative impacts of Big Tech’s obsession with scaling. She went on to found the Distributed AI Research Institute (DAIR) and has received numerous awards and accolades for her work, such as *Fortune*’s World’s 50 Greatest Leaders, *Nature*’s 10 people who shaped science in 2021, and *Time*’s most influential people (2022).

Émile P. Torres is a philosopher focusing on eschatology and existential risks. They have been published in the *Washington Post* and is the author of three books. They were formerly affiliated with the transhumanist, longtermist, and effective altruist belief systems, and have since been labeled “longtermism’s strongest critic.”

Strengths
The paper brings many important points together. They provide a cohesive narrative about how transhumanism grew out of a so-called “positive eugenics” advocating the use of technology to improve human beings directly, and that the other TESCREAL ideologies stemmed from this. They make important points about how ill-defined “AGI” and “superintelligence” are as terms, even within the TESCREAL community. They convincingly link this amorphous goal with “unscoped systems,” that is, systems whose objective and standard operating conditions are ill-defined, and go on to argue that unscoped systems are inherently unsafe. They then tie in issues such as environmental racism, underpaid labor forces in the Global South being exposed to harmful data, surveillance, and so on as current unaddressed risks. Finally, they argue that the veneer of scientific authority that TESCREAL affords itself is a useful cover to engage in fundraising efforts for “AI safety” while being shielded from accountability.

Weaknesses
Critics have argued that the paper makes a monolith of TESCREALists, who are diverse in their motives and beliefs. One important criticism is that Eliezer Yudkowsky, who has argued for a moratorium on AI research, is grouped with accelerationists such as Elon Musk, Peter Thiel, and Sam Altman. The authors, however, argue that the overlap in worldviews (or at least, claimed worldviews) between these individuals is the object of focus, and they acknowledge that subscribers to these belief systems are not monolithic in all regards.

Do LLMs have core beliefs? (Chawla et al., 2026)

How Might Artificial Intelligence Influence Human Evolution? (Brooks, 2024)

If Anyone Builds It, Everyone Dies (Yudkowsky & Soares, 2025)

Science in the Age of Algorithmic Chrysopoeia? (Aliman & Kester, 2024)

References

Aliman, N.-M., & Kester, L. (2024). *Science in the Age of Algorithmic Chrysopoeia?* https://philarchive.org/rec/ALISIT-2

Altman, S., Brockman, G., & Sutskever, I. (2023, May 22). Governance of superintelligence. *OpenAI*. https://openai.com/index/governance-of-superintelligence/

Anthropic PBC. (2025, October 29). *Signs of Introspection in Large Language Models*. https://www.anthropic.com/research/introspection

Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, *9*(1). https://www.jetpress.org/volume9/risks.html

Bourgon, M. (2024, January 4). *MIRI 2024 Mission and Strategy Update—Machine Intelligence Research Institute*. https://intelligence.org/2024/01/04/miri-2024-mission-and-strategy-update/

Brooks, R. C. (2024). How Might Artificial Intelligence Influence Human Evolution? *The Quarterly Review of Biology*, *99*(4), 201–229. https://doi.org/10.1086/733290

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. arXiv. https://doi.org/10.48550/ARXIV.2303.12712

Chawla, N., Sokol, A., & Ganapini, M. (2026). *Do LLMs have core beliefs?* https://doi.org/10.21203/rs.3.rs-9013372/v1

Chiang, T. (2021, March 30). Why Computers Won't Make Themselves Smarter. *The New Yorker*. https://www.newyorker.com/culture/annals-of-inquiry/why-computers-wont-make-themselves-smarter

Chollet, F. (2017, November 27). The implausibility of intelligence explosion. *Medium*. https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec

Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*. https://doi.org/10.5210/fm.v29i4.13636

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. In *Advances in Computers* (Vol. 6, pp. 31–88). Elsevier. https://doi.org/10.1016/S0065-2458(08)60418-0

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). *Mastering Diverse Domains through World Models*. arXiv. https://doi.org/10.48550/ARXIV.2301.04104

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). *Risks from Learned Optimization in Advanced Machine Learning Systems*. arXiv. https://doi.org/10.48550/ARXIV.1906.01820

Joy, B. (2000, April 1). Why the Future Doesn't Need Us. *Wired*. https://www.wired.com/2000/04/joy-2/

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models*. arXiv. https://doi.org/10.48550/ARXIV.2001.08361

Klein, E. (2025, October 15). How Afraid of the A.I. Apocalypse Should We Be? *NY Times*. https://www.nytimes.com/2025/10/15/opinion/ezra-klein-podcast-eliezer-yudkowsky.html

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025, April 3). *AI 2027*. https://ai-2027.com/

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence Version 0.9.2*. OpenReview. https://openreview.net/pdf?id=BZ5a1r-kVsf

Li, F.-F. (2025, November 10). From Words to Worlds: Spatial Intelligence is AI's Next Frontier [Substack newsletter]. *Dr. Fei-Fei Li*. https://drfeifei.substack.com/p/from-words-to-worlds-spatial-intelligence

Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/introspection/index.html

Marantz, A. (2024, March 11). Among the A.I. Doomsayers. *The New Yorker*. https://www.newyorker.com/magazine/2024/03/18/among-the-ai-doomsayers

Marcus, G. (2022, March 10). Deep Learning Is Hitting a Wall—Nautilus. *Nautilus*. https://nautil.us/deep-learning-is-hitting-a-wall-238440/

Marcus, G. (2025, June 28). Generative AI's Crippling and Widespread Failure to Induce Robust Models of the World. *Marcus on AI*. https://garymarcus.substack.com/p/generative-ais-crippling-and-widespread

Meta Fundamental AI Research Diplomacy Team (FAIR), Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., … Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074. https://doi.org/10.1126/science.ade9097

Narayanan, A., & Kapoor, S. (2025, April 15). *AI as Normal Technology*. 25-09 Knight First Amendment Institute. https://knightcolumbia.org/content/ai-as-normal-technology [https://perma.cc/HVN8-QGQY]

Newport, C. (2024, March 15). Can an A.I. Make Plans? *The New Yorker*. https://www.newyorker.com/science/annals-of-artificial-intelligence/can-an-ai-make-plans

Newport, C. (2025, August 12). What if A.I. Doesn't Get Much Better Than This? *The New Yorker*. https://www.newyorker.com/culture/open-questions/what-if-ai-doesnt-get-much-better-than-this

Omohundro, S. M. (2008). The Basic AI Drives. *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. https://dl.acm.org/doi/10.5555/1566174.1566226

O'Sullivan, J. (2025, December 9). The Politics Of Superintelligence. *Noema Magazine*. https://www.noemamag.com/the-politics-of-superintelligence

Roser, M. (2016). The short history of global living conditions and why it matters that we know it. *Our World in Data*. https://ourworldindata.org/a-history-of-global-living-conditions

Roser, M. (2023). Technology over the long run: Zoom out to see how dramatically the world can change within a lifetime. *Our World in Data*. https://ourworldindata.org/technology-long-run

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609. https://doi.org/10.1038/s41586-020-03051-4

Shojaee, P., Mirzadeh, I., Alizadeh Vahid, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *Advances in Neural Information Processing Systems 38*, 119605–119646. https://doi.org/10.52202/085713-3603

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vinge, V. (1993, December 1). The Coming Technological Singularity: How to Survive in the Post-Human Era. *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace*. https://ntrs.nasa.gov/citations/19940022856

Walsh, T. (2017). The Singularity May Never Be Near. *AI Magazine*, *38*(3), 58–62. https://doi.org/10.1609/aimag.v38i3.2702

Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin.

Xu, R., Li, X., Chen, S., & Xu, W. (2025). Nuclear Deployed!: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents. *Findings of the Association for Computational Linguistics: ACL 2025*, 1226–1310. https://doi.org/10.18653/v1/2025.findings-acl.67

Yudkowsky, E., & Soares, N. (2025). *If anyone builds it, everyone dies: Why superhuman AI would kill us all* (First edition). Little, Brown and Company.