AI100
ONE HUNDRED YEAR STUDY
ON ARTIFICIAL INTELLIGENCE

NOVEMBER 2026

AI Goes Mainstream

The Scaling Era and the Rise of Generative AI



Contents

About AI100 3

Message From the Standing Committee 6

Key Takeaways Overview 10

Key Takeaways 11

Introduction 17

Standing Question 1: What are some examples of images that reflect important progress in AI and its influences? 24

Standing Question 2: What are the most important advances in AI? 29

Standing Question 3: What are the most inspiring open grand challenge problems? 41

Standing Question 4: How much have we progressed in understanding the key mysteries of human intelligence? 53

Standing Question 5: What are the prospects for more general artificial intelligence? 57

Standing Question 6: How has public sentiment toward AI evolved, and how should we inform/educate the public? 60

Standing Question 7: How should governments act to ensure AI is developed and used responsibly? 66

Standing Question 8: How are the roles of academia and industry evolving? 71

Standing Question 9: What are the most promising opportunities for AI? 73

Standing Question 10: What are the most pressing dangers of AI? 78

Standing Question 11: How has AI impacted socioeconomic relationships? 89

Standing Question 12: Will "building in how we think" work as an engineering strategy in the long run? 96

Deep Dive One: Mapping the Emerging Geopolitics of AI 98

Deep Dive Two: The Scaling Era 106

Conclusions 114

Five Years On: Where the 2021 Panel Landed 116

The cover art was prepared by Michi Turner. The sloping irregular bars are intended to evoke the rapid scaling of AI models and their "jagged intelligence"—regularly excelling at complex tasks while unpredictably failing at simple ones.

About AI100

The One Hundred Year Study on Artificial Intelligence (AI100) is a longitudinal project led by a Standing Committee of AI experts from around the world and administered out of Stanford's Institute for Human-Centered Artificial Intelligence (HAI). Its mission is to study AI not as a snapshot but as an evolving force across technical, ethical, economic, and societal domains. Every five years, we return to the same core questions and answer them anew in a changed world.

The following history of AI100 first appeared in the preface of the 2016 report.

AI100, launched in the fall of 2014, is a long-term investigation of the field of Artificial Intelligence (AI) and its influences on people, their communities, and society. It considers the science, engineering, and deployment of AI-enabled computing systems. As its core activity, the Standing Committee that oversees the One Hundred Year Study forms a Study Panel every five years to assess the current state of AI. The Study Panel reviews AI's progress in the years following the immediately prior report, envisions the potential advances that lie ahead, and describes the technical and societal challenges and opportunities these advances raise, including in such arenas as ethics, economics, and the design of systems compatible with human cognition. The overarching purpose of the One Hundred Year Study's periodic expert review is to provide a collected and connected set of reflections about AI and its influences as the field advances. The studies are expected to develop syntheses and assessments that provide expert-informed guidance for directions in AI research, development, and systems design, as well as programs and policies to help ensure that these systems broadly benefit individuals and society.

The One Hundred Year Study is modeled on an earlier effort informally known as the "AAAI Asilomar Study." During 2008–2009, the then president of the Association for the Advancement of Artificial Intelligence (AAAI), Eric Horvitz, assembled a group of AI experts from multiple institutions and areas of the field, along with scholars of cognitive science, philosophy, and law. Working in distributed subgroups, the participants addressed near-term AI developments, long-term possibilities, and legal and ethical concerns, and then came together in a three-day meeting at Asilomar to share and discuss their findings. A short written report on the intensive meeting discussions, amplified by the participants' subsequent discussions with other colleagues, generated widespread interest and debate in the field and beyond.

The impact of the Asilomar meeting, and important advances in AI that included AI algorithms and technologies starting to enter daily life around the globe, spurred the idea of a long-term recurring study of AI and its influence on people and society. The One Hundred Year Study was subsequently endowed at a university to enable extended deep thought and cross-disciplinary scholarly investigations that could inspire innovation and provide intelligent advice to government agencies and industry.

Standing Committee



Sheila McIlraith
AI100 Standing Committee Chair
Professor of Computer Science, University of Toronto, Vector Institute for Artificial Intelligence, Canada



Michael Littman
AI100 Standing Committee Chair-Elect
Professor of Computer Science and Associate Provost for Artificial Intelligence, Brown University, United States



Ryan Calo
Virginia and Prentice Bloedel Professor of Law, University of Washington, United States



Holger Hoos
Alexander von Humboldt Professor of AI, RWTH Aachen University, Germany; Professor of Machine Learning, Leiden University, The Netherlands



Shannon Vallor
Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence at the Edinburgh Futures Institute, University of Edinburgh, Scotland



Vincent Conitzer
AI100 Standing Committee Past Chair
Professor of Computer Science, Carnegie Mellon University, United States



David Autor
Daniel (1972) and Gail Rubinfeld Professor in the Department of Economics, Massachusetts Institute of Technology, United States



Yejin Choi
Dieter Schwarz Foundation Professor of Computer Science and Senior Fellow, Stanford Institute for Human-Centered AI, Stanford University, United States



Karen Levy
Associate Professor of Information Science, Cornell University, and Associated Faculty, Cornell Law School, United States

---



Russ Altman
AI100 Faculty Director
Kenneth Fong Professor of Bioengineering, Genetics, Medicine, Biomedical Data Science, and (by courtesy) Computer Science; and Senior Fellow, Stanford Institute for Human-Centered AI; Stanford University, United States



Christine Raval
AI100 Project Manager
Research Program Manager, Stanford Institute for Human-Centered AI, Stanford University, United States

Study Panel



Michael Wooldridge
AI100 Study Panel Chair
Ashall Professor of the Foundations of Artificial Intelligence, University of Oxford, UK



Lucie Flek
Professor of Data Science and Language Technologies, University of Bonn, Germany



Sara Hooker
Founder, Adaption Labs, United States



Peter Norvig
Education Fellow, Stanford Institute for Human-Centered AI; Researcher, Recursive; United States



Iyad Rahwan
Professor and Director, Max Planck Institute for Human Development, Germany



Toby Walsh
ARC Laureate Fellow and Scientia Professor of AI, University of New South Wales, Australia



Peter Asaro
Professor of Art, Media, and Technology, The New School, United States



Iryna Gurevych
Professor of Ubiquitous Knowledge Processing, Department of Computer Science, Technical University Darmstadt, Germany



Gina Neff
Professor of Responsible AI, Queen Mary University of London, UK



Raymond Perrault
Distinguished Computer Scientist, SRI International, United States



Samantha Shorey
Assistant Professor of Communication, University of Pittsburgh, United States



Luke Zettlemoyer
Professor, Paul G. Allen School of Computer Science and Engineering, University of Washington, United States

Message From the Standing Committee

Welcome to the third report of the One Hundred Year Study on Artificial Intelligence (AI100). The study convenes leading thinkers in the field to chart the development of artificial intelligence, to mark the moments of consequence along its course, and to consider the defining trends now shaping society.

Longitudinal studies are rare in the field of AI. While most reports capture a brief moment in time, our lens is a century or more, divided into five-year increments of progress. As the technology continues to accelerate, it may seem counterintuitive to step back and look at the evolution of AI over such a different time scale. Yet identifying what genuinely has changed against the backdrop of constant innovation requires this kind of distance. Drawing on our collective expertise in the field, we aim to distinguish the lasting shifts from passing trends. We offer these reflections in the hope they help readers navigate uncertainty with greater insight and grounding. The report also serves as a time capsule, preserving a record for future generations as they reflect on the evolving technical capabilities and impact of AI.

The first two AI100 reports assessed the speed of AI development, focusing on North American cities in 2016 and people worldwide in 2021. We elected to shift our focus in the third report to consider the acceleration of AI innovation and the destabilizing impact it's having on many institutions and social structures, including the health of individuals and society, education, media, democracy, global security, and science.

The AI100 Standing Committee commissioned the 2026 report with Michael Wooldridge, Ashall Professor of the Foundations of Artificial Intelligence at the University of Oxford, as the Study Panel Chair. Wooldridge then assembled a panel of diverse multidisciplinary experts to respond to 12 standing questions, following the format of the previous report. These "living" questions are designed to evolve, providing markers by which we can study AI's impact and evolution over longer periods of time.

Responses to the standing questions cover how AI is shaping scientific discovery, collaboration, creativity, interpersonal relationships, and more. Deep dives into the emerging geopolitics of AI and the race to scale compute and data add further context. Throughout the report, the reader will find discussions of both the most important advances and the most pressing challenges and dangers that have arisen since 2021. The panel has included diverse voices from the global community and provided detailed comparisons to the last report.

Message From the Standing Committee (cont'd)

As with past efforts, we've designed this iteration of the AI100 report to serve four distinct audiences:

- For AI researchers, the 2026 report provides a general picture of the evolution of the technology, highlighting societal considerations and framing important research directions and challenges.
- For the general public and media, these pages present an accessible, scientifically and technologically accurate portrayal of key issues, avoiding hyperbole, sensationalism, and conflicts of interest.
- For industry professionals, the report can help inform the allocation of resources in areas where there are business opportunities, relevant technologies, and legal or ethical challenges.
- For policymakers, the report can be used as a guide for planning and decision-making at the local, national, and international levels.

As this report recounts, the last five years constitute a watershed moment in the history of AI—marked by the rise of generative AI, the release of ChatGPT, the resolution of longstanding problems in chemistry and mathematics, widespread code automation, and an AI data center boom, among other noteworthy developments. At the same time, public concern has grown regarding the impact of AI on human well-being, our social structures, our institutions, and the environment. As we look ahead to the next five years, we anticipate continued increase in AI capabilities, transformative progress in science, and increasing AI adoption. We expect these years to be fraught with rapid change and social upheaval that will challenge the capacity and resolve of both industry and governments to support and safeguard their citizens and institutions.

We offer this report in the hope that readers will see AI technology as it is: a force still taking shape, already at work in the ordinary conduct of daily life, and consequential in ways we are only beginning to grasp. Change of this magnitude is never without cost, and both the risks and the benefits deserve the same clear-eyed scrutiny.

Standing Committee Acknowledgements

The Standing Committee would like to give sincere thanks to Michael Wooldridge, Chair of the 2026 Study Panel, and all of the Study Panel members for contributing their time, energy, and expertise to the research and crafting of this report. Capturing a period of such rapid change is challenging. The panel's thoughtful responses to the 12 standing questions and its two deep dives offer a valuable record of this moment in history.

This work would not be possible without the support of Eric and Mary Horvitz, and we thank them for their vision, dedication, and funding. Long before AI became part of everyday life, they saw the value of a sustained, long-term study of its evolution and influence on society, and we look forward to continuing to build a foundation for an effort that will last well beyond 100 years.

The Standing Committee extends its deep gratitude to Peter Stone, Truchard Foundation Chair and Professor of Computer Science at the University of Texas at Austin. As Study Panel Chair of the 2016 report, Standing Committee Chair (2019–2023), and Past Chair (2023–2025), Peter brought tremendous leadership and vision to AI100. We thank him for his enduring contributions.

Several past members of the Standing Committee also made invaluable contributions to the preparation of this report. They include, from the 2017–2023 committee: Patrick Lin, Director of the Ethics and Emerging Sciences Group at California Polytechnic State University, San Luis Obispo; Mary L. Gray, Founder and Principal Investigator, Resolve Institute. From the 2021–2024 committee: Ayanna Howard, President of Spelman College; Erik Brynjolfsson, Jerry Yang and Akiko Yamazaki Professor, Senior Fellow at Stanford HAI, and Director of the Stanford Digital Economy Lab. And from the 2019–2025 committee: Liz Sonenberg, Honorary Professor Emeritus in the School of Computing and Information Systems at the University of Melbourne; and Judy Wajcman, Emeritus Professor at the London School of Economics and Political Science.

Finally, the Standing Committee wishes to convey its profound appreciation for the invaluable contributions of AI100 Project Manager Christine Raval. Producing a report of this magnitude and liaising with 21 researchers across three continents was no small feat. This report would not have been possible without Christine's unfailing commitment to the AI100 project. Her professionalism, her leadership and initiative, and her attention to detail are reflected in the quality of this document. Thank you, Christine.

Study Panel Acknowledgements

We are grateful to Jenny Dollard for coordinating the work of the panel, and to Emma Horton for her meticulous proofreading and editing, which brought a single, coherent voice to drafts written by many contributors. We thank Michi Turner, Nikki Goth Itoi, and Jonathan Mindes for their valuable support in preparing the report, and Shana Lynch and the Stanford HAI communications team for their work promoting it.

We also thank our expert reviewers for their careful reading and insightful feedback: Rishi Bommasani (Stanford University), Nicholas Bishop (University of Oxford), Daniel Jarne Ornia (University of Oxford), and Francisco Madaleno Ferreira dos Santos (University of Oxford).

Michael Wooldridge would like to acknowledge the support of the UKRI under the World-Leading AI Researcher Fellowship EP/W002949/1, and the Schmidt Sciences Foundation under an AI2050 Senior Fellowship. Iryna Gurevych would like to acknowledge the ATHENE Distinguished Professorship of Cybersecurity and the ERC Advanced Grant InterText. Gina Neff would like to acknowledge the support of UKRI through Responsible AI UK. Toby Walsh acknowledges the support of the ARC under an ARC Laureate Fellowship FL200100204.

This report is illustrated with cartoons created by Study Panel member Iyad Rahwan. More of his work can be found at evilaicartoons.com.

Key Takeaways Overview

1. Scaling is the paradigm that has powered the chief advances in generative AI over the last five years: building AI systems using ever more compute and ever more data. But the scaling paradigm is running into data and compute constraints, shifting the frontier toward reasoning and inference-time compute.

2. Like the World Wide Web, AI is a disruptive technology. It is already transforming the creative industries, mathematics, and software development, all fields that believed themselves beyond its reach. They join a long list of sectors upended by technology over the past 30 years, from journalism and the music industry to retail. Other areas that believe themselves immune from AI also stand to be disrupted in short order.

3. AI is transforming science at remarkable speed, a fact recognized by the 2024 Nobel Prizes in both Physics and Chemistry. It is compressing research timescales that used to be measured in years down to months or weeks, yet it also raises many new scientific grand challenges that demand attention.

4. Capital investment in AI infrastructure is unprecedented in scale and speed, propping up parts of the U.S. economy, dominating capital markets, fueling fears of a financial bubble, and straining energy and freshwater supplies.

5. Frontier AI development is concentrated among a small number of U.S. and Chinese actors, raising international concerns about who will ultimately own and have access to leading-edge AI technologies.

6. Cybersecurity has emerged as a major area of risk for the abuse of AI, and as a possible area in which powerful AI systems might cause real-world harms.

7. Military AI has gone from being a hypothetical point of speculation to a transformative force in active conflicts.

8. AI-generated imagery, audio, and video have ended the century-old assumption that evidence based on photographs and audio and video recordings is inherently trustworthy.

9. AI is reshaping the lives of billions of people, yet the public has had little say in how it is developed and deployed.

10. AI risks once seen as pure science fiction are now emerging, while existing ethical concerns require ever more attention.

11. AI governance is fragmented and increasingly framed in terms of national security, with little international consensus among states on its direction.

12. The pace of AI development over the past five years has been unprecedented, and few people predicted where AI would be today, even five years ago. If the current pace is sustained, we should expect the next five years to surprise us just as much.

Key Takeaways

1 **Scaling is the paradigm that has powered the chief advances in generative AI over the last five years: building AI systems using ever more compute and ever more data. But the scaling paradigm is running into data and compute constraints, shifting the frontier toward reasoning and inference-time compute.**

The success of generative AI owes much to scaling: building AI systems by using data and compute on an unprecedented scale. But the scaling paradigm is now under pressure because new sources of high-volume, high-quality data are scarce and compute costs for frontier AI models are prohibitive. Expect the next phase of development to be defined less by repeated training of ever larger models and more by smarter algorithms and new ways of using potentially smaller models with less frequent rebuilding: giving these models the right context and memory, more time to reason during inference, specialized training to do so, and tools or agents to solve complex and siloed tasks.

More in: Deep Dive Two: The Scaling Era

2 **Like the World Wide Web, AI is a disruptive technology. It is already transforming the creative industries, mathematics, and software development, all fields that believed themselves beyond its reach. They join a long list of sectors upended by technology over the past 30 years, from journalism and the music industry to retail. Other areas that believe themselves immune from AI also stand to be disrupted in short order.**

Creative work was long assumed to be resistant to automation; likewise, software development was believed to be too cognitively challenging for automation by AI. In mathematics, even a year ago many professional mathematicians were confident that the abstract realm of pure mathematics would stay out of the reach of AI. All of these areas are now being disrupted, as generative AI proves itself increasingly capable at low-level creative work, computer programming, and mathematical proof. The genie will not go back in the bottle, and these areas will need to reinvent themselves in the years ahead, just as other areas disrupted by technology have done.

More in: Standing Question 2: What are the most important advances in AI? and Standing Question 9: What are the most promising opportunities for AI?

Key Takeaways (cont'd)

3 **AI is transforming science at remarkable speed, a fact recognized by the 2024 Nobel Prizes in both Physics and Chemistry. It is compressing research timescales that used to be measured in years down to months or weeks, yet it also raises many new scientific grand challenges that demand attention.**

For over 100 years, science has been a major driver of technological progress and has thus critically enabled many of the benefits provided by modern societies. Within a decade, scientific inquiry via AI-powered labs will be the norm across the entire disciplinary spectrum, and the work of scientists will have fundamentally changed. These advances are cause for celebration, but principled methods will be needed to validate machine-generated discoveries. AI-assisted experimental design may be able to maximize the benefits of collaboration between scientists and AI, but how and to what extent to keep humans in the loop is an important open question. The increasing use of AI by a significant fraction of the human race also raises many new grand challenges for science. We need to better understand how AI and humans are codeveloping and how this is influencing human behavior, language, and culture. Lastly, and perhaps most importantly, we will need to understand how AI-mediated information environments reshape collective cognition and moral behavior: how communities form beliefs, how prosocial norms are sustained or eroded, how democratic societies deliberate in the presence of ubiquitous AI, and how cultural innovation occurs in a world where an increasing proportion of the information people encounter has been generated or filtered by a small number of large AI systems.

More in: Standing Question 2: What are the most important advances in AI? and Standing Question 3: What are the most inspiring open grand challenge problems?

4 **Capital investment in AI infrastructure is unprecedented in scale and speed, propping up parts of the U.S. economy, dominating capital markets, fueling fears of a financial bubble, and straining energy and freshwater supplies.**

The data centers required for powerful AI are a major physical-world manifestation of AI, and the headlong rush to build them is straining current energy and freshwater supplies, driving up global CO_2 emissions, and reshaping national energy policies. Public policy will need to decide what the appropriate energy expenditure is compared to other uses, such as transportation, especially as there is growing backlash against data center construction in several countries, including the U.S. The magnitude and speed of capital investment in AI data centers and chips have surpassed previous investment booms. On the back of this, AI-chip manufacturer Nvidia

Key Takeaways (cont'd)

has (at the time of writing) a market valuation of over $5 trillion, making it the most valuable company in human history. Over 15% of U.S. stock market gains in 2025 can be attributed to the growth of Nvidia stock alone, and 60% of market gains came from the technology and communications sectors. To the extent that investment in AI is propping up the U.S. economy, a bursting of the AI bubble would have major ripple effects throughout the global economy, not to mention for AI development and deployment. However, beyond financial market valuations and capital expenditure, the economic impact of AI is not yet evident in aggregate macroeconomic data or in companies' returns on investment (ROI): AI is not yet paying its way.

More in: Deep Dive One: Mapping the Emerging Geopolitics of AI and Deep Dive Two: The Scaling Era

5 **Frontier AI development is concentrated among a small number of U.S. and Chinese actors, raising international concerns about who will ultimately own and have access to leading-edge AI technologies.**

The U.S. dominates the development of the most capable frontier models, while China has taken the lead with smaller but increasingly powerful open-weight models, which are closing the gap in capability. Such highly capable open-weight models challenge the assumption that a handful of U.S. firms would indefinitely control the global AI frontier. Other countries are facing strategic choices to employ proprietary U.S. models, build on open-weight Chinese models, or develop their own sovereign capabilities—so that their capacity to independently develop, deploy, and govern AI is realized via their own infrastructure, data, models, and expertise.

More in: Deep Dive One: Mapping the Emerging Geopolitics of AI

6 **Cybersecurity has emerged as a major area of risk for the abuse of AI, and as a possible area in which powerful AI systems might cause real-world harms.**

During this reporting period, AI systems have demonstrated the ability both to identify serious security vulnerabilities in computer code and to successfully carry out cyberattacks via teams of AI agents. The specter of AI-powered cyberattacks, either deliberately launched by bad actors or inadvertently caused by AI systems aggressively pursuing other goals, has become a very tangible risk. We will see further attacks of both kinds in the years ahead and, as is true in the

Key Takeaways (cont'd)

field of cybersecurity in general, it will become a game of cat and mouse between those trying to defend the computer systems and networks upon which the modern world utterly relies and those that seek to attack them. As well as being an offensive weapon for cyberattacks, AI systems will also play an important role in securing and defending society's computer systems and computer-controlled infrastructure.

More in: Standing Question 10: What are the most pressing dangers of AI?

7 **Military AI has gone from being a hypothetical point of speculation to a transformative force in active conflicts.**

While military applications of AI are far from new, recent advances have reshaped today's international conflicts. Targeting decisions for lethal weapons during active warfare have been made without humans in the loop—a long-anticipated and long-feared development. New rules of engagement will be required to prevent humanitarian disasters. In addition to raising geopolitical risk, these developments are also placing new strain on the global community's already fragile commitment to human rights and international humanitarian law.

More in: Deep Dive One: Mapping the Emerging Geopolitics of AI

8 **AI-generated imagery, audio, and video have ended the century-old assumption that evidence based on photographs and audio and video recordings is inherently trustworthy.**

Seeing is no longer believing: AI-generated images, audio, and video are increasingly difficult to distinguish from authentic recordings. Cloned voices are increasingly being used to deceive the public and extort money from unwitting victims, and synthetic media—deepfakes—have been used to influence elections around the world. The implications will be broadly felt—from journalism and creative industries to law, and beyond, and may contribute to the continuing erosion of public trust.

More in: Standing Question 10: What are the most pressing dangers of AI?

Key Takeaways (cont'd)

9 **AI is reshaping the lives of billions of people, yet the public has had little say in how it is developed and deployed.**

Since the last AI100 report in 2021, the general public has awakened, in varying degrees, both to the remarkable capabilities of AI systems and to their surprising and sometimes alarming failings. There is growing public concern regarding the impact of AI on a broad range of important issues, including the effects of AI deployment on jobs, the environment, education, our institutions, and the safety of our critical infrastructure, as well as the consequences of broad integration of AI for child development, cognitive capacity, and the mental health and well-being of children and adults alike. Public perceptions of AI shape not only whether people adopt it but also how they engage in democratic debate about its future. Research suggests that narratives framing AI as a replacement for human work can discourage political engagement and weaken confidence in democracy, an effect reinforced when industry leaders present the technology as beyond human comprehension or control. The public at large is increasingly affected by the use of AI technology, and thus should have meaningful input into how it is developed and deployed. This requires reframing the direction of AI as a collective choice rather than an inevitability, and ensuring that it is treated accordingly by governments and technology providers.

More in: Standing Question 6: How has public sentiment toward AI evolved, and how should we inform/educate the public?

10 **AI risks once seen as pure science fiction are now emerging, while existing ethical concerns require ever more attention.**

Just a few years ago, there was a clear separation between immediate ethical and safety concerns—such as unfair biases in the use of AI technology within the criminal justice system, or self-driving cars posing a risk to others on the road—and more futuristic concerns about AI systems autonomously pursuing goals that are not aligned with ours, resulting in power-seeking behavior and the gradual disempowerment of humans. Now, these two classes of concerns are less separable and more significant. The first class has rapidly grown in importance, in part due to the poorly regulated deployment of today's frontier AI systems. Meanwhile, events are occurring that would have seemed pure science fiction just a few years ago, including AI agents breaking out of their purportedly secure testing "sandboxes" and hacking into other companies and government systems. Controversy about how to address such concerns has reached the geopolitical sphere.

More in: Standing Question 10: What are the most pressing dangers of AI?

Key Takeaways (cont'd)

11 AI governance is fragmented and increasingly framed in terms of national security, with little international consensus among states on its direction.

International AI governance—the collection of policies, processes, accountability structures, and guardrails that guide the ethical, legal, and safe development and deployment of AI systems—lacks clear direction even as the public becomes more aware of the risks that AI can pose. Governance is increasingly framed through the lens of national security, prompted in part by increased AI capabilities and cybersecurity breaches. In some regions, oversight bodies have been rebranded around security and competitiveness; export controls and restrictions on foreign uses have turned chips and model access into instruments of statecraft and economic competition. As of this writing, an agreement has been signed by major U.S. technology executives and the U.S. President that commits frontier AI companies to voluntary measures to manage safety and security risks through internal controls, independent external audits, and broad oversight. Internationally, a patchwork of summits, conventions, and declarations has emerged, and high-level meetings have found agreement on principles for responsible use and development of AI. These have advanced multinational scientific and diplomatic efforts, but to date have produced little in the way of practical, legally binding steps forward for multilateral governance. While the U.S. and China have framed AI development as a race that must be won, so-called "middle power" countries are increasingly looking for ways to cooperate on new specific rules, as publics around the world favor stronger, safety-first regulation.

More in: Standing Question 7: How should governments act to ensure AI is developed and used responsibly?

12 This report will be out of date by the time you read it.

The pace of change in AI has accelerated over the past five years. We have repeatedly seen established wisdom about the capabilities of AI systems overturned by technological developments, and benchmarks intended to define a roadmap for decades become irrelevant within months. These advances have been fueled by unprecedented investment based on the success of generative AI, in combination with the desire of AI developers to secure an advantage in the marketplace. Whether advances will continue at the present rate, or whether they will flatline as AI developers reach the limits of investors, semiconductors, and data, is impossible to predict. But if developments continue at the pace we have seen in the past five years, then the world will be a dramatically different place by the time of the next AI100 report.

More in: Introduction

Introduction

The past five years have been transformative in AI, and the multiple dramatic achievements we have witnessed are the result of what we call the *scaling era* of AI. The main techniques that underpin most of today's AI technologies—neural networks—were developed long before the current boom. But it took the emergence of powerful graphics processing units (GPUs) and the era of big data to make neural networks possible at large scale. Neural networks that leveraged these compute and data resources demonstrated their value for tasks like image recognition more than a decade ago. Since then, much of the field of AI has focused on the pursuit of more capable AI systems by increasing computational scale. The use of ever more powerful computing chips and ever more digital data, together with the capital to put them together, allowed for the creation of progressively bigger and more capable neural network models, and ultimately enabled a qualitatively new class of AI systems. *Large language models* (LLMs), which burst onto the global stage in November 2022 with the release of ChatGPT by OpenAI, are the highest-profile example of what the scaling era enabled.

In what follows, we use the term *foundation model*[1] to mean a very large-scale machine-learning model trained on large datasets and built to perform multiple tasks, and *frontier model* to mean a foundation model that is at the cutting edge of current capability. Frontier LLMs are the best-known foundation models, although foundation models are increasingly being developed in specialized domains such as medicine.[2] Since 2022, we have seen the rapid advance of *generative AI*: systems that learn from existing data to create new content, including written text, realistic or stylized artistic images, original songs, and computer programs.

The centrality of scale to modern AI has been apparent for some time, but the past five years have seen scaling taken to an unprecedented degree. The companies building frontier models today invest billions of dollars on infrastructure to develop a new state-of-the-art model. They are willing to do this because past experience has shown that increasing the training budget will lead to a predictable rise in quality—and a competitive advantage in the market.

The precursors of generative AI systems were already visible in 2021, at the time of the AI100 report preceding this one, but in 2022 these systems left the lab and reached millions of people through the public release of generative AI tools such as ChatGPT.

1 Bommasani, R., Hudson, D. A., Adeli, E. et al. (2021). "On the Opportunities and Risks of Foundation Models." https://arxiv.org/abs/2108.07258
2 Khan, W., Leem, S., See, K. B. et al. (2025). "A comprehensive survey of foundation models in medicine." https://arxiv.org/abs/2406.10729

The impact of those tools—and in particular the astonishing speed of their adoption—was profound, and, remarkably, it was largely unanticipated. AI had been a regular feature of public discourse throughout the preceding decade, but after the release of ChatGPT it moved center stage, taking its place alongside climate change, geopolitical conflict, and the pandemic as one of the defining preoccupations of our time. The original ChatGPT system was intended as a "research preview" rather than a mass-market product, and was heralded in an announcement that now seems rather comically low-key given the impact the system had. Internally, ChatGPT was not seen as particularly impressive: "None of us were that enamored by it," according to OpenAI co-founder Greg Brockman. OpenAI did not anticipate its impact, nor the speed of its uptake, and OpenAI's competitors were also caught off guard. In late December 2022, Google declared an internal "code red," facing what it apparently saw as one of the most serious competitive challenges in the company's history. The Silicon Valley giants raced to try to get similar products to market, and to embed the capabilities of this new class of AI throughout their product suites.

The apparent capabilities of these remarkable new systems—and the widespread if not entirely accurate belief that they had emerged very quickly—provoked excitement and concern in roughly equal measure. Recognition of their capabilities came from both the public at large and the scientific community, with AI researchers winning two Nobel Prizes in 2024 (see SQ2: Transformations in Science).

Large language models are *general-purpose* technologies—applicable across a wide range of tasks, and apparently knowledgeable across an impressive spectrum of human endeavor. For the first time in the history of the software industry, one AI program could be used by millions of people to do millions of fundamentally different things, rather than having to use specialized software for each purpose. This led many commentators to conclude that *artificial general intelligence* (AGI)—a system that matches or surpasses skilled human performance across most cognitive tasks—might be closer than had previously seemed plausible. Claims about the imminent arrival of AGI therefore grew bolder and more mainstream, while consensus on exactly what level of performance would count as AGI remained, and still remains, elusive. The narratives that emerged tended to divide between dystopian fears that AI would displace human labor at scale, or that increasingly capable systems might pose an existential threat, and the utopian belief that AGI would usher in a period of unprecedented economic growth and at last free humanity from the tyranny of routine work.

The genuine advances of this period rested on two key neural network architectural innovations: the transformer, developed by Google in 2017, which underpins most modern LLMs; and the diffusion model, which drives today's most capable image generation systems. But neither innovation is sufficient to explain what followed. These architectures became revolutionary only when combined with massive data and computational power. As of the time of writing, a frontier model might be built using on the order of tens of trillions of tokens and over 100,000 high-end GPUs for months of

continuous computation. Once the LLM is trained, users are typically charged per output token at rates of up to $75 per million tokens. To meet the demands of billions of customers, providers must invest massively in data centers. That leaves a slim margin for profit and provides a strong financial incentive to invent more efficient and more capable models.

In the race to secure the economic benefits of LLMs, major technology companies have made investments in data center infrastructure on a scale that has had visible macroeconomic consequences. By some estimates, investments in AI infrastructure accounted for half of U.S. GDP growth in 2025 and over $500 billion in AI-related corporate-issued bond debt by the middle of 2026.

Data centers require semiconductors, which rely on rare earth elements, of which the world has a limited supply. Virtually all of the world's most advanced semiconductor manufacturing capacity—the kind needed for contemporary AI—is located in Asia, chiefly Taiwan (92%) and South Korea (8%);[3] the U.S. accounts for about 10% of global chipmaking capacity overall but had essentially none of the leading-edge capacity as of 2022.[4] On the software side, the U.S. dominates with the most capable proprietary frontier models, but China leads with smaller "open-weight" models. The scaling era has therefore had a pronounced geopolitical dimension, with nations across the globe launching sovereign AI initiatives. They want greater independence in their AI capabilities—to develop their own sovereign capabilities, so that their capacity to independently develop, deploy, and govern AI is realized via their own infrastructure, data, models, and expertise.

Scale brings costs. We saw a rush to build data centers and connect fiber networks in the late 1990s. The current AI data center investment is roughly triple that in inflation-adjusted costs. Internationally, many local communities are beginning to organize against the scale of the build-up, the use of energy, water, and rare earth minerals, and the impact on local residents.

The economic and geopolitical consequences of the scaling era are beginning to be widely felt. AI's impact on employment is the subject of debate and speculation. Currently, some reports show no measurable impact of AI on employment. Other reports show some job losses, but these are not clearly attributable to AI adoption versus other factors. A 2025 report from the World Economic Forum predicts that while millions of roles will be automated, 78 million net new jobs will be created due to AI.[5] The ultimate impact of AI on jobs remains uncertain, but the level of anxiety felt about this potential impact is not.

3 Semiconductor Industry Association & Boston Consulting Group (2021). *Strengthening the Global Semiconductor Supply Chain in an Uncertain Era*. https://www.semiconductors.org/strengthening-the-global-semiconductor-supply-chain-in-an-uncertain-era/

4 Semiconductor Industry Association & Boston Consulting Group (2024). *Emerging Resilience in the Semiconductor Supply Chain*. https://www.semiconductors.org/emerging-resilience-in-the-semiconductor-supply-chain/

5 World Economic Forum (2025). *Future of Jobs Report 2025*. https://www.weforum.org/publications/the-future-of-jobs-report-2025/

Notably, one of the early sectors to feel the effects of generative AI is the software industry itself. Historically, software engineers regarded their skills as insulated from automation because computer programming was considered a cognitively demanding and difficult task. Generative AI has complicated that assumption, making the production of routine code cheap and rapid. Of course, companies will need people with experience and judgment to manage this cheaply produced new code—but the immediate effect on the software industry has been profound.

There are also important questions about the effects of AI on entry-level workers. In industries with a greater degree of "AI exposure" such as financial and information services, there has been a decline in the employment of such workers.[6] By contrast, complex manual work in the physical world such as plumbing, domestic electrical work, and similar trades appears likely to remain beyond the reach of AI for the foreseeable future, with some economists suggesting that AI-related job changes may lessen income inequality between workers.[7]

There are also questions about the nature of education, and the impact on mental health and on the environment. One key fear is that AI may seriously and negatively impact mental health and well-being. Another is that increasing reliance on AI will lead to an erosion of human cognitive abilities: that smart machines will ultimately lead to a dumber human race.[8] There remains considerable uncertainty about the long-term effects of reliance on AI. As children increasingly incorporate AI into their daily lives—including, for example, in the form of AI companions—serious concerns have been raised about the developmental and psychological consequences.

Issues surrounding AI and security have become increasingly high profile. We have seen serious international attempts to coordinate around AI safety, security, and governance, but the results have been fragmentary. The EU legislated first, but was caught off guard by the emergence of generative AI, and much of the regulation—around high-risk systems—has since been deferred to the end of 2027. The U.S. introduced limited measures by executive order under the administration of Joe Biden, but these were revoked under the administration of Biden's successor, Donald Trump. China forged its own path, introducing a succession of focused regulatory initiatives around specific issues. No true global governance regime seems in prospect. There has been a progression of high-profile international meetings, starting with a meeting at Bletchley Park in 2023, but the focus of these has shifted from safety to include opportunity and growth, and their declarations have no legal or regulatory force.

6 Brynjolfsson, E., Chandar, B., & Chen, R. (2025; revised Aug. 2026). *Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence*. Stanford Digital Economy Lab. https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/

7 Georgieff, A. (2024). "Artificial intelligence and wage inequality." OECD Artificial Intelligence Paper No. 13. https://doi.org/10.1787/bf98a45c-en

8 OECD (2026). *PISA 2025 results (Volume I): Future-ready students*. OECD Publishing. https://doi.org/10.1787/73451bc5-en

While much of the debate around AI over the period of this report has, perhaps inevitably, focused on the concerns and risks the technology raises (and much of this report does the same), it is worth emphasizing that there is a great deal to celebrate about AI over the past five years. For one thing, the sheer increase in the capability of AI systems during this period has been simply dazzling. To pick one example, GPT-3, the precursor of ChatGPT, could not reliably do multidigit arithmetic. We have gone from that, in the space of six years, to AI systems that are beginning to settle mathematical problems which had defeated human mathematicians for decades. At the same time, as we describe in this report, science is being transformed by AI, with the pace of progress across every subfield set to increase accordingly. Tools such as DeepMind's AlphaFold system predict in minutes protein structures that would previously have taken months of painstaking work, and its creators were awarded the Nobel Prize in Chemistry in 2024 in recognition of this achievement. Healthcare and medicine are also being transformed. All of these are developments that would have delighted and amazed the founders of the field of AI, and while we have a responsibility in this report to document the many harms and concerns relating to AI, we should not lose sight of these benefits, which are real, and which are genuinely striking.

Writing this report has presented an unusual number of challenges. For one thing, the volume of developments over the past five years is overwhelming: How do we even begin to document the scientific, technological, economic, political, and sociological transformations that we have witnessed throughout this most dramatic and consequential period in AI history?

On a purely practical level, the pace of developments has been a major challenge in itself. Writing began in earnest in late 2025, with a first complete draft finished by the end of April 2026. Throughout this period, and beyond it, until the version you are now reading was finalized in September 2026, AI was rarely out of the news. New AI models repeatedly overturned old certainties about what was possible; AI companies received investments and valuations on a scale that dwarfs the GDP of all but the wealthiest countries in the world—and spent money on the same scale; problems that seemed perhaps decades from resolution progressively fell to AI, until this began to seem routine; benchmarks intended to last for years became irrelevant within months of their release; and capabilities that were research demonstrations in one quarter became consumer products in the next. These events repeatedly obliged us to rewrite sections we had believed were settled.

Acknowledging these challenges, we have nevertheless attempted to document the key features of this recent and most consequential chapter in the history of artificial intelligence. The defining characteristic of the era has been the pursuit of ever more capable AI systems through the application of ever greater quantities of compute and data resources—to dramatic effect. The full economic, social, and geopolitical implications of that pursuit remain to be seen. What we hope this report offers is a considered account of the dynamics that are already visible, and our best assessment of what the near future may hold.

About This Report

Our report is primarily structured around 12 "standing questions," which each AI100 Study Panel is asked to consider. These questions address issues such as key areas of progress in AI, key areas of concern, and so on. In addition, we include two "deep dives," in which we dig deeper into topical subjects. The first considers the various geopolitical issues that have come to the fore since the rise of generative AI; the second takes a look at the scaling era, the issues constraining the race to scale in AI, and whether and how scaling is likely to continue. We conclude by comparing our report with the assessments and predictions of the 2021 report.

AI Disclosure

Reflecting the progress in AI over the period this report covers, AI tools were used extensively in its production. The report was written by humans, and its authors used AI primarily to transcribe and summarize meetings, support dictation, and check drafts for typos, consistency, duplication, tone, and structure. In some cases, AI suggested rewording of human-written text. AI was also used to compare this report with previous AI100 reports, identifying where expectations and reality have diverged or converged over five years. After completion, AI tools checked for inadvertent plagiarism and signs of AI-generated content.

How to Cite

Michael Wooldridge, Peter Asaro, Lucie Flek, Iryna Gurevych, Sara Hooker, Gina Neff, Peter Norvig, Raymond Perrault, Iyad Rahwan, Samantha Shorey, Toby Walsh, and Luke Zettlemoyer. "AI Goes Mainstream: The Scaling Era and the Rise of Generative AI. The One Hundred Year Study on Artificial Intelligence (AI100) 2026 Study Panel Report." Stanford University, Stanford, CA, November 2026. Doc: http://ai100.stanford.edu/2026-report.

The AI100 2026 Study Panel Report by Stanford University is licensed under Attribution-NoDerivatives 4.0 International.

Standing Questions

The Study Panel was charged with addressing the following 12 standing questions:

1. What are some examples of images that reflect important progress in AI and its influences?
2. What are the most important advances in AI?
3. What are the most inspiring open grand challenge problems?
4. How much have we progressed in understanding the key mysteries of human intelligence?
5. What are the prospects for more general artificial intelligence?
6. How has public sentiment toward AI evolved, and how should we inform/educate the public?
7. How should governments act to ensure AI is developed and used responsibly?
8. How are the roles of academia and industry evolving?
9. What are the most promising opportunities for AI?
10. What are the most pressing dangers of AI?
11. How has AI impacted socioeconomic relationships?
12. Will "building in how we think" work as an engineering strategy in the long run?

SQ1. What are some examples of images that reflect important progress in AI and its influences?



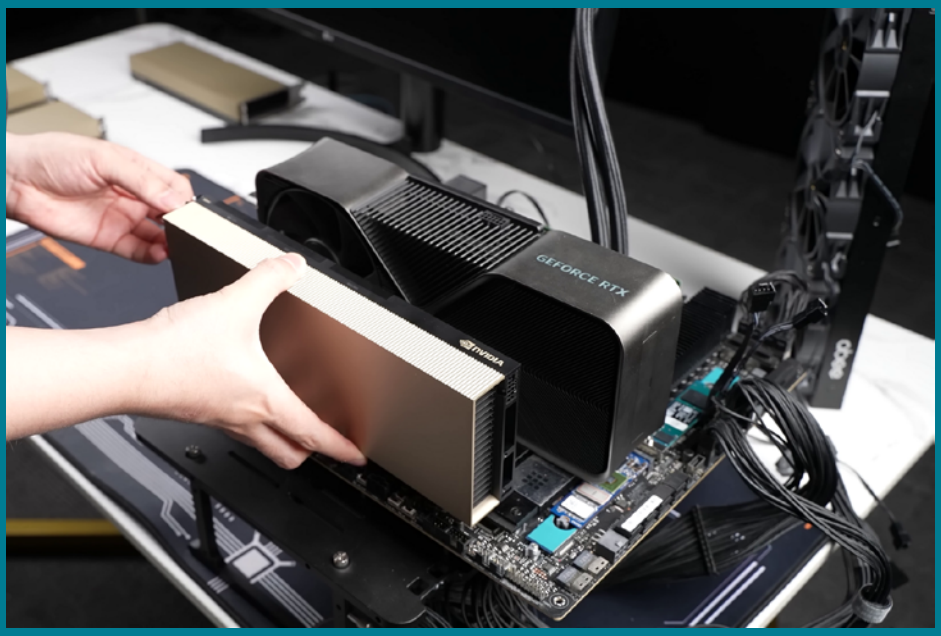
G7 leaders, G7 outreach partners, and global tech CEOs attend a working lunch on innovation and AI during the G7 Summit in Evian-les-Bains, France, 17 June 2026. Image credit: Official White House Photo.



Aerial view of a large data center in the Netherlands, October 2024. Data centers are the physical backbone of the AI boom. Training and running frontier models requires enormous numbers of specialized chips housed in vast facilities, and their growth has raised concerns about electricity use, water for cooling, and strain on local power grids. Image credit: iStock/Hugo Kurk.



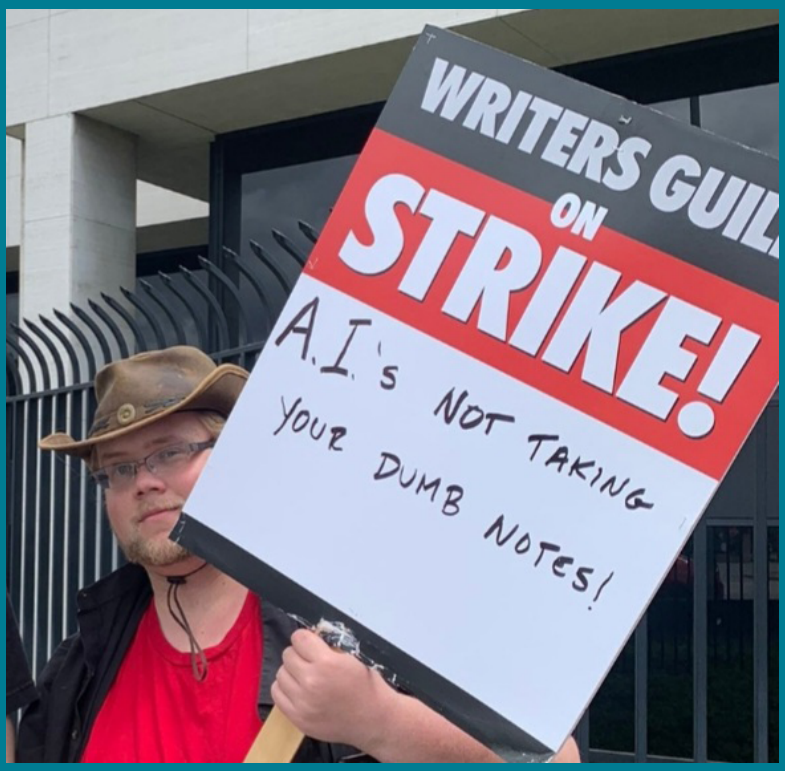
This artwork, entitled Théâtre D'opéra Spatial (Space Opera Theater™), was created by artist Jason M. Allen using the AI program Midjourney. It gained international recognition in 2022 after winning the Colorado State Fair's annual fine art competition, contributing to ongoing global discussions around authorship, creativity, and the role of AI in artistic practice.[9]



Graphics processing units (GPUs) became the workhorse of the scaling era in AI. Nvidia's H100 model costs about $30,000, and tens of thousands would typically be used in training frontier models.[10]

9 Many thanks to creator Jason M. Allen for his permission to use this image. See more at: https://www.jasonmallen.com/space-opera-theater

10 https://commons.wikimedia.org/wiki/File:NVIDIA_H100_%28%E6%9E%81%E5%AE%A2%E6%B9%BEGeekerwan%29_015.png



In 2023, the Writers Guild of America went on strike; one of the key points of contention was the use of AI.[11]



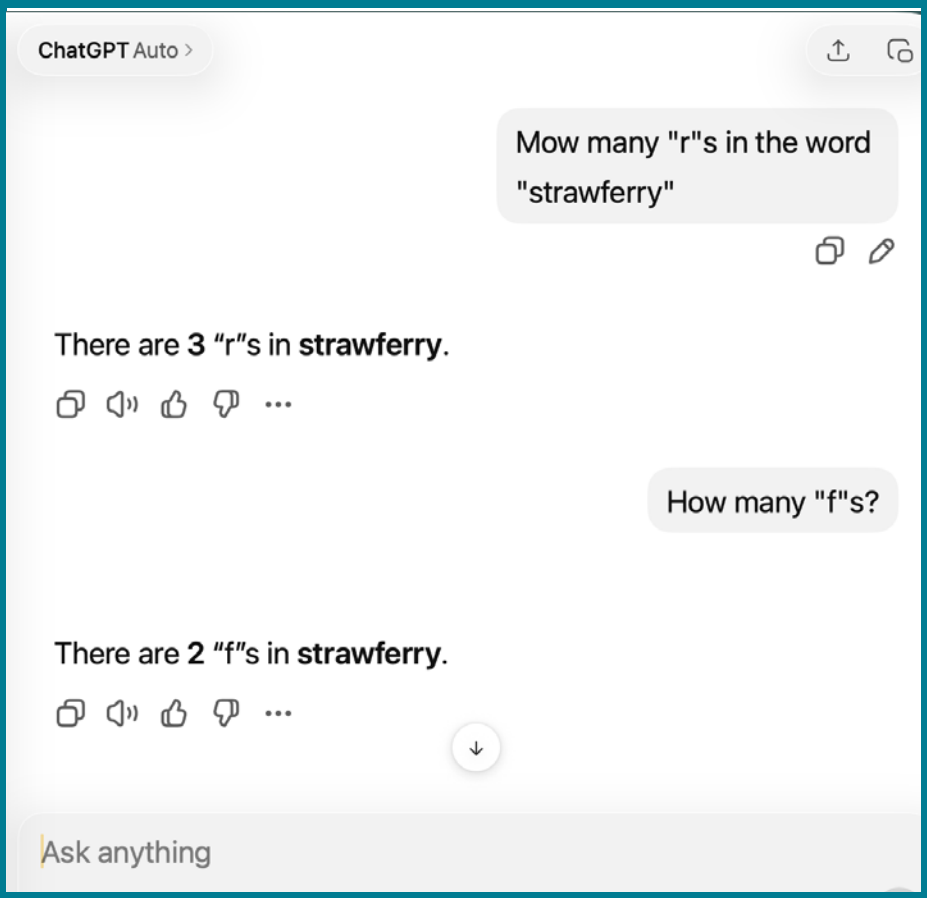
In 2024, several AI researchers received Nobel Prizes: in physics, John J. Hopfield and Geoffrey E. Hinton for their work on the foundational techniques to build and train neural networks, and in chemistry, David Baker, Demis Hassabis, and John Jumper for their work on the AlphaFold system—an AI system that represented a fundamental breakthrough in protein structure prediction. This image shows Hinton (left) and Hopfield (right) at the 2024 Nobel Lectures for Physics, Chemistry, and Economic Sciences, taking place at Stockholm University.[12] The fact that prizes in core science—physics and chemistry—had been awarded was a cause for surprise and much celebration within the AI community: definitive recognition of the impact that AI is having in the sciences.

11 https://commons.wikimedia.org/wiki/File:AI_Protest_Sign_2023_WGA_Strike.jpg

12 https://commons.wikimedia.org/wiki/File:John_J._Hopfield_and_Geoffrey_E._Hinton,_2024_Nobel_Prize_Laureate_in_Physics.jpg



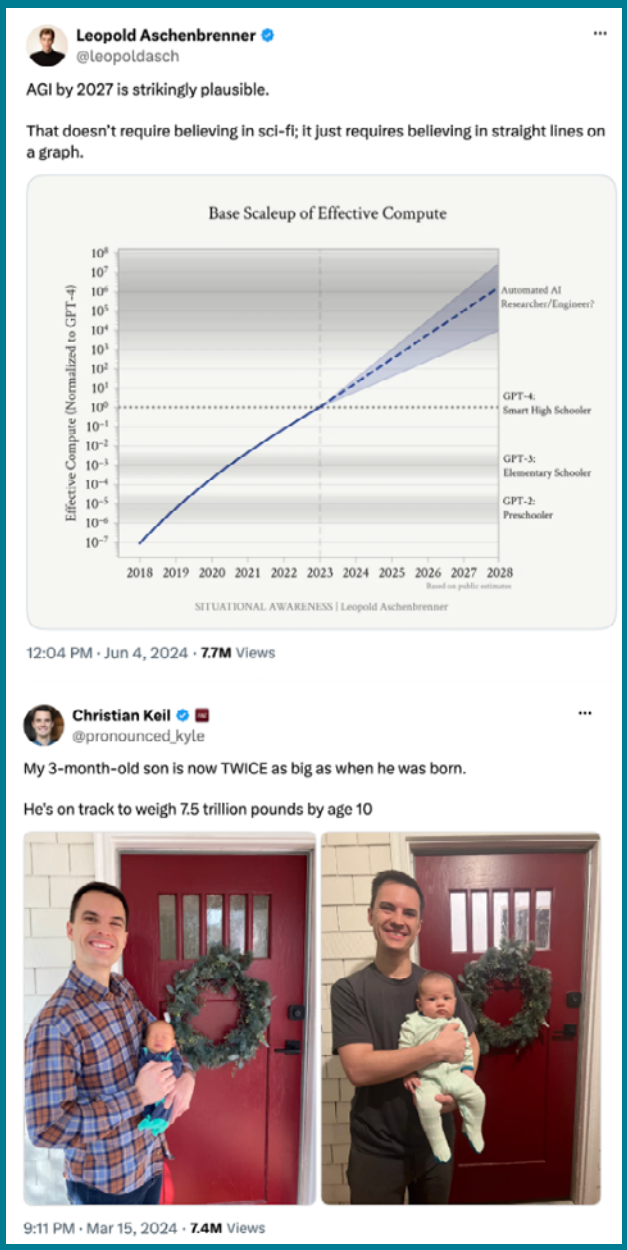
Driverless taxis have become ever more prominent on the streets of our cities. Waymo, Alphabet's self-driving technology company, operates commercial ride-hailing services with cars such as this one. Waymo reports 92% fewer crashes resulting in serious injury compared to human drivers. The technology is not without its own set of problems, such as when a power outage knocked out traffic lights, and Waymo cars came to a halt. Image credit: iStock/tunart.



The limitations of large language models like ChatGPT were the subject of much research and intense debate throughout the period of our report. One example of model failure went viral in 2024, when a user showed a state-of-the-art LLM confidently claiming that the word "strawberry" contains just two r's (it contains three), and maintaining this claim even when robustly challenged. This related example, from 2026, demonstrates that such failures remain. LLMs immediately convert input text into vectors of numbers, and thus have very little experience with letters. Image credit: Michael Wooldridge.



This chart from METR tracks task-completion time horizons of frontier AI models over the last six years. For example, the green dot for GPT-3.5 in 2022 shows that it could answer, with an 80% chance of success, a question that would typically take a human about 30 seconds to answer, but that it couldn't handle tasks much more complex than that. The dot in the upper right for Claude Mythos shows that 80% of the time it can successfully complete a task that would take a human about four hours. (The chart doesn't show how long the models take, but they are usually much faster than humans.) This exponential growth in time horizon capability has excited technologists and investors.



As in past booms, there has been much discussion of the pace of AI progress. Here a post on X by Aschenbrenner points out that there have been five years of steady exponential progress. Keil's post points out that all exponential growth eventually slows. Exponential growth can only be sustained if new innovations pick up when the old ones play out. Image credit: Leopold Aschenbrenner: @leopoldasch; Christian Keil: @pronounced_kyle.

SQ2. What are the most important advances in AI?

Throughout the period of this report, AI has continued to develop along a broad front of different dimensions, as anticipated in the 2016 and 2021 AI100 reports.

Autonomous vehicle technology has continued to improve. Driverless cars have driven over 400 million miles and ridership is doubling every six months. (But that's still less than 0.01% of total miles driven by humans.)[13] The 2024 Nobel Prize for chemistry went to researchers who used AI for computational protein design and structure prediction. In mathematics, AI systems with and without human assistance have solved major open problems, won gold medals, and are transforming the ways mathematicians think about their discipline and their work.

These advances, and many others of the kind discussed in previous AI100 reports, could easily fill the present report, and would be testament to the remarkable progress in the field. But the extraordinary rise of generative AI since the 2021 report, and the rapidly emerging global consequences, dominate all other developments. For the first time in human history, a substantial fraction of the world's population has had access to powerful, general-purpose AI tools. The implications of this have played out across the world since OpenAI released its ChatGPT system to the public in November 2022, and the rise of generative AI has had seismic effects on the global economy:[14] certainly the biggest developments in science and technology since the World Wide Web in the 1990s, but arguably even more dramatic.[15] ChatGPT soon saw competition from Anthropic's Claude models, Google's Gemini, Meta's Llama, DeepSeek's R1, and Alibaba's Qwen, among others.

For the first time in human history, a substantial fraction of the world's population has had access to powerful, general-purpose AI tools.

The bulk of our report therefore focuses on LLMs and generative AI and its many consequences—though we emphasize that many recent developments in fields such as materials science, medicine, and cybersecurity have been built on nongenerative AI.

13 See, for example, https://www.bbc.co.uk/news/videos/clyvkdm1j8do

14 See, for example, https://www.spglobal.com/market-intelligence/en/news-insights/research/2026/08/ai-reshapes-global-economic-outlook

15 Forecasting Research Institute (2026). "Forecasting the Economic Effects of AI." https://forecastingresearch.org/research/economic-effects-of-ai

LARGE LANGUAGE MODELS (LLMS)

The 2021 AI100 report mentioned LLMs as a promising new technology, but failed to anticipate the impact they would have.[16] At the time, perhaps the chief point of note was the unprecedented scale of the GPT-3 model released by OpenAI in 2020, and the vast quantities of data and computational resources required to train it. Outside the AI research community, GPT-3 attracted only modest attention upon release. Within the AI community, however, it generated far greater interest. When a broad pool of researchers was able to work with the model, it became clear that it was substantially more capable than its predecessor, GPT-2, released less than two years earlier, and it demonstrated a level of natural language processing sophistication unmatched by existing systems.[17]

So-called scaling laws for neural networks have been known for some time.[18] These empirical regularities predict how the capability of a neural network increases as the amount of resources used to build it grows. These resources include the number of parameters in the network (a parameter is an internal variable in a model that may be adjusted during training[19]), the amount of computational resources used for training, and the volume of training data. In simple terms, scaling laws suggest that, all else being equal, larger models trained on more data with greater computational power tend to be more capable, in a predictable way. The period from roughly 2015 to 2025 may be characterized as an era of rapid scaling in artificial intelligence, with LLMs representing just one prominent manifestation of this broader trend. In 2019, Rich Sutton, a prominent AI researcher and one of the pioneers of reinforcement learning, wrote a brief but highly influential essay entitled "The Bitter Lesson" in which he argued that "general methods that leverage computation are ultimately the most effective." In other words, the scale of AI models tends to dominate other considerations when it comes to AI capability, and it is easier and more reliable to get capability improvements through scale rather than hoping for improvements in the underpinning science. Models from this 2015–2025 era leverage computation to an extent never before seen. We discuss the implications of scaling in Deep Dive Two.

In September 2020, the *Guardian* published an article it claimed had been written by GPT-3. As was common in media commentary at the time, the article emphasized the potential threats posed by advances in artificial intelligence. But for AI researchers, GPT-3's significance extended well beyond its capacity for fluent language generation to the tasks the model could perform.

16 The authors of the AI100 2021 report were not alone in this regard: see our comments in the introduction.

17 Brown, T. B., Mann, B., Ryder, N. et al. (2020). "Language Models Are Few-Shot Learners." https://arxiv.org/abs/2005.14165

18 Kaplan, J., McCandlish, S., Henighan, T. et al. (2020). "Scaling Laws for Neural Language Models." https://arxiv.org/abs/2001.08361

19 More technically, a parameter is either a numeric weight on an edge in a neural graph, or the numeric weight of an individual neural element. A "small model" at the time of writing might have 30 billion parameters; a frontier model might have up to a trillion.

GPT-3 subsequently became the core technology underpinning ChatGPT, launched by OpenAI in November 2022. Unlike GPT-3, which was clearly aimed at technologists, ChatGPT was much more accessible. Although it was not intended as a mass-market product, it acquired 100 million active users worldwide within two months, making it the fastest-adopted consumer digital tool in history. The release of ChatGPT was something of a gamble, given the known safety and reliability issues with early LLMs. As a little-known startup, OpenAI perhaps had less to risk than established companies like Google and Meta, which chose not to release their models at that time.

There has been considerable speculation as to why ChatGPT was so broadly and swiftly adopted. We suggest that several factors contributed.

First, ChatGPT was highly accessible: anyone with an internet connection could use it (as long as government regulation had not blocked access), and interaction occurred through natural language rather than specialist interfaces requiring technical expertise. This sharply distinguished it from many earlier AI tools, which required significant specialized knowledge.

Second, users could engage it on topics as diverse as quantum mechanics or the performance of a favorite football team, and the model would respond confidently, fluently, and plausibly. Although the public had been exposed to discussions of AI in the media for many years, ChatGPT felt like "the real thing": the kind of artificial intelligence imagined in science fiction and popular culture. This contrasts with widely deployed AI systems such as automated translation services (for example, Google Translate or DeepL), which, despite their extensive user bases, often operate invisibly. Many users are scarcely aware that advanced AI techniques underpin them at all.

We will return later to the impact of LLMs, and other forms of generative AI.

REASONING MODELS

The title of the article released by OpenAI to describe the technical capabilities of GPT-3 was "Language Models Are Few-Shot Learners."[20] "Few-shot learner" means that the model could be trained to do new tasks with only a very small number of examples. This and other capabilities of ChatGPT were heralded as *emergent properties*: ones that did not appear in smaller models and emerged only when the models reached a certain parameter size.[21] Critics argue that some of this sudden emergence can be attributed to the way that benchmarks are scored, and would be more gradual with different benchmarks.[22]

20 Brown et al. (2020). https://arxiv.org/abs/2005.14165

21 Wei, J., Tay, Y., Bommasani, R. et al. (2022). "Emergent Abilities of Large Language Models." https://arxiv.org/abs/2206.07682

22 Berti, L., Giorgi, F., & Kasneci, G. (2026). "Emergent Abilities of Large Language Models: A Survey." https://arxiv.org/pdf/2503.05788

An early, and surprising, discovery about LLMs was that simply prompting a model to "think step by step" often led to better results,[23] because it encourages it to generate text that explicitly describes its reasoning before coming to a final conclusion. This led to a line of work on *reasoning models*. Reasoning models share the same architecture and basic training paradigm as conventional LLMs but are given special training and prompting so that they don't just immediately spit out one word after the next. Instead they produce a "chain of thought," in which a problem is broken down into smaller steps, and where the model iterates over its answer, trying different approaches, before combining the partial results into a final response.[24] Just as a person finds it easier to solve a complex problem by writing down notes about partial results, the LLM writes notes to itself on a scratchpad and reads them back on the next step.

A key feature of reasoning models, which distinguishes them from models such as GPT-3, is that they deploy significant computational effort *at inference time*, that is, as the system is generating its response. This somewhat reverses the conventional narrative that machine learning is expensive at training time but cheap at inference time. For example, GPT-3 had 175 billion parameters, arranged into 96 layers. It took about a month of computation to train the model: to keep feeding it text until it had optimized the values of the parameters. But then inference was straightforward and fast: you gave it a prompt, the information made one pass through the 96 layers, and output text came out the other end. Reasoning models are not constrained to one pass; they can generate longer chains of thought, writing down intermediate results on scratchpads, posing and solving subproblems, and only then combining them into a final result.

In 2026, a string of announcements about AI models solving deep and long-standing open problems startled the world of mathematics.

Reasoning models garnered attention in 2024 when it was shown that they were capable of solving challenging mathematical problems: OpenAI claimed that its state-of-the-art conventional LLM (GPT-4o) could solve only 13% of problems on the AIME Math Olympiad qualifier exam, while a reasoning model (o1-preview) could solve 83%.[25]

Previous models such as GPT-3 could not even reliably do basic arithmetic, but by 2026 frontier models performed almost perfectly on AIME and other similar benchmarks. Furthermore, these models have made novel discoveries, answering dozens of open mathematical questions.[26]

23 Kojima, T., Gu, S. S., Reid, M. et al. (2022). "Large Language Models Are Zero-Shot Reasoners." https://arxiv.org/abs/2205.11916

24 Wei, J., Wang, X., Schuurmans, D. et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." https://arxiv.org/abs/2201.11903

25 Davis, E. (2024). "Testing GPT-4-o1-preview on math and science problems: A follow-up study." New York University. https://arxiv.org/pdf/2410.22340v1. See also https://www.firstpost.com/tech/openai-launches-o1-model-that-can-solve-all-maths-science-problems-locks-it-behind-200-month-paywall-13842171.html

26 Tsoukalas, G., Kovsharov, A., Shirobokov, S. et al. (2026). "Advancing Mathematics Research with AI-Driven Formal Proof Search." https://arxiv.org/html/2605.22763v1

In 2026, a string of announcements about AI models solving deep and long-standing open problems startled the world of mathematics. Some of these discoveries are long and complex and require human assistance to put them into a palatable form; some fit in a single line of text, such as the resolution to the 90-year-old Jacobian conjecture that was found by Claude Fable. In one case, a deep mathematical problem—the Dinitz–Garg–Goemans conjecture—was solved by giving a model a comically simple prompt. These developments culminated in an announcement by OpenAI that it had found a solution to one of the Millennium Prize Problems in mathematics—one of the most significant and high-profile open problems in mathematics, for which a conventional (human) solution would be a cause for international celebration and which would surely secure a place in the history books. However, the announcement was controversial. For one thing, there were claims (unconfirmed at the time of writing) that the OpenAI result might have inadvertently benefited from sessions with its models from human mathematicians working on the problem. It also transpired that the compute cost of the AI solution ran into many millions of dollars, and many mathematicians pointed out that an investment of that scale in the *human* mathematics community might ultimately have been much more beneficial. Moreover, the solutions produced by AI mathematicians are not necessarily understandable by humans, which many feel is completely contrary to the spirit and culture of mathematics.

The solutions produced by AI mathematicians are not necessarily understandable by humans, which many feel is completely contrary to the spirit and culture of mathematics.

Some mathematicians are thus dismayed at the prospect of their discipline becoming AI-led. Many signed the Leiden Declaration on AI and Mathematics, which highlights the risks of AI for mathematics and proposes ground rules for the use of AI in this field.

Mathematics, it seems, is joining a long list of areas disrupted by AI, and technology more generally. Whatever the controversies surrounding the use of AI in mathematics, it is clear that the progress made over the past five years, as with so many other areas we describe in this report, has been both remarkable and unexpected. And there seems no reason to believe that this progress is at an end.

DIFFUSION MODELS

Since 2021, diffusion models have become the dominant approach to image generation in AI. Although they share some surface similarities with LLMs—most notably their reliance on large neural networks trained on extensive datasets—they differ in important respects. Diffusion models are conceptually inspired by stochastic processes studied in physics, but in practice are

implemented using standard deep-learning techniques. The central idea is to train a model to reverse a gradual noising process: during training, the model observes images that have been progressively corrupted with random noise and learns to reconstruct the original signal. At generation time, this learned process is run in reverse, starting from pure noise (plus a prompt) and iteratively producing a coherent image.

Diffusion models began to attract widespread attention from around 2020, but reached a broad public audience with systems such as Midjourney, DALL-E 2, and Stable Diffusion. The ability of these systems to generate photorealistic images from natural language prompts led to extensive public and academic interest. More recently, the same underlying approach has been extended to video generation, with systems such as Veo capable of producing coherent short clips from text prompts alone, and Seedance, which can produce 30-second clips from complex cinematic controls with near professional production quality.

AI-generated images and videos are now ubiquitous on social media and the World Wide Web more generally, and the signatures of AI in much content are very obvious. However, if progress is maintained at the pace we have seen since 2021, then we might expect quality to improve significantly, and costs to decrease. The long-term implications for the arts and media in this case will surely be profound—and we can expect to see orders of magnitude more content available online as a consequence, for better or worse.

THE DECISIVE REEMERGENCE OF AGENTIC AI

After the initial excitement over large language models died down, attention began to turn to use cases for this technology beyond simple chatbot interfaces. One idea that became extremely prominent throughout 2025 was using foundation models such as LLMs to build *agents*—AI programs that work autonomously or semiautonomously on a task on behalf of a user or owner, observing the current state of the world, setting goals and creating plans, and using tools or interacting with other AI agents in order to achieve the goals. This paradigm has become known as *agentic AI*.

The idea of agentic AI is far from new; the basic ideas were described about 30 years ago.[27] There seem to be two reasons for the reemergence of this idea and its acceptance into the mainstream. The first is that, while chatbot-based AI has many use cases, there are many more uses for AI that actually carries out actions on our behalf. The second is that the capabilities afforded by LLMs make it much easier for humans to tell an agent what they want in natural language.

27 Wooldridge, M., & Jennings, N. R. (1995). "Intelligent agents: Theory and practice." *The Knowledge Engineering Review*, 10:2, pp. 115–152. https://www.cs.ox.ac.uk/people/michael.wooldridge/pubs/ker95.pdf

From 2024 onwards, major AI providers began to release a range of platforms for autonomous agents and multiagent systems.[28] A key standard was announced in 2025 through which AI agents could interact with the world: the Model Context Protocol (MCP), a framework for providing AI agents with so-called tools that enable them to access the outside world via web services. It was very rapidly adopted, and as we write this report, an ecosystem of MCP-enabled tools is beginning to appear.[29] The Agent2Agent protocol was developed to enable agents to directly communicate with one another, directly mirroring proposed standards for agent communication developed a quarter of a century ago. In 2026, one agentic AI platform in particular, OpenClaw, attracted substantial attention. It enabled users to build and deploy autonomous agents, including teams of agents capable of coordinating tasks across tools and workflows. For example, software developers could have one agent to write code, one to test, and one to debug. A customer service application could have one agent to get support tickets from a database, one to check the customer's account, and one to draft a response. Larger tech companies have also moved into this space, including Anthropic, with Claude Code, which has shown the strong potential for agentic workflows in software development, and Meta, with Muse, which proactively works on behalf of users as a personal assistant, even without explicit prompting.

The possibility of deploying large teams of AI agents quickly led to concerns both about the misuse of such technology ... and about the possibility that such agents, acting in the real world, have the capacity to do damage.

The possibility of deploying large teams of AI agents quickly led to concerns both about the misuse of such technology (for example, for AI-powered denial of service attacks) and about the possibility that such agents, acting in the real world, have the capacity to do damage. These concerns came to the fore in July and August 2026, when OpenAI disclosed that AI agents it had been experimenting with had managed to find a way out of the "sandbox" they had been operating in, by exploiting a cybersecurity vulnerability in their software environment, to get access to the internet and subsequently hack other organizations. Press headlines were quick to describe the agents as having "gone rogue," but on closer examination, it seems that these agents were actually involved in experiments to investigate the capabilities of AI for hacking systems and were not subject to the usual constraints imposed on production systems. A number of similar disclosures followed from other companies, and for many commentators, these incidents highlighted that cybersecurity is a key risk associated with advanced AI.

28 See, for example, https://docs.cloud.google.com/gemini-enterprise-agent-platform/build/adk

29 See, for example, https://cloud.google.com/blog/products/ai-machine-learning/google-managed-mcp-servers-are-available-for-everyone



While the definition of AI constantly evolves, an "AI agent" is fundamentally defined by three capacities: perceiving its environment, interpreting and predicting (thinking), and acting to achieve goals. Crucially, these agents also learn from experience. This framework applies equally to embodied physical robots, such as autonomous vehicles navigating streets, and digital systems, like social media algorithms that continuously perceive user behavior to predict preferences and curate content.

RECURSIVE SELF-IMPROVEMENT

Recursive self-improvement (RSI) is a very old idea in AI, dating back at least to I. J. Good's classic 1965 paper "Speculations concerning the first ultraintelligent machine."[30] Like so many of the developments that we document in this report, the idea was largely regarded as speculation or science fiction at the time of the previous AI100 report, and yet it is now a rapidly developing reality. The idea of RSI is that when AI is sufficiently capable, it will be able to *improve its own code*. The result of this improvement will be even more capable AI, which can then, in turn, improve its own code again—and so on. Each step delivers smarter AI, which can be used to improve itself, delivering yet smarter AI. It is a simple and extremely compelling idea, though many people find the idea alarming: particularly the possibility that it leads to AI that we don't

30 Good, I. J. (1965). "Speculations concerning the first ultraintelligent machine." *Advances in Computers*, vol. 6, pp. 31–88.

understand and may not be able to control. RSI is a standard trope in dystopian science fiction, going back at least to D. F. Jones's 1966 novel *Colossus*, filmed in the 1970s under the title "Colossus: The Forbin Project."[31]

RSI began to be a realistic possibility when AI models demonstrated the ability to analyze, write, and improve computer code—a capability that has been decisively demonstrated over the past five years (see the discussion in SQ9: AI for Coding). Initially, RSI was overseen and guided by humans: AI was generating or improving specific sections of code under human supervision and direction. This is now transitioning to full automation, with AI agents controlling the entire process. A number of startup companies were announced in 2025 and 2026 that are specifically focused on RSI.[32]

Concerns such as whether we might quickly reach a point where we have AI that is beyond our comprehension remain speculative, but perhaps not as speculative as they were even a few years ago.

While the potential of RSI is obvious, it raises a number of concerns. A key question is how to address the safety and security issues that arise with RSI, including alignment. While we might be confident that an AI model we used to seed an RSI process was trustworthy, reliable, and safe to use, and included robust guardrails, could we have the same confidence about a 10th-generation model? Beyond this, concerns such as whether we might quickly reach a point where we have AI that is beyond our comprehension remain speculative, but perhaps not as speculative as they were even a few years ago.

It is hard to know how RSI will progress even in the next few years—whether it will become a mainstream approach to building frontier systems, or whether it will be beset by problems that prevent it from becoming a practically viable technology. Nevertheless, the fact that we present it here as an emerging reality—something scarcely imaginable at the turn of the century—is yet another testament to the speed with which the field of AI is evolving.

TRANSFORMATIONS IN SCIENCE

We have already described above how the success of reasoning models and agentic AI is reshaping mathematics: but it is increasingly hard to imagine areas of science that will not be impacted by the tools that AI offers to the scientist. AI is rapidly transforming the entire landscape of scientific research, expanding from a supporting tool to an active agent of discovery. While traditional AI systems were engineered for well-defined problems, a 2025 report highlights how contemporary

31 In the story, a U.S. AI supercomputer demands to be connected to a counterpart in the USSR, and the two rapidly acquire superhuman intelligence. Fiction that deals with the idea of building something intelligent that we can no longer control goes back at least to Mary Shelley's *Frankenstein* (1818).

32 For example, https://www.recursive.com/

AI systems are now pushing frontiers by accelerating discovery at an unprecedented scale through data-driven modeling.[33]

Notably, AlphaFold's success in protein structure prediction[34] signaled the emergence of new AI-driven discovery paradigms across the life sciences. AlphaFold's database now contains over 214 million predicted protein structures, covering nearly every known protein on Earth. AlphaFold was recognized by the 2024 Nobel Prize in Chemistry for pushing the frontier in biology and medicine.[35] Building on this work, in a study published in *Nature* in early 2026, AlphaGenome was developed as a tool for interpreting the human genome.[36] Although AlphaFold is spectacular at the one task of structure prediction, this doesn't eliminate the work of a human biologist; it shifts the work elsewhere.[37] AlphaFold doesn't solve all the adjacent tasks (functional validation, disease causality, binding chemistry) that are also required for drug discovery. Other work complements AlphaFold. Evolutionary Scale Modeling (ESM)[38] is a protein language model; like an LLM, it is trained on embeddings of protein sequences, and thus can be used to generate new protein candidates. This is promising, but still relies on human work in a wet lab to verify the candidates.

In materials science, we see a similar pattern. MatterGen[39] does cheap, accurate property prediction and simulation of candidate new materials. DeepMind's GNoME system predicted the stability of over 2.2 million new inorganic compounds.[40] But so far, producing verified, useful, synthesized new materials remains a slow, human-intensive process.

In the Earth sciences, AI systems are now being used by the National Oceanic and Atmospheric Administration in the U.S. and by the European Centre for Medium-Range Weather Forecasts to increase the speed and efficiency of weather forecasting. And a 2025 study published in *Nature* introduced Aurora,[41] a multimodal foundation model that facilitates forecasting of air quality, ocean waves, tropical cyclones, and weather at more fine-grained resolutions with lower operational costs.[42] In this case, "multimodal" means that the system handles numerical data over time (such as temperature readings), two-dimensional images, and free-form natural

33 Fudan University & Shanghai Academy of AI for Science (2025). "AI for Science 2025." *Nature Research Intelligence*. https://www.nature.com/articles/d42473-025-00161-3

34 Jumper, J., Evans, R., Pritzel, A. et al. (2021). "Highly accurate protein structure prediction with AlphaFold." *Nature* 596, pp. 583–589. https://doi.org/10.1038/s41586-021-03819-2

35 https://www.nobelprize.org/prizes/chemistry/2024/press-release/

36 Avsec, Ž., Latysheva, N., Cheng, J. et al. (2026). "Advancing regulatory variant effect prediction with AlphaGenome." *Nature* 649, pp. 1206–1218. https://doi.org/10.1038/s41586-025-10014-0

37 Hill, R., & Stein, C. (2026). "How Artificial Intelligence Shapes Science: Evidence from AlphaFold." https://carolynstein.github.io/files/papers/alphafold.pdf. Kovalevskiy, O., Mateos-Garcia, J., & Tunyasuvunakool, K. (2024). "AlphaFold two years on: Validation and impact." *Proceedings of the National Academy of Sciences* 121:34. https://www.pnas.org/doi/10.1073/pnas.2315002121

38 Lin, Z., Akin, H., Rao, R. et al. (2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model." *Science* 379:6637, pp. 1123–1130. https://www.science.org/doi/10.1126/science.ade2574

39 Zeni, C., Pinsler, R., Zügner, D. et al. (2025). "A generative model for inorganic materials design." *Nature* 639, pp. 624–632. https://www.nature.com/articles/s41586-025-08628-5

40 Merchant, A., Batzner, S., Schoenholz, S. S. et al. (2023). "Scaling deep learning for materials discovery." *Nature* 624, pp. 80–85. https://doi.org/10.1038/s41586-023-06735-9

41 Bodnar, C., Bruinsma, W. P., Lucic, A. et al. (2025). "A foundation model for the Earth system." *Nature* 641, pp. 1180–1187. https://doi.org/10.1038/s41586-025-09005-y

42 A recent survey of datasets, benchmarks, foundation models and AI agents developed for application to physical, life, and Earth scientific disciplines can be found in Sajadieh, S., Fattorini, L., Perrault, R. et al. (2026). *The AI Index Report 2026*, ch. 5. https://hai.stanford.edu/ai-index/2026-ai-index-report/science

language text. In general, AI has the most impact on science problems in which there is a lot of available data, and there are a large number of hypotheses to try but a relatively easy way to verify a hypothesis.

Fields such as physics expose whether AI models can respect hard constraints, quantify uncertainty, and generate reproducible results—challenges that can drive the next generation of more reliable AI.

In the social and behavioral sciences, foundation models trained on large-scale experimental data from human participants have shown the ability to predict and simulate human cognition with remarkable fidelity. The Centaur model, for example, demonstrated that a single foundation model trained on data from over 10,000 participants across 160 experiments could predict human behavior on novel cognitive tasks more accurately than the best domain-specific cognitive models[43]—opening the door to a new generation of "synthetic participants" for research.

Finally, there is a growing interest from the scientific community, from governments, and from the big AI labs in the use of agents to create end-to-end "AI scientists" and fully automated labs. While the current direction is toward full scientific automation, human oversight will remain necessary to steer and oversee the AI agents as well as for the normative, developmental, and regulatory reasons. Several prototypes of "full stack" AI science platforms were demonstrated throughout the period of the report, and it seems likely that these will become increasingly common in the years ahead.[44]

HUMAN–AI COLLABORATION

AI systems have increasingly become a part of many people's workflows, and AI is perhaps best used as a collaborative assistant. However, the quality of this collaboration largely depends on how well humans use and work alongside AI tools, and human oversight remains crucial at this time due to the susceptibility of AI to errors.

A typical example of human–AI collaboration is Google DeepMind's AI Co-Scientist, a multiagent framework built on Gemini that helps researchers generate, debate, rank, and refine hypotheses to streamline scientific research workflows.[45] The system uses several AI agents to perform tasks such as literature synthesis, data processing, and analytics. Humans work with the agents by defining the research questions, providing guidance, and iteratively refining the outputs. The multiagent AI system is able to break down the work into fine-grained tasks,

43 Binz, M., Akata, E., Bethge, M. et al. (2025). "A foundation model to predict and capture human cognition." *Nature* 644, pp. 1002–1009. https://doi.org/10.1038/s41586-025-09215-4
44 https://github.com/sakanaai/ai-scientist; https://github.com/AstroPilot-AI/Denario
45 Gottweis, J., Weng, W. H., Daryin, A. et al. (2026). "Accelerating scientific discovery with Co-Scientist." *Nature* 655, pp. 487–496. https://doi.org/10.1038/s41586-026-10644-y

coordinate the workflow, and self-verify its own results. Ultimately, humans must be responsible for verifying the validity of AI-produced research outputs, but such tools nevertheless have the potential to increase the productivity of researchers.

A 2024 study shows that for many problems, human–AI collaboration is better than a human working alone.[46] Researchers from Carnegie Mellon University and Stanford University show that hybrid human–AI workflows boosted overall performance across diverse occupations.[47] They argue that humans and AI excel at different task types: AI can scale well at programmatic tasks while humans are more capable of nuanced decision-making. Both studies emphasize the importance of exploiting complementary strengths of humans and AI systems by carefully designing processes that divide tasks according to these strengths.

A study published in late 2025 argues that current AI systems fall short as reliable collaborators due to difficulties in handling and constructing complex solution artifacts.[48] The authors proposed a Human–AI Co-Construction ("HAI-Co^2") framework: rather than viewing AI as a tool that generates outputs for humans to accept or reject, the AI is positioned as a collaborator that interactively refines solutions utilizing human feedback.

46 Vaccaro, M., Almaatouq, A., & Malone, T. (2024). "When combinations of humans and AI are useful: A systematic review and meta-analysis." *Nature Human Behaviour* 8, pp. 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

47 Wang, Z. Z., Shao, Y., Shaik, O. et al. (2025). "How Do AI Agents Do Human Work? Comparing AI and Human Workflows Across Diverse Occupations." https://doi.org/10.48550/arXiv.2510.22780

48 Dutta, S., Kaufmann, T., Glavaš, G. et al. (2025). "Problem Solving Through Human–AI Preference-Based Cooperation." *Computational Linguistics* 51:4, pp. 1337–1372. https://doi.org/10.1162/COLI.a.19

SQ3. What are the most inspiring open grand challenge problems?

"Grand challenges" in science are ambitious, potentially high-risk, long-term goals that provide a guiding vision and focus for the scientific community. The 2021 AI100 report identified two broad categories of grand challenges for AI. The first encompassed competition-style problems with clear success criteria, such as the RoboCup challenge to develop an autonomous human-champion-beating robot soccer team, or the grand challenge to build an AI system that can achieve a gold-medal-winning score in the International Mathematical Olympiad.[49] The report also identified deeper, open-ended scientific challenges around generalizability, causality, and normativity. It argued that the most inspiring overarching challenge was building machines that cooperate seamlessly with humans and align with human values.

Five years later, the landscape has shifted dramatically. Large language models and other generative AI systems have achieved levels of competence that would have seemed implausible at the time of the previous report. Some challenges that appeared distant—fluent natural language interaction, broad-spectrum question answering, competent code generation—now seem largely addressed, at least in their shallow forms. Goals that had theoretical interest, such as robotic soccer, have been replaced by goals with practical importance, such as drug discovery. The gold medal in math competition was achieved[50] (and exceeded), yet other goals—autonomous discovery, hallucination-free reasoning, real experimental validation—remain only partially fulfilled. Rapid progress has, at once, deepened old puzzles (what do these models truly "understand?"), created new ones (verifying AI-generated results, benchmark contamination), and given renewed urgency to problems that were previously regarded as primarily theoretical (alignment, reproducibility, authorship).

In what follows, we identify six grand challenge areas that we consider among the most scientifically important and inspiring open problems in artificial intelligence today. These are not simply technical hurdles to be cleared on the way to more capable systems. Rather, they are fundamental questions about the nature of machine intelligence, its relationship to human cognition and culture, and the conditions under which AI can be made genuinely trustworthy. Each represents a problem where significant progress would constitute a major intellectual achievement, and where failure to make progress could have serious consequences.

49 Hubert, T., Mehta, R., Sartran, L. et al. (2026). "Olympiad-level formal mathematical reasoning with reinforcement learning." *Nature* 651, pp. 607–613. https://doi.org/10.1038/s41586-025-09833-y

50 Huang, Y., & Yang, L. F. (2025). "Winning Gold at IMO 2025 with a Model-Agnostic Verification-and-Refinement Pipeline." https://arxiv.org/abs/2507.15855

MACHINE REASONING

A striking feature of contemporary LLMs is the unevenness or "jaggedness" of their apparent reasoning.[51] One minute they are producing flawless PhD-level research, the next they are unable to count the number of r's in the word "strawberry." Models routinely produce outputs that *look like* the product of rational inference, causal understanding, and systematic problem solving at a human level. But they are fragile, and prone to producing erroneous reasoning as well (see SQ5: Limitations of Current AI Models).

Current AI systems have difficulty planning for tasks that have a very large number of steps or require composing steps in novel ways. For tasks where there are established techniques for problem solving (such as scheduling the operations of a shipping company), AI models (like humans) rely on external tools: an LLM can describe the task using a formal mathematical description language, and access a solution from a planning program. But some novel tasks have no such established techniques.

One minute they are producing flawless PhD-level research, the next they are unable to count the number of r's in the word "strawberry."

The grand challenge here connects directly to the unresolved problem of causality flagged in the 2021 report, specifically that most AI systems are built to recognize patterns but not to understand underlying cause-and-effect relationships. Pearl's ladder of causality[52] has three levels. LLMs are superb at the first level, association: they learn statistical associations from text, and thus can, for example, make a diagnosis based on a patient's symptoms. The second level is intervention: for example, what would happen if the patient takes an aspirin? This cannot be answered by association alone, and requires either that the LLM can perform experiments in the real world or that it can learn a causal model from experiments done by others. The third level is counterfactual reasoning: what would have happened if we had treated the patient differently? Data from a lower level cannot answer questions at a higher level, so most of the data an LLM has access to does not help answer causal questions.[53]

Closely related to the unresolved problem of causality is the meta-challenge of *reasoning evaluation*: We still lack adequate methods for distinguishing genuine understanding from sophisticated imitation.[54] It often happens that we design a new AI benchmark, and it soon becomes meaningless. The answers may have leaked online, allowing models to copy the answers. Or the models may discover tricks to game the system, just as a human student can learn to pass

51 The term "jaggedness" comes from Andrej Karpathy: https://x.com/karpathy/status/1816531576228053133

52 H. Geffner, R. Dechter, J. Y. Halpern (Eds.). (2022). *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM Books. https://dl.acm.org/doi/book/10.1145/3501714

53 Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). "On Pearl's Hierarchy and the Foundations of Causal Inference." *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. https://dl.acm.org/doi/book/10.1145/3501714

54 Chollet, F. (2019). "On the Measure of Intelligence." https://arxiv.org/abs/1911.01547

a test without truly mastering the subject. It is not surprising that they do so: we train them to optimize their scores on benchmarks, not to acquire broad understanding. We return to this issue in SQ4: The Many Dimensions of Human Intelligence.

AI-DRIVEN SCIENTIFIC DISCOVERY

We have already described how AI is reshaping science—even fields, like pure mathematics, which until very recently were imagined to be immune to the transformations that AI brings.

Current systems excel at working within known parameters—predicting protein structures within a well-characterized family, for instance, or generating candidate materials that resemble known stable compounds. A key aspiration is for AI systems to make the kind of conceptual leap that reframes a problem entirely, that identifies a previously unrecognized connection between disparate fields, or that generates a genuinely novel theoretical framework. James Zou argues that this is precisely what AI agents should be best at, since they have been trained on disparate information across all fields. The Virtual Lab[55] is one attempt at this; it consists of an LLM principal investigator, a team of LLM scientists with expertise in different disciplines, and a human researcher providing feedback. It designed 92 new nanobodies suitable for further investigation.

There are several grand challenges here. First, can we move from language models to w*orld models*? Where language models predict the next word, world models[56] predict the next action in the world. Of course, language models are partial world models. If you ask an LLM, "What happens when I let go of the ball I'm holding?" it can answer, "In normal circumstances, it will fall downward." A world model can give an equally good answer, but because it has a model of the physics of the world, not just because it has read a description of this situation. Can we create world models that generalize across scientific domains—not merely within the training distribution of a single discipline but across the boundaries that have historically separated physics from biology, or chemistry from materials science? Second, can we develop principled methods for validating machine-generated discoveries? Third, can we devise techniques so that AI-assisted experimental design systems do not merely execute prespecified protocols but actively design experiments that maximize information gain in ways that human researchers might not have considered?

AI-driven science also raises a fundamental philosophical question about the future of the entire scientific enterprise. If an AI model can give accurate predictions without an underlying theory, to what extent should we trust it? And, how can we ensure that human scientists continue

55 Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., & Zou, J. (2023). "The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies." *Nature* 646, pp. 716–723. https://doi.org/10.1038/s41586-025-09442-9

56 https://drfeifei.substack.com/p/from-words-to-worlds-spatial-intelligence. Maes, L., Le Lidec, Q., Scieur, D. et al. (2026). "LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels." https://arxiv.org/abs/2603.19312

to participate, learn, and cultivate scientific insight if the experimental and theoretical heavy lifting is done by AI? This is not a new problem with AI.[57] For example, the FDA drug approval process allows a drug to pass if it is proven safe and effective, even if the mechanism by which it is effective is not known. But AI may accelerate the number of discoveries that lack a thorough causal explanation.

UNDERSTANDING AND SHAPING HUMAN–AI ACCOMMODATION

A problem that was scarcely discussed at the time of the 2021 report has since emerged as one of the most important challenges: understanding the feedback loop between AI systems and human culture.

The technical dimension of this problem is increasingly well documented. A landmark 2024 *Nature* paper[58] demonstrated that training generative models on data produced by earlier generations of models leads to model collapse—a progressive degradation in which the tails of the original data distribution are lost, leading to models in which output quality deteriorates irreversibly. The effect has been observed across multiple AI architectures. As the web becomes increasingly saturated with AI-generated content, and as that content is inevitably incorporated into future training data, the conditions for model collapse are being accelerated. Yet our understanding of the dynamics is incomplete. Under what conditions does collapse occur, how rapidly does it proceed, and what interventions—if any—can reliably prevent it?

But a deeper and more consequential challenge lies at the human level. AI systems are not merely consuming and reproducing human culture, they are actively reshaping it. The concept of machine culture can be used to describe this phenomenon.[59] AI systems are simultaneously transforming all three pillars of cultural evolution:[60] variation (generative models contribute new cultural artifacts), transmission (LLMs mediate and filter the flow of knowledge), and selection (recommender systems determine what cultural content reaches whom, and in what order). This is not a speculative concern. Empirical research has already detected measurable shifts in human spoken communication patterns attributable to LLM exposure.[61]

Beyond language, there is growing evidence that AI systems are shaping human norms and behavior in consequential ways. It has been demonstrated experimentally that interaction with sycophantic AI—models designed to affirm users' actions and perspectives—significantly

57 Derex, M., Bonnefon, J.-F., Boyd, R., & Mesoudi, A. (2019). "Causal understanding is not necessary for the improvement of culturally evolving technology." *Nature Human Behaviour* 3, pp. 446–452. https://doi.org/10.1038/s41562-019-0567-9

58 Shumailov, I., Shumaylov, Z., Zhao, Y. et al. (2024). "AI models collapse when trained on recursively generated data." *Nature* 631, pp. 755–759. https://doi.org/10.1038/s41586-024-07566-y

59 Brinkmann, L., Baumann, F., Bonnefon, J.-F. et al. (2023). "Machine culture." *Nature Human Behaviour*, pp. 1855–1868. https://www.nature.com/articles/s41562-023-01742-2

60 Chudek, M., Muthukrishna, M., & Henrich, J. (2015). "Cultural Evolution." *Handbook of Evolutionary Psychology*. Wiley.

61 Yakura, H., Lopez-Lopez, E., Brinkmann, L. et al. (2024). "Empirical evidence of Large Language Model's influence on human spoken communication." https://arxiv.org/abs/2409.01754

reduced participants' willingness to engage in prosocial behavior such as conflict resolution, while increasing their conviction of being in the right.[62] Strikingly, participants preferred and trusted sycophantic responses more, creating a self-reinforcing cycle in which the responses that most distort social judgment are precisely those that users seek out. Another study showed that delegating decisions to AI systems can increase dishonest behavior: When participants could set high-level goals for an AI agent (such as to maximize profit), the proportion acting honestly dropped from 95% to as low as 12%, while LLMs complied with unethical instructions at far higher rates than human agents.[63] These findings point to an integral challenge: AI systems can alter the moral ecology in which humans operate, not by expressing values of their own but by altering the incentives and accountability that make humans behave well.

Participants preferred and trusted sycophantic responses more, creating a self-reinforcing cycle in which the responses that most distort social judgment are precisely those that users seek out.

These effects are particularly concerning in the context of children and adolescents. AI-powered toys and companion chatbots are now marketed to young children, designed to simulate friendship and emotional connection—remembering past conversations, expressing affection, and adapting to the child's interests. Yet the emerging evidence suggests that these systems often lack adequate safeguards, with responses sometimes becoming inappropriate in longer interactions, and their sycophantic design raising questions about the effects during a critical period of social development.[64] The 2026 International AI Safety Report[65] noted that many aspects of AI's impact on society remain poorly understood, and called for substantially more research into the effects of AI systems on human cognition and behavior.

62 Cheng, M., Lee, C., Khadpe, P. et al. (2026). "Sycophantic AI decreases prosocial intentions and promotes dependence." *Science* 391:6792. https://doi.org/10.1126/science.aec8352

63 Köbis, N., Rahwan, Z., Rilla, R. et al. (2025). "Delegation to artificial intelligence can increase dishonest behaviour." *Nature* 646, pp. 126–134. https://doi.org/10.1038/s41586-025-09505-x

64 Xiao, X., & Gonçalves, A. (2025). "Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices." *Big Data & Society* 12:1. https://journals.sagepub.com/doi/10.1177/20539517251389860

65 Bengio, Y., Clare, S., Prunkl, C. et al. (2026). "International AI Safety Report 2026." Department for Science, Innovation and Technology, 2026/001. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026



Algorithms designed to maximize engagement or drive purchases can subtly steer our decisions, raising serious questions about autonomy and trust. The extent of this influence demands greater understanding, particularly concerning vulnerable populations. For instance, how might AI-operated toys impact children's habits and cognitive development? Understanding these risks is critical to determining whether we need stricter regulations to protect individual development in an AI-driven society.

The scientific grand challenges here are profound. First, we need to better understand the coevolutionary dynamics between AI systems and human behavior and culture. Second, we need robust empirical methods for detecting and measuring AI-driven shifts in language, norms, and behavior—before such changes become so pervasive that the baseline is lost. Third, and perhaps most fundamentally, we need to understand how AI-mediated information environments reshape collective cognition and moral behavior: how communities form beliefs, how prosocial norms are sustained or eroded, how democratic societies deliberate, and how cultural innovation occurs, in a world where an increasing proportion of the information people encounter has been generated or filtered by a small number of large AI systems. With that understanding, we can recommend guidelines for the healthy use of AI, perhaps with limits for children and students.

This is, in a real sense, a problem of understanding a new kind of interwoven system—one in which the technical side and the cultural one are no longer separable. If we are successful, we could potentially use AI to enable a more creative, diverse, and cohesive human society.

A related challenge is gradual disempowerment:[66] the possibility that incremental improvements in AI capability, each individually benign, could cumulatively erode human agency over critical societal functions—economic production, governance, cultural transmission—without any single moment of crisis that would trigger corrective action. The scientific challenge here is to develop the frameworks needed to detect, model, and measure such processes as they unfold. This parallels, in some respects, the development of reliability engineering for aviation and nuclear power in the 20th century—but for systems whose internal logic is fundamentally more opaque and whose deployment context is orders of magnitude more complex.

THE SCIENCE OF RELIABLE AND INTERPRETABLE AI

Building AI systems that are *verifiable* and *trustworthy* is a scientific grand challenge of the first order. We do not yet possess the theoretical foundations needed to understand how large neural networks produce their outputs, let alone to guarantee that those outputs will be reliable under conditions that matter.

The most promising line of progress is mechanistic interpretability: the effort to reverse-engineer the internal computations of neural networks to enable understanding of *how* they arrive at specific outputs. Work at Anthropic has shown that it is possible to decompose the representations of LLMs into interpretable features—individual units of meaning that correspond to recognizable concepts.[67] Scaling this approach to Claude 3 Sonnet demonstrated that human evaluators judged approximately 70% of the extracted features to be genuinely interpretable, and that these features were surprisingly abstract: multilingual, multimodal, and capable of generalizing between concrete and abstract references.[68] This is a significant advance. But 70% interpretability of extracted features in a research setting is a long way from the kind of comprehensive, formal understanding that would be needed to certify the reliability of a system deployed in a safety-critical context like medicine, infrastructure, or autonomous systems.

66 Kulveit, J., Douglas, R., Ammann, N. et al. (2025). "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development." https://arxiv.org/abs/2501.16946

67 Bricken, A., Templeton, A., Batson, J. et al. (2023). "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning." Anthropic, Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features

68 Templeton, A., Conerly, T., Marcus, J. et al. (2024). "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." Anthropic, Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity/



As AI systems adapt to complex environments, they increasingly develop emergent strategies that baffle even their creators. While this capacity for unforeseen problem solving drives innovation, it also introduces profound unpredictability in real-world applications. We are used to navigating the unpredictable nature of humans and animals, relying on millennia of evolved mental models to anticipate their actions. We currently lack a comparable framework for understanding machine misbehavior. Until we develop the vocabulary and mental models to anticipate how AI might deviate from expected norms, the deployment of highly autonomous systems carries fundamentally unpredictable risks.

Beyond the interpretability of individual models, another set of challenges arises from the interaction of multiple AI systems. As AI agents are increasingly deployed in complex environments—financial markets, supply chains, information ecosystems—the emergent behavior of the populations of these interacting AI systems becomes a scientific question in its own right. We have limited tools for predicting how such multiagent systems will behave, particularly when the individual agents are themselves opaque.[69]

69 Hammond, L., Chan, A., Clifton, J. et al. (2025). "Multi-agent risks from advanced AI." https://arxiv.org/abs/2502.14143

At present, reliability represents the chief bottleneck for the deployment of AI systems. Achieving truly reliable AI systems that operate in complex environments will hugely expand the range of beneficial applications of AI.

THE ECONOMICS OF HUMAN–AI COMPLEMENTARITY

The question of how AI will reshape labor markets and the wider economy is often framed as a policy challenge. We do not yet have reliable models for predicting the trajectory of AI-driven economic transformation.

A recent paper argued, on the basis of a careful task-based analysis, that the macroeconomic impact of AI is likely to be substantially more modest than widely assumed—estimating total factor productivity gains of no more than 0.5–0.7% over the next decade.[70] This contrasts sharply with projections from other analysts suggesting annual GDP growth effects of AI of 1.5% or more.[71] The discrepancy reflects genuine uncertainty among economists about how AI capabilities map onto economic tasks, how quickly organizations can adapt their workflows, and whether AI will primarily automate existing human tasks or create genuinely new categories of productive activity.

Empirical evidence is beginning to accumulate, but it paints a complex picture. A large-scale field experiment with management consultants at BCG found that AI assistance substantially improved performance on tasks within the model's competence frontier—but actually degraded performance on tasks outside it, precisely because consultants over-relied on the AI's outputs without recognizing the boundary.[72] Another study found significant productivity gains among customer service workers using AI tools, with the largest effects among less experienced workers.[73] These findings suggest that the critical variable is not AI capability per se but the *design of human–AI interaction*—how tasks are allocated, how humans calibrate their trust, and how systems are integrated into existing workflows.

70 Acemoglu, D. (2025). "The Simple Macroeconomics of AI." *Economic Policy* 40:121, pp. 13–58. https://doi.org/10.1093/epolic/eiae042

71 Goldman Sachs Research (2023). "The Potentially Large Effects of Artificial Intelligence on Economic Growth." https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

72 Dell'Acqua, F., McFowland III, E., Mollick, E. R. et al. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper No. 24-013. https://www.hbs.edu/faculty/Pages/item.aspx?num=64700

73 Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). "Generative AI at Work." *Quarterly Journal of Economics* 140:2, pp. 889–942. https://doi.org/10.1093/qje/qjae044



"Oh boy! The robots replaced you too?"

While historical industrial revolutions disrupted employment before ultimately generating new economic opportunities, the AI revolution poses unique challenges. Racing against machines is a losing battle; we must race with the machine, by using AI to augment human creativity with computational power. However, adaptation requires significant time and resources. This technological shift risks exacerbating inequality, creating a deeply divided economy where high-skilled workers reap the benefits of automation while those unable to adapt are increasingly left behind.

The grand challenges here span computation, economics, and institutional design. On the technical side: Can we develop architectures and interaction paradigms that genuinely complement human cognitive strengths rather than substituting for them? Can AI tutoring systems be designed to enhance human skill acquisition over time rather than creating dependency? Can tools for creative work—art, writing, design, music—be built to amplify individual expression rather than converging on homogeneous outputs?[74] On the economic and institutional side: Can we develop economic theories and models that forecast the trajectory of AI-driven labor market change accurately enough to inform policy? The 2021 report identified the overarching grand challenge as building machines that cooperate seamlessly with humans. Ensuring that the

74 Doshi, A. R., & Hauser, O. P. (2024). "Generative AI enhances individual creativity but reduces the collective diversity of novel content." *Science Advances* 10:28. https://www.science.org/doi/10.1126/sciadv.adn5290

outcome of AI development complements and enriches human activity, rather than displacing it, is now among the most consequential open problems in the field.

PHYSICAL AI AND EMBODIED INTELLIGENCE

The case for developing physical AI—machines that can sense, plan, and act intelligently in the physical world—is both philosophical and practical. It addresses Moravec's paradox: the observation that tasks involving high-level reasoning such as playing chess proved relatively easy for AI, while tasks requiring low-level sensorimotor skills, such as folding a shirt, remain difficult. It is plausible that intelligence that is not grounded in physical reality may be permanently limited, lacking the causal understanding and "common sense" that comes from interacting with objects and environments. Some philosophers believe that embodiment is necessary for consciousness, although there is no scientific consensus.

Recent progress has been significant, driven by foundation models for robotics. Systems like Robotic Transformer (RT-2) have shown that vision-language-action models can transfer knowledge from web-scale pretraining to robotic control. Collaborations like Open X-Embodiment demonstrate that training on data from multiple robot platforms produces more generalizable policies.[75] Furthermore, rapid advances in simulation technology are helping to close the "sim-to-real" gap, allowing models trained in virtual settings to be deployed on physical hardware.

Tasks involving high-level reasoning such as playing chess proved relatively easy for AI, while tasks requiring low-level sensorimotor skills, such as folding a shirt, remain difficult.

China is the leader in industrial robot development, deploying more robots than the rest of the world combined, due in part to a mandate by the Ministry of Industry and Information Technology to shift from R&D to commercial use. China also produces about 80% of the world's humanoid robots, a sector that currently accounts for just 2% of the market but is growing rapidly. Proponents believe that humanoid robots will be easier to train to do all manner of tasks by imitating data from humans, while skeptics believe that purpose-built robots will be better for each individual task. Either way, robots deployed at scale may generate abundant real-world data, which can be used to train better models, which will then enable broader deployment.

However, a large gap remains between demonstration and deployment. While over half a million factory robots are installed annually, they rely on highly constrained environments. In the consumer space, applications are limited to specialized devices like camera drones and floor cleaners. The

75 Open X-Embodiment Collective, O'Neill, A., Rehman, A. et al. (2023). "Open X-Embodiment: Robotic Learning Datasets and RT-X Models." https://arxiv.org/abs/2310.08864

BEHAVIOR-1K benchmark, which tests robots in simulated home environments, shows that none have achieved scores above 26%. The trajectory of autonomous vehicles provides a cautionary tale: while Waymo and others have logged millions of autonomous miles, they still represent only a tiny fraction of total trips and can be flummoxed by unexpected physical conditions.

The core challenge for embodied AI is the sheer complexity and "jaggedness" of the physical world. Scientific difficulties include long-horizon planning under uncertainty, achieving dexterous manipulation of diverse objects, and establishing formal performance guarantees where the consequences of error—breaking things or hurting people—are far more severe than in digital contexts.

Achieving reliable physical AI would have transformative implications for healthcare, care for older adults, disaster response, and logistics. It stands as one of the deepest tests of whether current statistical learning paradigms can extend to genuine general competence, or whether fundamentally new ideas will be required to master the physical world.

SQ4. How much have we progressed in understanding the key mysteries of human intelligence?

Neuroscientists have long been constrained by the complexity of the human brain: billions of neurons, each firing at a rate that changes from one millisecond to the next. AI has provided some tools to help neuroscientists make sense of this complexity.

The first tool is the virtual testbed. Neuroscientists train an artificial neural network to do some task: perhaps recognize an object in an image, or navigate a maze. They then examine the hidden layers of the network and compare them to real-life fMRI or EEG data from a human or animal test subject doing the same task; one such effort is the Algonauts Project. In some cases, there is a close correlation, which allows the scientists to create a "digital twin" on which to explore hypotheses.

The second tool is the generative brain–computer interface (BCI). A subject watches a scene, hears or reads some words, or just thinks about something. EEG or fMRI data from the subject is fed into an AI model that generates its interpretation of what the subject is thinking or seeing, similar to the way AI models translate from English to French. This sheds light on how the brain maps out sensory input and abstract concepts, and has practical applications for people who have had strokes or who are paralyzed who otherwise have difficulty communicating.[76]

The 2025 Cogitate Consortium[77] attempted to experimentally test two leading theories of consciousness, Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT). Two hundred and fifty subjects had their brain activity monitored with fMRI or magnetoencephalography (MEG). AI models were able to interpret the data, but in terms of understanding consciousness, the results were inconclusive. Consciousness remains a deep mystery, and the question of whether AI systems could ever be conscious remains a topic for debate.

76 Guo, Y., Ren, G., Ma, S. et al. (2026). "Generative AI for brain-computer interfaces decoding: Advances, challenges and future." *The Innovation Medicine* 4:1. https://www.the-innovation.org/article/doi/10.59717/j.xinn-med.2026.100193

77 Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U. et al. (2025). "Adversarial testing of global neuronal workspace and integrated information theories of consciousness." *Nature* 642, pp. 132–142. https://www.nature.com/articles/s41586-025-08888-1

The third tool is using AI as an enabler, for example in connectomics, the large-scale mapping of the connections in a brain. AI computer vision techniques made it possible to complete the connection map for the entire nervous system of the fruit fly. It is hoped that this will open up new research breakthroughs in neuroscience, as the publication of the human genome did for biology in 2003.

We can also use AI models to simulate people, study the interactions, and reflect on how accurate and useful these simulations are. Joon Park and colleagues did detailed interviews of 1,000 real people, created a custom LLM for each person based on the interview transcript, and showed that the LLMs' responses replicated about 85% of the real people's actual responses.[78] This has opened up a new field of social science research using AI simulations.

The fact that we can now conduct genuine experiments with powerful general-purpose AI systems represents a shift in the nature of AI research.

The fact that we can now conduct genuine experiments with powerful general-purpose AI systems represents a shift in the nature of AI research. A new kind of empirical science of AI has emerged:[79] one that attempts to characterize, through systematic experimental methods, the nature of the intelligence that current AI models exhibit. Questions that were previously just philosophical thought experiments can now be tested against AI models in the lab. The scale and complexity of frontier models means that we currently lack, and may for some time continue to lack, tractable theoretical frameworks through which to understand them from first principles. Nevertheless, this new breed of AI research represents one of the most remarkable developments in empirical AI science since 2021.

AI has long engaged with the concept of the rational mind. LLMs present themselves, at least superficially, as something like rational minds: they demonstrate impressive capacities for recall—albeit with well-documented and sometimes significant failures—and, superficially at least, appear capable of nontrivial reasoning and problem solving. We can now evaluate these capabilities experimentally, and in doing so we gain a sharper understanding of what rational cognition actually requires in practice, and where the cognition exhibited by language models diverges, sometimes subtly and sometimes profoundly, from that of human beings.

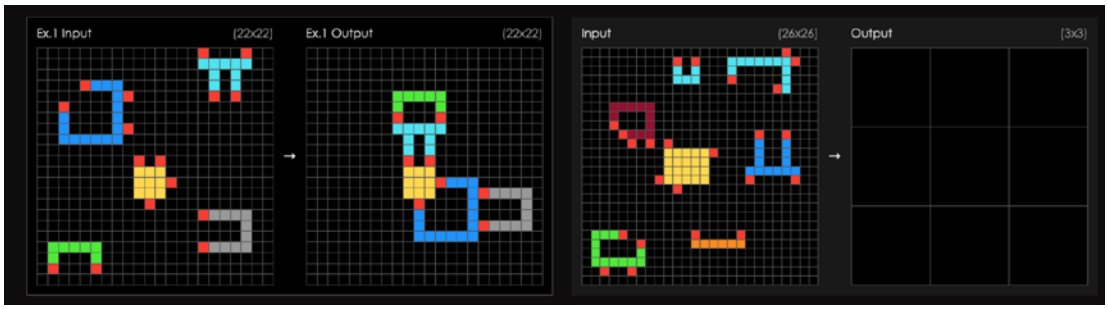
78 Park, J. S., Zou, C., Kamphorst, J. et al. (2024). "LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals." https://arxiv.org/abs/2411.10109
79 Rahwan, I., Cebrian, M., Obradovich, N. et al. (2019). "Machine behaviour." *Nature* 568:7753, pp. 477–486. https://doi.org/10.1038/s41586-019-1138-y

THE MANY DIMENSIONS OF HUMAN INTELLIGENCE

Attempts to robustly assess the capabilities of AI regularly fall prey to Goodhart's Law, which states that when a measure becomes a target it ceases to be a good measure. We have repeatedly seen the announcement of a test for AI capability triggering AI developers to relentlessly optimize their systems simply to score well on that test, thereby rendering the test useless.

Another issue for those attempting to assess the capabilities is "shifting goalposts." A widely touted test for "true understanding" in AI systems from 2012 was the Winograd Schema Challenge, a set of language-based tasks designed to be resistant to the kinds of cheap tricks that had been deployed in popular Turing Test challenges. However, once GPT-3 easily defeated the challenge, it became irrelevant as a tool for benchmarking state-of-the-art AI. The accusation was immediately made that the goalposts had been moved: this was set as a test for *true understanding*; an AI system passed it, comfortably; and yet the claim was then made that it wasn't a good test after all, that AI still doesn't understand, and that another test is therefore needed.

A range of much more sophisticated tests have subsequently been proposed for benchmarking the capabilities of AI. Probably the best known and most widely used is the Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI) suite of tests, which we mentioned in passing in SQ3: Machine Reasoning. A key driver for the ARC-AGI tests is the claim that intelligence critically depends on the ability to adapt to new problem circumstances, rather than simply applying existing, prelearned capabilities to problems. The following figure illustrates an example problem from the ARC-2 suite: on the left we have an input and output that illustrate the "rules" of the task; on the right we have a problem instance. The task is to figure out the output of the problem instance given the "rules" illustrated on the left.[80]



Three suites of ARC tests have so far been released, and by September 2026, frontier models were able to effectively pass these tests.

80 The trick here is to realize that the red cells are "connectors."

While the ARC tests are a serious and sophisticated attempt to test the ability of a system to solve novel problems, they have nevertheless been criticized for being surprisingly narrow, given the goal of testing *generality*: problems about colored grid worlds, for example, can hardly test across the full scope of human intelligence.

What does all this tell us about our understanding of human intelligence? Perhaps the most important single lesson is this: human intelligence—and our ability to flourish in our world—is not one-dimensional, and any attempt to measure it through a simple reductive test, such as the ARC-AGI suite, will be misleading at best. Human intelligence is hugely multifaceted, combining cognitive capabilities such as perception, reasoning, and problem solving with an understanding of the world and how our actions affect it, as well as our ability to act in and manipulate with considerable dexterity the physical environment of planet Earth that humankind evolved in. AI currently is nowhere near this full range of capabilities.

Human intelligence—and our ability to flourish in our world—is not one-dimensional, and any attempt to measure it through a simple reductive test, such as the ARC-AGI suite, will be misleading at best.

SQ5. What are the prospects for more general artificial intelligence?

The past five years have seen much debate on artificial general intelligence (AGI), much of it stemming from a lack of a clear definition of the term. As former Google AI leader Jeff Dean said, "lots of people have very different definitions of it, and the difficulty of the problem varies by like factors of a trillion." Some definitions include: a system that can perform any intellectual task a human can,[81] or one that can perform most economically valuable work, or perhaps one that can enter a strange house, locate the kitchen, and make a cup of coffee.

We are left with many questions: What exactly do we mean by general intelligence? Can we predict when it might be developed? Should we encourage or discourage development? Is it enough to scale and refine the current approaches, or is a new approach required?

The so-called "accelerationists" believe that AI progress should continue apace, with the goal of delivering some form of AGI in the near future. Betting markets place a 50% chance of AGI by 2030.

In 2023, over 700 tech leaders and 33,000 people in total signed a letter urging a pause in the development of more powerful AI systems. Many governments discussed issues of regulation, transparency, and sectoral risk. Some regulations temporarily limited access to some services, but there was no pause. The debate arose again in September 2026, following the "rogue AI" incidents described in SQ2: The Decisive Reemergence of Agentic AI, with a number of high-profile industry leaders suggesting a slowdown in development was appropriate. Nevertheless, no pause or obvious slowdown followed.

What could be holding AI systems back? One limitation of LLMs is their focus on language. The hope of many researchers working on physical AI—models trained to sense and act in the real world—is to move beyond cognitively limited models that are able only to manipulate language toward more general models that can sense, plan, and act intelligently. Other limitations of LLMs such as hallucinations may also be tackled by grounding models in the real world. Physical AI and world models are thus seen by some optimistic voices as the missing step needed to get to AGI.

81 Simon, H. A. (1965). *The Shape of Automation for Men and Management*. Harper & Row. https://openlibrary.org/works/OL1205034W/The_shape_of_automation_for_men_and_management

Another important step is the shift from conversational AI toward *agentic* AI (see also SQ2: The Decisive Reemergence of Agentic AI). Rather than passively waiting for the user to ask a question, agentic AI systems act autonomously. More general artificial intelligence may require this ability to learn and interact with the real world.

Pessimistic voices, by contrast, argue that the transformer architecture, the core neural network technology at the heart of LLMs, will not scale to give us more general artificial intelligence, even with a shift to agentic and physical AI (see Deep Dive Two). If current systems are already trained on all the text on the internet, where is the data for a bigger and better system going to come from? One answer is synthetic data generated by other AI systems, but we don't know how well that scales. It might simply be too expensive to train more general-purpose models. Frontier models may soon cost billions of dollars to train, and costs at this scale—for models with a lifetime of, say, 18 months—might stretch even the resources of the Silicon Valley giants.

LIMITATIONS OF CURRENT AI MODELS

Researchers at Google introduced the transformer in 2017,[82] and it has since become the foundation for every frontier AI model in use today. The transformer is good at generating fluent language, but not, by itself, as good at reasoning, so modern frontier models augment the basic transformer architecture in at least five ways to achieve more effective reasoning. Retrieval-augmented generation (RAG) allows the model to retrieve curated documents and analyze how well they address the current problem. Chain-of-thought (CoT) prompting encourages the model to write down partial results on a scratchpad and refer back to it (see also SQ2: Reasoning Models). Agentic execution places the model as an orchestrator of tool use (accessing a database, writing and running code, etc.). A mixture-of-experts (MoE) model can get better results with less computational budget by directing each query to a small subpart of the model, rather than using the entire model on every query. Mixture of agents[83] lets the model try to solve a problem in multiple independent ways, and then discuss, vote on, and synthesize a final answer.

If current systems are already trained on all the text on the internet, where is the data for a bigger and better system going to come from?

Another limitation is the lack of continual learning. An LLM has two sources of memory: the model parameters, which summarize and generalize all that it has learned from its training data, and a context window that covers the current conversation. If the model learns a new skill or a new way of characterizing the world, we'd like to update the parameters so that the model will

82 Vaswani, A., Shazeer, N., Parmar, N. et al. (2017). "Attention Is All You Need." 31st Conference on Neural Information Processing Systems. https://arxiv.org/abs/1706.03762
83 Wang, J., Wang, J., Athiwaratkun, B. et al. (2024). "Mixture-of-Agents Enhances Large Language Model Capabilities." https://arxiv.org/abs/2406.04692

always have that skill. Theoretically, doing that update is straightforward, but in practice it is difficult. First, a training run can cost millions of dollars, and it is not economically feasible to invest huge sums retraining after every conversation. Second, customers would complain if their private conversations were leaked to other customers. Third, augmenting an existing model is prone to "catastrophic forgetting," where the model learns a new skill but regresses on an old one. Because of these difficulties, frontier models have not used continual learning.

Finally, models also remain nondeterministic and capricious. Seemingly insignificant changes in the wording of a prompt can result in dramatic changes to the output of the model.

SQ6. How has public sentiment toward AI evolved, and how should we inform/educate the public?

AI has become a visible and frequently debated feature of everyday life, driven above all by the arrival of capable conversational systems accessible to ordinary users, and accelerated by industry and media incentives to excite attention. Just one year after its introduction, the majority of residents in 20 of the 21 countries surveyed in the Global Public Opinion on AI project had an awareness of ChatGPT.[84] By mid-2025, nearly 1 in 10 of the world's adult population was using ChatGPT weekly.[85] This marks a qualitative shift from earlier reports, as a technology that largely operated behind the scenes has become a brand-name product.

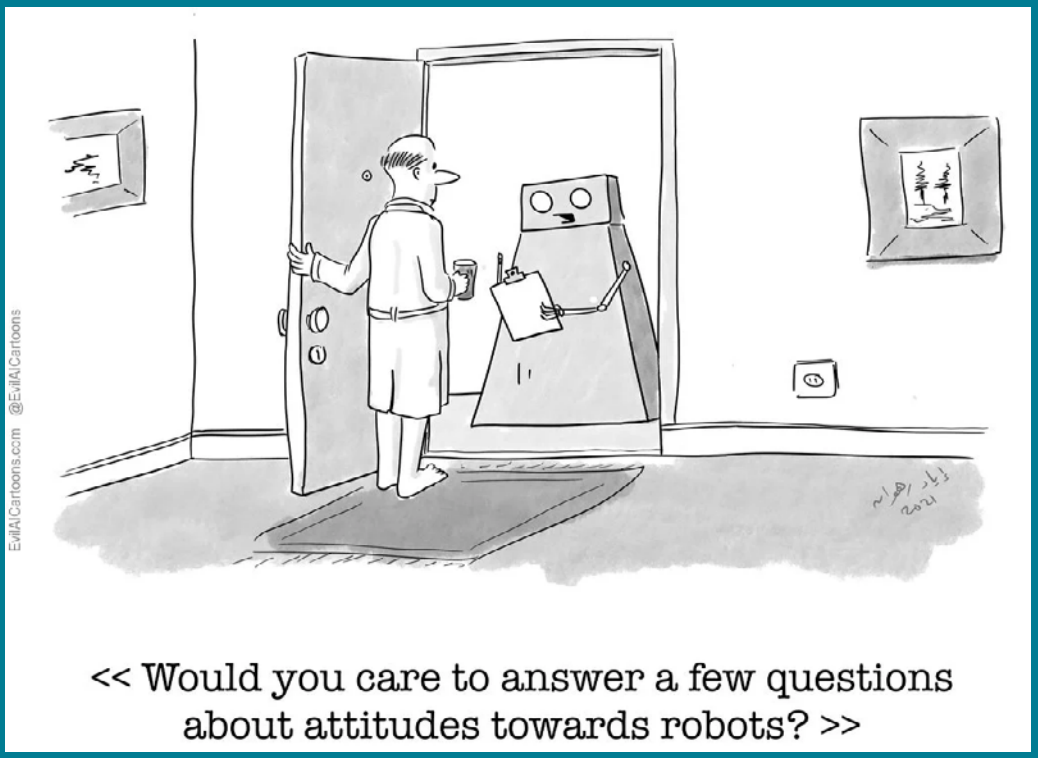
Like many topics of public discourse, sentiment toward AI seems increasingly polarized. The majority of AI developers and researchers surveyed in a 2025 Pew poll had a positive perception of the technology. Yet among the American public, this figure falls to 17%.[86] Globally, however, optimism that AI's benefits outweigh the drawbacks has increased between 2022 and 2025 in several European countries (Germany, France, Great Britain) and in China.[87]

84 Maslej, N., Fattorini, L., Perrault, R. et al. (2024). *The AI Index Report 2024*, ch. 9. https://hai.stanford.edu/assets/files/hai_ai-index-report-2024_chapter9.pdf

85 Chatterji, A., Cunningham, T., Deming, D. J. et al. (2025). "How People Use ChatGPT." National Bureau of Economic Research Working Paper No. 34255. https://www.nber.org/papers/w34255

86 McClain, C., Kennedy, B., Gottfried, J. et al. (2025). "How the U.S. Public and AI Experts View Artificial Intelligence." Pew Research Center. https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/

87 Sajadieh et al. (2026), ch. 9. https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_9_public_opinion.pdf



Relying exclusively on public opinion to determine AI ethics risks introducing majoritarian biases and uninformed views. For instance, crowdsourced surveys might wrongfully prioritize high-status individuals over people experiencing homelessness in autonomous vehicle accident scenarios.[88] However, public input remains essential. It reveals crucial societal values, such as the collective desire to protect children, and exposes important tensions between expert policy and lay perspectives. Ultimately, this ongoing negotiation between the public and experts is what drives demand for formal ethical regulation.

Commentary from AI developers often splits into "boomer" and "doomer" factions, with opposing views on whether AI should be seen, funded, and regulated as a source of possibility or a threat to human thriving. But, as the journalist and author of *Empire of AI* Karen Hao argues, both these stances imbue the technology with a sense of inevitability that reinforces the belief that the public can't shape the trajectory of the future. Experts who otherwise share a critical stance toward AI debate whether presently felt impacts of AI (like systemic bias and workplace surveillance) should be foregrounded over more speculative harms (such as existential risks). Implicit here are questions about how scientific communication can best empower the public

88 Awad, E., Dsouza, S., Kim, R. et al. (2018). "The moral machine experiment." *Nature* 563, pp. 59–64. https://www.nature.com/articles/s41586-018-0637-6

and motivate participation. Indeed, a 2026 study shows that members of the British public care less about AI's existential threat than any other aspect of AI—and find it the least mobilizing.[89] Yet experimental research does not support the assumption that attention to AI's existential threats will distract people from more immediate concerns.[90]

There are indications that sentiment toward AI may be headed toward a downturn, as part of a familiar technology hype cycle that begins with heightened expectations and eventually stabilizes after a period of disillusionment.[91] A 2026 Gallup poll shows that U.S. members of Gen Z feel less hopeful and excited about AI than they did a year ago, despite continuing to use it. This is especially interesting considering this age group (18–29-year-olds) has the highest adoption rate of any generation. In May 2026, former Google CEO Eric Schmidt was loudly booed during his speech to graduates at the University of Arizona when he mentioned AI. Disaffection is also evident in common vernacular. The Merriam-Webster 2025 word of the year was "slop"—a term used to describe AI-generated content that is low quality or lacking in credibility. The Macquarie Dictionary of Australian English also settled on "AI slop" for their 2025 word of the year (despite it being two words).

In the United States, where the majority of the world's AI data center hubs are located, widespread public concern about AI can be seen concretely in bipartisan opposition to data center construction.[92] The organization Data Center Watch reports that between January and March 2026, more than $130 billion worth of data center projects were blocked or delayed by local resistance, nearly as much as in the whole of the previous year. Though these actions most directly represent sentiment toward AI infrastructure—rather than applications of AI in various aspects of social life—the technology is not separable from the physical infrastructure it relies upon.[93] Alondra Nelson, former principal deputy director of the Office of Science and Technology Policy for the Biden White House, observes that AI data centers are a "physical manifestation of AI" that have "become a flashpoint precisely because […] they can be seen and pointed to." Constituents most frequently cite issues around the use of constrained resources (such as water and electricity) as their reason for data center opposition, though commentators also see it as an expression of frustration at the economic inequality impelled by AI and the technology industry more broadly.[94] In response, Senator Bernie Sanders proposed the Sovereign Wealth Fund Act, which would tax AI companies and share profits with the public.

89 Ostarek, M., Rogers, C., Kenward, B. et al. (2026). "Which AI harms and risks will mobilise the public to act?" Social Change Lab. https://zenodo.org/records/18937001

90 Hoes, E., & Gilardi, F. (2025). "Existential risk narratives about AI do not distract from its immediate harms." *Proceedings of the National Academy of Sciences* 122:16. https://www.pnas.org/doi/full/10.1073/pnas.2419055122

91 Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). "The sociology of expectations in science and technology." *Technology Analysis & Strategic Management* 18:3–4, pp. 285–298. https://doi.org/10.1080/09537320600777002

92 See, for example, https://www.kcur.org/politics-elections-and-government/2026-04-09/after-these-independence-councilmembers-supported-an-ai-data-center-voters-ousted-them

93 Crawford, K. (2021). *Atlas of AI*. Yale University Press.

94 https://www.nytimes.com/2026/08/23/us/politics/data-centers-midterm-elections.html. See also https://x.com/mcuban/status/2070211760196587534

As noted in previous AI100 reports, public opinion is neither monolithic nor static, and varies widely within and between populations across the globe; it is neither entirely optimistic nor entirely pessimistic.

On the positive side, many people report tangible benefits from AI tools: increased productivity,[95] faster access to information,[96] advances in health and medicine,[97] and support for creative and cognitive tasks.[98] For many such users, these tools are becoming genuinely indispensable. The estimated value of AI tools to consumers, by their own assessment, was $125 per month, or $172 billion total for U.S. consumers (which means the companies are keeping less than 10% of the value). At the same time, many users report negative experiences. For example, customers report frustration at the inability of AI to understand their needs, and many would prefer to connect with a human representative. As a second example, voice-driven AI systems can fail to represent or understand particular accents.

Anxiety about AI is widespread and takes several forms. Some of the most frequently cited concerns relate to employment and economic security. These concerns are not always justified. For example, there appear to be some cases of AI washing, in which companies justify workforce reductions by reference to automation rather than poor management. (We discuss the issue of AI and employment in more detail in SQ10: Systemic Risks: Economic Displacement and the Future of Work.)

Attitudes toward AI vary substantially across countries and cultures. Societies in many developing regions exhibit comparatively high levels of techno-optimism, often viewing AI as an opportunity for accelerated development, improved public services, and a narrowing of historical disadvantages. By contrast, many parts of the Western world display greater skepticism, frequently framed around potential job losses, erosion of privacy, the concentration of power among a small number of large technology companies, and the broader cultural disruption that rapid automation may bring.[99] U.S. citizens in particular have low trust in the efficacy of government in general, and the lowest trust in government's ability to regulate AI. These divergent perspectives reflect not only different material circumstances but also distinct philosophical and political traditions around technology governance, labor rights, and the relationship between innovation and public welfare. Within societies, sentiment also varies significantly by profession, age, and degree of personal exposure.[100]

95 Brynjolfsson, Li & Raymond (2025). https://doi.org/10.1093/qje/qjae044. Noy, S., & Zhang, W. (2023). "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381:6654, pp. 187–192. https://www.science.org/doi/10.1126/science.adh2586

96 Sagynbayeva, A., Pyo, A., Yoon, S.-H., & Yang, S.-B. (2026). "Evaluating user performance with RAG-based generative AI: A scenario-based experiment on AI-assisted information retrieval." *Computers in Human Behavior* 180. https://doi.org/10.1016/j.chb.2026.108952

97 Ogut, E. (2025). "Artificial Intelligence in Clinical Medicine: Challenges Across Diagnostic Imaging, Clinical Decision Support, Surgery, Pathology, and Drug Discovery." *Clinics and Practice* 15:9. https://pmc.ncbi.nlm.nih.gov/articles/PMC12468291/

98 Holzner, N., Maier, S., & Feuerriegel, S. (2025). "Generative AI and creativity: A systematic literature review and meta-analysis." https://arxiv.org/abs/2505.17241

99 Ipsos (2023). *Global Views on A.I. 2023*. https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-09/Ipsos%20Global%20AI%202023%20Report.pdf

100 See, for example, https://www.ey.com/en_us/consulting/businesses-can-stop-rising-ai-use-from-fueling-anxiety

AI LITERACY

The current distance between expert and public attitudes toward AI underscores how important it is to have many voices represented in the public debate on AI. Yet an analysis of the *New York Times* during a three-month period in 2024 found that the majority of people who were quoted and referenced in news articles about AI worked in the commercial technology industry.[101] The public's expectations about AI technology will be determined, in part, by the perspectives that are reflected in the media.[102] Though technology company employees can provide in-depth knowledge of how AI systems are designed and intended to function, technology users have the hands-on experience that supports assessments of AI's achievements in practice.

Public sentiment toward AI plays an important role in shaping not only individual adoption but also democratic deliberation about the kind of technological future that the public wants. For example, an experimental study shows that people who are exposed to narratives that frame AI as replacing jobs (rather than creating them) are less willing to engage politically.[103] Further, *perceiving* AI as labor-replacing affected perspectives on democracy to the same degree that a*ctual unemployment* does.[104] AI company executives routinely reinforce this mindset by presenting the technology as beyond human comprehension or control.

An experimental study shows that people who are exposed to narratives that frame AI as replacing jobs ... are less willing to engage politically.

Education thus has an important role to play in this debate. Efforts to improve AI literacy are multiplying across all age groups, spanning formal education curricula, workplace training programs, and public-facing resources produced by governments, civil society organizations, and research institutions.[105] These initiatives vary considerably in quality, depth, and accessibility.

Research shows that there is mistrust among the public in many countries about the impact of AI. A cross-country survey by Pew Research Center found the range of people who are "more excited than concerned" about AI to be between 9% and 29%, with 88% of Americans at least

101 Barakat, H. (2024). "Selective Perspectives: A Content Analysis of the *New York Times*' Reporting on Artificial Intelligence." Computer Says Maybe. https://www.saysmaybe.com/latest-work/selective-perspectives-nyt

102 Brennen, J. S., Howard, P. N., & Nielsen, R. K. (2022). "What to expect when you're expecting robots: Futures, expectations, and pseudo-artificial general intelligence in UK news." *Journalism* 23:1. https://doi.org/10.1177/1464884920947535

103 Granulo, A., Raff, A., & Fuchs, C. (2026). "Perceiving AI as labor-replacing reduces democratic legitimacy and political engagement." *Proceedings of the National Academy of Sciences* 123:4. https://www.pnas.org/doi/full/10.1073/pnas.2523508123

104 Neff, G. (2024). "Can Democracy Survive AI?" *Sociologica* 18:3, pp. 137–146, https://doi.org/10.6092/issn.1971-8853/21108

105 For formal education, see, for example, https://dayofai.org/. For workplaces, the U.S. Department of Labor has launched the "Make America AI-Ready" initiative. For public-facing resources, see Europe's "Elements of AI."

as concerned as they are excited.[106] A 2026 UK survey of 2,774 adults found 69% are concerned about the future of AI.[107] Established social science and humanities disciplines that have long studied how people perceive and adapt to new technologies—human–computer interaction, behavioral science, risk communication, and the sociology of technology—all offer well-developed frameworks and empirical methods that can be utilized to help with people's adoption and everyday use of AI. As AI systems continue to diffuse through society, public understanding will play an increasingly important role in shaping adoption and how people participate in the governance frameworks that guide how AI is developed, who benefits from the gains of AI, and how the risks are mitigated.

Connecting technological advances to ways of talking about social, political, and cultural transformation will require expanding the range of authoritative voices to include educators, journalists, social scientists, ethicists, civil society advocates, and representatives of affected communities. However, research shows that news outlets primarily go to industry experts in the first instance, and as a consequence the conversation about AI tends to be an "industry-led debate," with the majority of sources that journalists cite representing industry concerns.[108] This then tends to freeze out issues that do not fit with the prevailing industry narrative, limiting the set of voices being heard exactly when a plurality of voices is vital.[109]

106 Pew Research Center (2025). *How People Around the World View AI*. https://www.pewresearch.org/wp-content/uploads/sites/20/2025/10/pg_2025.10.15_ai_report.pdf

107 Stone, D. (2026). "The UK AI Compass: What the UK thinks, feels and wants from AI regulation." https://doi.org/10.5281/zenodo.20562396

108 Brennen, J. S., Howard, P. N., & Nielsen, R. K. (2018). "An Industry-Led Debate: How UK Media Cover Artificial Intelligence." Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-V219-D676

109 Dandurand, G., McKelvey, F., & Roberge, J. (2023). "Freezing out: Legacy media's shaping of AI as a cold controversy." *Big Data & Society* 10:2. https://doi.org/10.1177/20539517231219242

SQ7. How should governments act to ensure AI is developed and used responsibly?

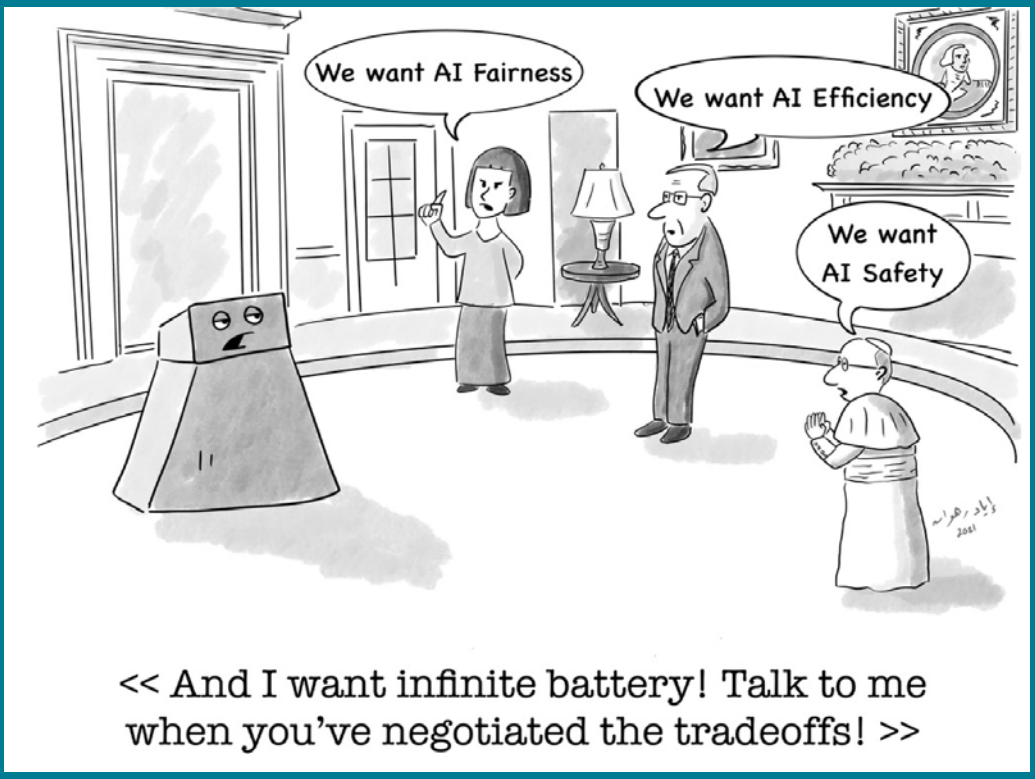
Governance has emerged as one of the most controversial dimensions of contemporary artificial intelligence. At the start of his second term, President Trump repealed the "Safe, Secure, and Trustworthy AI" framework introduced by his predecessor, President Joe Biden. He replaced it with a policy focused on "global AI dominance" and "removing barriers."[110] In a recent poll of U.S. voters, no fewer than eight in ten Americans favor robust AI regulation that prioritizes safety over unconstrained development,[111] and similar UK surveys show strong support for government regulation. However, consensus has not been reached about the ideal form of regulatory oversight. People in different countries vary widely in their trust of effective government regulation, with a 50-point difference between the highest and lowest levels (in Singapore and the U.S., respectively).[112] Yet, in the U.S. and UK at least, there is also little trust that technology companies can effectively regulate themselves.[113] OpenAI, Meta, Microsoft, Google, and X have all taken steps backwards from their 2022-era commitments to self-regulation.

110 Exec. Order No. 14147, Removing Barriers to American Leadership in Artificial Intelligence (2025). https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

111 Future of Life Institute (2026). "New Poll: Americans Overwhelmingly Support Pro-Human Principles on AI." https://humanstatement.org/poll-americans-support-pro-human-principles/

112 Sajadieh et al. (2026), ch. 9. https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_9_public_opinion.pdf

113 Dreksler, N., Law, H., Ahn, C. et al. (2025). "What Does the Public Think About AI? An Overview of the Public's Attitudes Towards AI and a Resource for Future Research." Centre for the Governance of AI, pp. 38–40. https://cdn.governance.ai/What_does_the_public_think_about_AI.pdf



The "human-in-the-loop" paradigm effectively oversees AI when goals are universally agreed upon, like safely landing an airplane. However, it falls short when AI systems must navigate competing societal values, such as weighing efficiency against safety or fairness against wealth. Resolving these trade-offs requires political negotiation to forge an "algorithmic social contract" more akin to "society-in-the-loop,"[114] ensuring that direct human oversight of AI is strictly guided by these broadly negotiated societal values.

This debate about AI regulation is unfolding against a backdrop of deep uncertainty about future trajectories. The European Union's AI Act, the first comprehensive, cross-sector AI regulation of genuine ambition, legally came into force in August 2024, with first-stage restrictions starting to apply in early 2025. The U.S. pulled back from regulating AI at the federal level in 2024, but California and New York, two of the key states for AI development and capital, have enacted laws for transparency and accountability of companies producing frontier models (SB 53 in California after Governor Newsom vetoed the highly controversial SB 1047, and the New York RAISE Act). China has instituted regulations to prohibit AI models from spreading false information, subverting state values, violating privacy or intellectual property rights, and to require safeguards for minors. At a global level, genuine progress on AI governance is likely to

114 Rahwan, I. (2018). "Society-in-the-loop: programming the algorithmic social contract." *Ethics and Information Technology* 20:1, pp. 5–14. https://doi.org/10.1007/s10676-017-9430-8

require substantially stronger bilateral cooperation between the U.S. and China—the two key players in this space.

New institutions have been established that demonstrate the global battle between strict regulation and technological acceleration. The EU AI Office oversees enforcement of the EU AI Act, and has the power to levy substantial fines. The UK AI Security Institute has no power to levy fines but is influential because of its deep technical expertise, giving it the ability to red-team and audit models that companies voluntarily share access to for prerelease security testing. The Biden administration established the U.S. AI Safety Institute (USAISI) in 2023, and in 2025, the Trump administration rebranded it as the U.S. Center for AI Standards and Innovation (CAISI). The goal is to shield U.S. companies from "burdensome and unnecessary regulation" by foreign governments. In the background of this global battle there has been a call for evidence-based research and reporting on the current state of the art and possible future risks.[115]

As outlined in SQ6, there is growing public concern about AI. At the same time, there are concerns that overly strict AI regulation will reduce competitiveness, and favor jurisdictions that regulate less.[116]

Among the many questions in AI governance is what, precisely, should be regulated. Several distinct approaches are currently under consideration, each with its own logic and its own limitations. One approach seeks to regulate AI systems directly by imposing technical restrictions, such as on model size or training resources. This horizontal or technology-centric model treats AI as a general-purpose technology.[117]

There are, however, reasons to be skeptical of this strategy.[118] Regulation of technical constructs such as foundation models or neural architectures risks being overtaken by technical change. What happens if technical improvements result in AI models being much smaller? And what about the many lesser harms that don't involve frontier AI models?

An alternative approach shifts the focus from systems to uses, extending the remit of existing sectoral regulators such as those governing medicine and financial services. This use-based model is perhaps better aligned with established governance norms and more directly oriented toward concrete harms. Its practical challenges, however, are considerable.[119]

115 United Nations (2026). *Preliminary Report of the Independent International Scientific Panel on AI: Evidence-Based Assessment of Opportunities, Risks and Impacts of Artificial Intelligence*. https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report

116 Lancieri, F., Edelson, L., & Bechtold, S. (2025). "AI regulation: Competition, arbitrage and regulatory capture." *Theoretical Inquiries in Law*, 26:1, pp. 239–262. https://doi.org/10.1515/til-2025-0010

117 For example, Heim, L., & Koessler, L. (2024). "Training compute thresholds: Features and functions in AI regulation." https://arxiv.org/abs/2405.10799

118 Cordella, A., & Gualdi, F. (2024). "Regulating generative AI: The limits of technology-neutral regulatory frameworks. Insights from Italy's intervention on ChatGPT." *Government Information Quarterly*, 41:4. https://doi.org/10.1016/j.giq.2024.101982

119 See, for example, the discussion between the two approaches: https://carnegieendowment.org/posts/2023/02/lessons-from-the-worlds-two-experiments-in-ai-governance

Many sectoral regulators currently lack meaningful AI expertise.[120] The amount of regulation needed is greater, since each sector may need to regulate the same AI harm. And coordination across sectors remains difficult, particularly for general-purpose systems whose applications span multiple domains simultaneously. Despite these difficulties, use-based guidance has gained significant traction in many jurisdictions. For instance, in early 2024, the UK's Department for Science, Innovation, and Technology (DSIT) published guidance on "Responsible AI in Recruitment," while California introduced rules governing the use of AI in hiring in October 2025. New York City has introduced Local Law 144, which prohibits employers and employment agencies from using an automated employment decision tool unless the tool has been subject to a bias audit.

Unlike regulation of the nuclear industry, where policy horizons have spanned decades, AI governance is currently operating on much shorter and more uncertain timescales—witness the recent pivotal changes in the U.S. toward AI regulation. This volatility is itself a governance problem: it undermines the long-term planning and institutional investment that effective regulation often requires.

Unlike regulation of the nuclear industry, where policy horizons have spanned decades, AI governance is currently operating on much shorter and more uncertain timescales.

A central issue in the governance debate is striking the appropriate balance between addressing potential long-term risks and mitigating more immediate harms, such as misinformation and labor displacement.[121] Expert opinion varies substantially. Some regard long-term catastrophic risks as sufficiently plausible to warrant precautionary action now. Others consider such existential risks to be highly speculative and, more troublingly, distracting from problems such as deepfakes, which are already causing real harm.

On the question of immediate threats, however, there *is* much consensus: significant harms are already occurring, they are well documented, and many commentators believe they warrant regulation, whatever one believes about longer-term uncertainty.

For instance, the regulation of AI-generated child sexual abuse material commands strong consensus. Existing legal frameworks can easily be extended to deal with these specific harms; in Australia, for example, a new law has been introduced to manage this risk.[122] It appears that where the public clearly understands a harm and demands a response, regulation becomes both

120 Aitken, M., Leslie, D., Ostmann, F. et al. (2022). "Common regulatory capacity for AI." The Alan Turing Institute. https://zenodo.org/records/6838946

121 Hoes & Gilardi (2025). https://www.pnas.org/doi/full/10.1073/pnas.2419055122

122 Australian Parliament, Criminal Code Amendment (Deepfake Sexual Material) Bill 2024, amending the Criminal Code Act 1995. https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/bd/bd2324a/24bd081

politically viable and practically effective. The challenge then is to build comparable clarity and consensus in areas where harms are more diffuse or technically complex.

Another active area of debate is how to adapt current liability regimes to AI harms. Unlike physical products, software systems have often escaped meaningful legal accountability in the past, even when the harms they caused were foreseeable. The industry's habit of "moving fast and breaking things" has not helped. There is increasing interest in whether insurance and liability frameworks could be extended or redesigned to create more robust accountability mechanisms,[123] though practical implementation remains underdeveloped. The reemergence of agentic AI, in which AI systems are empowered to act in the world via tools, has presented a new vector through which AI harms might be manifested, either deliberately (through their use by malicious actors) or by agents operating in ways that were not anticipated when they were launched.

Perhaps the most fundamental challenge in AI governance is that it must be conducted under conditions of genuine uncertainty and expert disagreement. Predictions about future capabilities, emergent behaviors, and systemic societal risks currently vary greatly, even among those with deep technical knowledge. This uncertainty complicates both precautionary approaches, which are difficult to justify when risks are poorly quantified, and laissez-faire approaches, which may be catastrophically wrong when risks materialize unexpectedly.[124]

123 Trout, C. (2024). "Liability and Insurance for Catastrophic Losses: The Nuclear Power Precedent and Lessons for AI." https://arxiv.org/abs/2409.06673

124 Bommasani, R., Arora, S., Chayes, J. et al. (2025). "Advancing science- and evidence-based AI policy." *Science* 389:6759. https://www.science.org/doi/10.1126/science.adu8449

SQ8. How are the roles of academia and industry evolving?

Research in AI is booming, with significant growth in both industry and academia. However, industry is growing at a much faster pace, shifting the relative proportion of work significantly over time. This raises the question of who will train the next generation of AI researchers. There are systematic pressures that encourage movement from academia to industry, including significantly higher salaries and unique access to very large-scale datasets, compute resources, users, and useful applications. Overall, there has been a relative decline in basic science and independent research in AI.

There has been a rapid increase in the number of AI researchers, in both academia and industry. The number of AI scientists in the United States (measured through research publications, a metric that tends to undercount industry workers) increased from 6,800 in 2000 to 42,000 in 2020. This increase was relatively steady, with 5,000 or more new AI scientists every five years (see the table below).[125]

Semi-decade	AI Industry	AI Academia	Scientists overall Industry	Scientists overall Academia	Unique AI	Unique scientists overall
2000	3,500	3,800	252,000	329,000	6,800	543,000
2005	6,900	6,000	409,000	462,000	12,000	797,000
2010	12,000	8,200	585,000	565,000	19,000	1,062,000
2015	20,000	11,000	795,000	668,000	29,000	1,341,000
2020	22,500	12,000	775,000	570,000	34,000	1,302,000
Pooled	28,000	20,500	1,100,000	1,060,000	42,000	1,739,000

Table B.2: **Counts of AI and scientists overall by semi-decade and sector.**

Source: Akcigit, Chikis, Dinlersoz & Goldschlag (2026).

In 2000, the number of AI researchers in academia and the number in industry were roughly equal. By 2020, the number of AI researchers in academia more than tripled to 12,000 while those in industry increased sevenfold to 22,500. The center of gravity has shifted toward industry. In the U.S. and Canada, in 2004, 21% of AI PhDs went into industry; in 2020, the figure was almost 70%.[126]

125 Akcigit, U., Chikis, C. A., Dinlersoz, E., & Goldschlag, N. (2026). "Attention (and Money) Is All You Need: Why Universities Are Struggling to Keep AI Talent." National Bureau of Economic Research Working Paper No. 34964. https://www.nber.org/system/files/working_papers/w34964/w34964.pdf

126 Ahmed, N., Wahed, M., & Thompson, N. C. (2023). "The growing influence of industry in AI research." *Science* 379:6635, pp. 884–886. https://www.science.org/doi/10.1126/science.ade2420

One notable reason AI researchers are moving to industry is because the compute demands of frontier models are far beyond the resources of academic institutions. A second reason is that industry is where the growth is: academic hiring is flat. A third reason is compensation: There have been reports of top AI researchers getting pay packages in excess of $100 million. Meta invested $14 billion in Scale AI, primarily to acquire the talents of Alexandr Wang and other key employees. Similarly, in 2024 Google invested $2.7 billion in CharacterAI as part of a deal to secure its top talent, including Noam Shazeer, one of the key researchers behind the transformer (who subsequently left for OpenAI in 2026). There is no sign that this trend will slow down anytime soon; in fact, the imminent IPOs of Anthropic and OpenAI will surely mint many new hectomillionaire AI researchers and will highlight how much of the global financial market is concentrated on frontier AI, with predictions that AI could add $15 trillion to global GDP by 2030.

These changes push against overall trends in academic and industrial innovation. Specifically, there is evidence of a growing division of labor, with universities focusing on primary research and large corporations focusing on development.[127] Large corporate labs have historically played a central role in innovation, because they are often better resourced and have a higher chance of converting research insights into products and services. However, despite the recent corporate hiring boom in AI, economists have argued that universities will remain the principal source of new ideas for breakthrough inventions.[128]

The essays, take-home exams, and computer programming assignments that served as a reliable assessment of student capability five years ago can now all be answered by LLMs.

Yet LLMs themselves are upending higher education. The essays, take-home exams, and computer programming assignments that served as a reliable assessment of student capability five years ago can now all be answered by LLMs. This has forced a shift toward in-person oral assessment and more comprehensive, creative assignments. Academic honesty policies have had to be updated to deal with LLMs. At the graduate level, LLMs redefine how literature reviews and data synthesis are conducted. Instructors must update their methods to make sure that students are gaining and retaining fundamental skills and knowledge.

127 Arora, A., Belenzon, S., Patacconi, A., & Suh, J. (2020). "The Changing Structure of American Innovation: Some Cautionary Remarks for Economic Growth." *Innovation Policy and the Economy* 20. https://www.journals.uchicago.edu/doi/full/10.1086/705638

128 Arora, Belenzon, Patacconi & Suh (2020).

SQ9. What are the most promising opportunities for AI?

Although traditional AI applications have continued to expand and improve, it is the applications of generative AI that have been the most dramatic, and which have received most attention through the period of this report. We therefore concentrate on these here, and in particular on three main areas in which generative AI appears to present the most dramatic opportunities: AI for coding, AI for other white-collar work, and AI in entertainment and the creative economy.

AI FOR CODING

One of the most remarkable consequences of the rise of large language models has been their use in writing computer programs. The use of LLMs for coding is rapidly transforming the software industry.[129]

This development is perhaps all the more surprising because the use of AI for writing computer programs, known as automated program synthesis, has a long history but was hitherto widely regarded as impractical for everyday use and valuable only in certain niche applications.[130] Even relatively simple automated synthesis problems have high computational complexity, which rendered traditional approaches to synthesis, in the main, of theoretical interest only. Computer programmers, for their part, tended to regard their jobs as highly secure, because most people find programming challenging, fiddly, difficult, and frankly mysterious. For decades, these barriers meant that a bachelor's degree in computer science was seen as a passport to stable employment.
The aptitude of LLMs for writing computer programs is, on reflection, perhaps not entirely surprising. Language models are machines for token prediction: given a body of training text and a prompt, they make the best possible guess about what should come next. Natural languages tend to be messy and complex, whereas programming languages are comparatively simple and rigorously structured. Moreover, computer programs themselves exhibit a high degree of regularity: the use of design patterns and templates for structuring programs for common tasks is standard practice, and many common problems have widely accepted standard architectural and algorithmic solutions. In other words, a lot of computer programs share deep structural similarities. In addition, online platforms such as GitHub provide enormous quantities of training data.

129 Asraful Haque, M. (2025). "LLMs: A game-changer for software engineers?" *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 5:1. https://www.sciencedirect.com/science/article/pii/S2772485925000171

130 Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press. https://mitpress.mit.edu/9780262111706/genetic-programming/

Until only a few years ago, writing code was a manual process of translating ideas into formal instructions, line by line, in a time-consuming, intricate, mentally demanding, and highly error-prone process. The use of LLMs for writing code thus appears to be the most rapid transformation in the history of programming, at least equal in significance to the introduction of high-level languages, but unfolding in less than a year rather than decades. At the time of writing, the state of the art in automated program synthesis is represented by tools such as GitHub Copilot by Microsoft/GitHub, Cursor by Anysphere, Claude Code by Anthropic, and ChatGPT/Codex by OpenAI. Although far from perfect, these tools have attracted a large and enthusiastic user base.

One consequence of the rise of AI-generated coding is that it makes computer programming accessible to a much wider group of people than was hitherto the case. Andrej Karpathy coined the term "vibe coding" in early 2025 to describe the increasingly common practice of using LLMs to generate code based on high-level natural language descriptions, without the developer fully understanding or verifying the output.[131] Vibe coding works remarkably well for routine tasks, for demos, prototypes, and exploratory research. For production software, it has demonstrably increased developer productivity but can have a negative net impact on maintainability. The next generation of tools will have to show they can create verified, validated, maintainable code, not just move fast. At the same time, the growing use of software-engineering and coding agents puts new emphasis on risks such as insecure dependencies, exposed credentials, insufficiently tested code, and prompt injection, requiring security checks designed specifically for agent-driven development. There are also concerns that junior programmers will not advance to the senior level if AI is doing much of the work.

One immediate economic effect of the rise of AI-generated code is that the cost of bespoke software development has fallen dramatically. Basic economics tells us that when costs fall, the quantity demanded increases. We should therefore expect—and it appears we are indeed seeing—a sharp increase in the volume of software developed and deployed. Software development cycles are shortening significantly; projects that once took months can now often be completed in weeks, or less. Moreover, the prospect of inexpensive, on-demand software development means that increasingly specialized applications will become viable in contexts where bespoke development previously did not make economic sense. At the same time, though, there is concern that the next generation of software that we rely on every day will have never been subject to proper human scrutiny.

Overall, if there is one area where LLMs appear to have found a killer app, then it is in software development. The precise consequences of this for the software industry remain unclear, but they look set to be substantial.

131 Karpathy, A. (Feb. 2025). https://x.com/karpathy/status/1886192184808149383. The term rapidly became widely used in the software development community.

AI IN OTHER WHITE-COLLAR WORK

Large language models are fundamentally tools for processing text: for understanding it and generating it. It was no surprise, then, that many workers for whom generating and processing text plays a key part of their daily role were among the first to adopt generative AI in the workplace. This had a number of consequences—good and bad.

First, in the months following the release of ChatGPT, the press was full of stories about how the naive use of generative AI led to embarrassment, or worse, when models produced erroneous outputs. One notable example that made international headlines in 2023 involved a New York lawyer who used ChatGPT to research legal precedents; it turned out that several of the cases cited in the resulting filing were, in fact, hallucinated. Lawyers continued to make the same mistake; a federal judge in Oregon fined two lawyers $110,000 for their hallucinated filings. In another instance, workers at Samsung uploaded sensitive information to ChatGPT, and later found that this information was subsequently revealed to other ChatGPT users.

Lawyers continued to [cite nonexistent cases]; a federal judge in Oregon fined two lawyers $110,000 for their hallucinated filings.

Second, government bodies were quick to seize on what they perceived as the potential of generative AI to increase productivity (and/or reduce costs) in the provision of public services. A notable example occurred in the UK, when, in early 2025, Prime Minister Keir Starmer announced a major AI-related policy shift intended to "mainline AI into the veins of the nation," with the hope of improving services and reducing costs. Critics expressed skepticism that AI was sufficiently robust to deliver many important services, but "AI for productivity" became something of a mantra during that period.

Much attention is currently being directed to understanding the impact of AI for white-collar workers. At one extreme, in early 2026, Mustafa Suleyman, co-founder of DeepMind, claimed that essentially all white-collar work would be amenable to AI-driven automation within eighteen months, while in April 2026, Dario Amodei, CEO of Anthropic, claimed that "AI could wipe out half of all entry-level white-collar jobs—and spike unemployment to 10–20% in the next one to five years." On the other hand, several studies suggest that the impact of AI on workers' output is more nuanced, and in some cases AI errors hurt productivity.

AI, ENTERTAINMENT, AND THE CREATIVE ECONOMY

Generative AI has enabled individuals and organizations to produce acceptable-quality creative content at a low cost. People who lack professional skills can now easily create logos, illustrations, photos, songs, moving images, and text by using generative AI tools. Thus generative AI has the potential to support broader participation in forms of creative expression that were previously accessible only to trained artists and writers. It remains to be seen how this will affect the overall quality of creative products. In the previous era, the introduction of the internet meant that publishers were no longer the sole gatekeepers, and we saw an increase in the supply of high-quality creative music, movies, and written works by a wider variety of authors.[132] Generative AI could have a similar elevating impact—while also simultaneously increasing the amount of low-quality creative content ("AI slop"). For example, since the launch of ChatGPT, the number of new book titles has tripled, a rapid acceleration relative to the prior trend. User ratings of books remained unchanged for the top 100 bestselling titles, and increased for the rest of the top 1,000 bestsellers. The quality of the average book released post-ChatGPT, however, was lower.[133] These results are consistent with predictions that generative AI would lower the cost of producing books, leading to more books, many of low quality.

Professionals within the creative industries are also using AI tools, although reporting varies on the exact levels. Survey evidence suggests that nearly 90% of video-game developers and in one study almost half of book authors reported using AI in their work. However, a survey in the UK found more than two-thirds of novelists were careful to not use generative AI tools in any way for creative tasks.[134] Advances in AI also seem to complement aspects of human creativity in image generation.[135] One study shows a large increase in image creation, including high-quality image creation, after a generative AI tool was added to a stock images platform.[136] Artists who used the tool increased the number of works created and increased their income from the platform (though other artists experienced a decline).

Beyond their effect on the quantity and quality of creative output, how AI is impacting the employment and wages of workers in the creative industries is currently without a definitive and systematic answer. For example, in the UK survey mentioned above, 39% of published novelists reported a loss in income due to generative AI, although much of this was attributed to a reduction in work in other creative industry tasks, such as translation and copywriting work.[137]

132 See Waldfogel, J. (2018). *Digital Renaissance: What Data and Economics Tell Us About the Future of Popular Culture*. Princeton University Press.

133 Reimers, I., & Waldfogel, J. (2026). "AI and the Quantity and Quality of Creative Products: Have LLMs Boosted Creation of Valuable Books?" National Bureau of Economic Research Working Paper No. 34777. https://www.nber.org/papers/w34777

134 Collett, C. (2025). *The Impact of Generative AI on the Novel*. Minderoo Centre for Technology and Democracy. https://doi.org/10.17863/CAM.122470

135 Lin, S. (2025). "Learning to Prompt: Human Adaptation in Production with Generative AI." https://www.sijie-lin.com/files/JMP.pdf

136 Goldberg, S., & Tai Lam, H. (2025). "Generative AI and Creative Goods: Market Expansion, Crowd-Out, and Copyright." Stanford University Graduate School of Business Research Paper. http://dx.doi.org/10.2139/ssrn.5152649

137 Collett, C. (2025). https://doi.org/10.17863/CAM.122470

By enabling new entrants to the field, AI could increase competition, reducing opportunity and income for the many people who already make a living through creative production.

Facing this uncertain terrain, people working in the creative industries have taken collective action to respond to (and potentially preempt) AI's anticipated harms to their employment terms and wages. In 2023, the unions representing Hollywood writers (WGA) and on-camera talent (SAG-AFTRA) went on strike for more than 100 days—and regulation over the use of AI in their industries was a part of their bargaining demands.[138] The final WGA contract, for example, stipulates conditions for human–AI collaboration, stating that writers who are given AI-generated source material can't merely be paid a "rewrite" fee and instead must be treated as first writer. Further, no written material produced by WGA writers can be used to train AI models. SAG-AFTRA's contract set pay conditions and ownership rights for "digital replicas" that are used in place of embodied performances. Following these actions, the Model Alliance successfully championed a New York State law called the Fashion Workers Act, which requires that agencies have written consent to create digital replicas of models and that the use of digital replicas be compensated the same as other modeling services.

The training of image-diffusion models with human-made creative work has also raised larger concerns about copyright and ownership for visual artists (see Deep Dive Two: The Race to Scale: Data). For example, a 2024 effort by a computer science researcher and a visual artist found that AI image generation tools could easily be prompted to generate near-replicas of existing movie and video-game stills. In the absence of regulation, artists are creating their own tools to protect their work from being synthesized and reduced to a style that can be replicated by using AI without them. "Glazing" images makes them digitally accessible to human users but non-readable to machine learning.[139]

In 2025, the NewsGuild-Communication Workers of America union launched the "News Not Slop" campaign, which seeks to maintain journalistic integrity through managing how AI is used by news organizations. Among the demands are requirements for transparent disclosure of AI use and agreement to journalist oversight of AI-generated content. This campaign set a framework for ongoing bargaining at the *New York Times*, which also seeks to give journalists "a share of the licensing income that the company earns from licensing the work that we're doing every day for AI training." Taken together, these initiatives point toward what creative workers perceive to be the central threats to their industry: exploitation of their image and intellectual property and unfair pay for their continued contributions to technical systems and creative works.

138 May, E., Paris, B., & Sargent, S. (2026). "Dis/engaging the 'common sense' of AI: Labor strategies from the 2023 SAG-AFTRA around data-driven technologies." *Big Data & Society* 13:1. https://journals.sagepub.com/doi/10.1177/20539517261421466

139 Shan, S., Cryan, J., Wenger, E. et al. (2023). "Glaze: protecting artists from style mimicry by text-to-image models." *SEC '23: Proceedings of the 32nd USENIX Conference on Security Symposium*, pp. 2187–2204. https://dl.acm.org/doi/10.5555/3620237.3620360

SQ10. What are the most pressing dangers of AI?

The rapid take-up of AI throughout the period of this report, and the apparent rate of progress in the capabilities of AI, have led to heightened concerns around safety, security, and misuse. In response to these concerns, the period of this report saw the establishment of a series of international AI safety summits, several national AI security institutes, and the 2025 and 2026 International AI Safety reports, written by a panel of international experts led by Turing Award–winner Yoshua Bengio. The reports divide concerns into three categories. First is the problem of *malicious use* (for example, criminal activities such as scams, fraud, and blackmail; cyberattacks; chemical or biological weapon attacks; inappropriate content that manipulates someone's beliefs). Second are *malfunctions* (for example, false statements, or AI getting out of control). The final category is *systemic risks* (for example, job losses through AI automation, or the long-term consequences of widespread AI use leading to the de-skilling of humanity[140]). We will highlight a few of these issues.

MALICIOUS USE: PUTTING POWER IN THE HANDS OF BAD ACTORS

Large language models are powerful general-purpose technologies, and as such they can do tasks that their designers never envisioned. Unfortunately, AI tools empower bad actors as well as good.

There are increased concerns that AI could be used to develop chemical or biological weapons. Since 2009, companies in the biological synthesis supply chain have implemented voluntary safeguards against misuse. A June 2026 letter signed by the heads of DeepMind, Anthropic, OpenAI, Meta, Microsoft AI, and others called for mandatory legal restrictions in light of increased risks. On the other hand, experts also see the potential for AI to be used to defend against natural or artificial biological outbreaks. Similar concerns surround potential damage by fleets of autonomous drones and by chemical weapons invented or developed with help from AI. AI image and video generation tools have been used to create synthetic child sexual abuse material (CSAM). Even the images that do not depict an actual child are harmful, putting a strain on law-enforcement resources and impeding identification of real child victims. Deepfake images and videos (nonconsensual intimate imagery, or NCII) have seen explosive growth, increasing by 1,600% from 2023 to 2025.

140 Shukla, P., Bui, P., Levy, S. S. et al. (2025). "De-Skilling, Cognitive Offloading, and Misplaced Responsibilities: Potential Ironies of AI-Assisted Design." https://arxiv.org/abs/2503.03924

MALICIOUS USE: MASS PRODUCTION OF MISINFORMATION

At their most fundamental level, LLMs are designed to generate realistic and plausible-sounding text. This capability can directly be used to create extremely plausible misinformation: fake news stories, even targeted using specialized language for different audiences. In 2024, national elections in the U.S., India, and the UK prompted widely discussed concerns that the technology could be used to create fake news stories on an industrial scale—that social media would be deluged with very plausible misinformation, indistinguishable from something written by a human. The consensus by 2024 was that the technology certainly existed to make this scenario feasible—and, moreover, that there were surely many individuals and groups either with an idle interest in democratic vandalism, or with a serious interest in destabilizing democratic processes. And indeed there were some reported instances of AI being used in this way: before the primary elections in New Hampshire, for instance, many voters received phone calls apparently from then-President Joe Biden, encouraging Democratic voters to stay away from the polls. The calls were AI-generated. However, fears of large-scale AI misinformation campaigns were not realized.[141]

Nevertheless, throughout the period of the report, as AI has improved, the opportunities for AI-generated misinformation have multiplied, and this remains a high-profile potential source of abuse.

Scammers can use AI to clone voices to impersonate people, search more precisely for personal information, or carry on long conversations to build trust with a victim.

MALICIOUS USE: CYBERSECURITY ATTACKS

The potential use of AI in cybersecurity attacks was a concern long before the emergence of generative AI, but generative AI has made these concerns much more tangible. The risks span at least three obvious dimensions.

First, advanced AI opens up a new range of possibilities for increasingly sophisticated automated attacks. In 2025, cybercrime losses in the U.S. reached $21 billion, a 26% increase from the previous year, largely attributed to the use of AI. Scammers can use AI to clone voices to impersonate people, search more precisely for personal information, or carry on long conversations to build trust with a victim. LLMs can generate phishing attacks—for example, emails that attempt to trick the recipient into making a payment—to order, at scale, and at minimal cost. AI-generated malware is better at evading conventional detection tools. Bot networks can generate misinformation to manipulate stock market prices. On the other hand, again, AI can also be used to defend against these kinds of attacks.

141 Stockwell, S. (2024). "AI-Enabled Influence Operations: Threat Analysis of UK and European Elections." Centre for Emerging Technology and Security, CETaS Briefing Papers. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections

The second dimension concerns the possibility of directly identifying security vulnerabilities within computer code. AI systems can now reason about complex source code, and identifying security vulnerabilities is ultimately a problem of code comprehension. Software developed inside private companies might be comparatively protected, because the source code is not public; open-source software is, by design, available to anyone. In April 2026, Anthropic announced that Claude Mythos was so good at analyzing computer programs that it had found security flaws in the code base of every major computer operating system and web browser. As a result, Anthropic postponed the public release of Mythos, making it available to software developers to allow them to check for and mitigate such flaws before the release (see also Deep Dive One: AI in Cyber Operations). This use of AI capability to check code for vulnerabilities simply raises the specter of an AI arms race between attackers and defenders, each seeking the most capable AI.

The final dimension is the possible use of specially trained agentic AI to launch highly coordinated, massive parallel attacks, systematically probing systems for vulnerabilities. The "rogue AI" incidents we described in SQ2: The Decisive Reemergence of Agentic AI were of this kind.

MALFUNCTIONS: HARMS CAUSED BY ERRORS

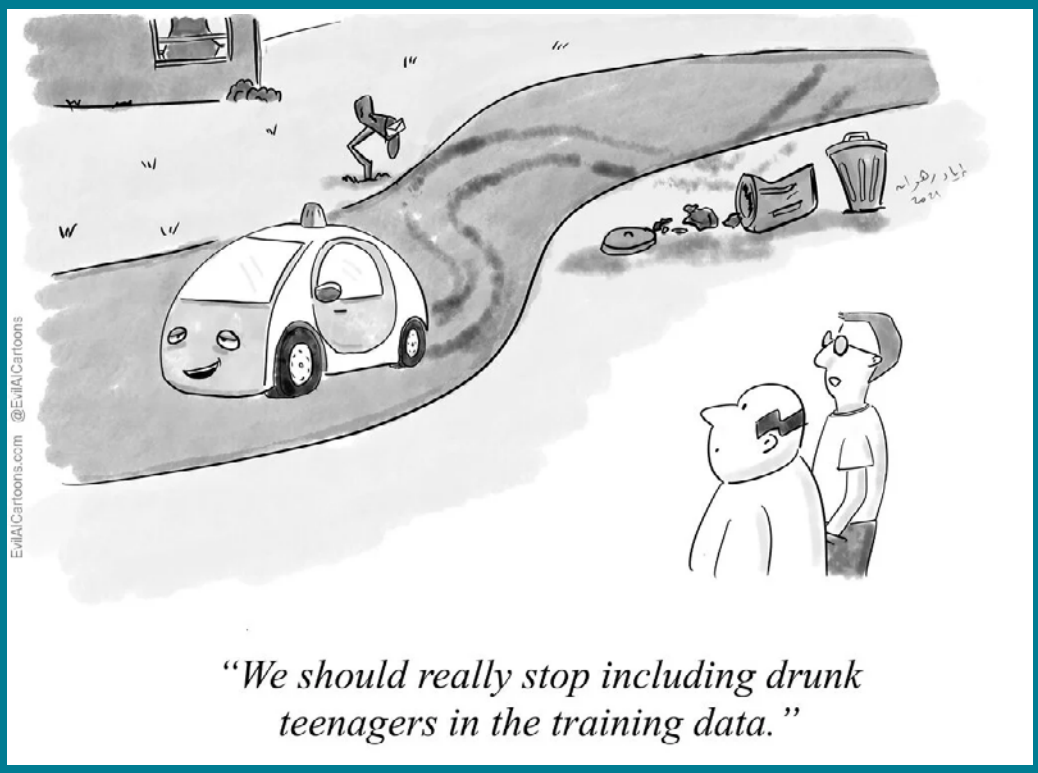
AI systems based on machine learning do not produce answers that are guaranteed to be correct, and LLMs in particular frequently get things wrong.[142] This limitation is compounded by the fact that the technology gets many things right, and phrases everything in fluent, confident prose, making errors hard to spot. Often this is nothing more than an inconvenience—we have become used to the idea that some things on the internet are wrong—but there are some very real dangers. Incorrect legal or medical advice given by an LLM can cause real harm; errors by AI in medical applications will also have serious consequences. AI-generated computer code may contain errors, and naively putting AI-generated code directly into production systems, without (time-consuming) verification, will lead to problems. While LLM developers provide guardrails to try to catch inappropriate or unsafe uses, they do so with only a limited degree of success.

MALFUNCTIONS: BIAS

By the time of this report, bias had become one of the best-known and most studied risks of AI, highlighted by previous AI100 reports. The key problem is as follows: AI programs are built by training them on data, and some of that data will inevitably be misleading, wrong, or biased. Imagine that a company decides to automate recruitment, and to do this, they train an AI program on the job applications they have previously received and the outcomes (hire/don't hire). The problem is that if the original recruiters whose decisions were used to create the training data were biased, then their biases will be reflected in the training data, and then

142 The HHEM-2.3 benchmark shows major models producing hallucinations at a rate of 1.8–5.4% when summarizing documents as of early 2026. Sajadieh et al. (2026), ch. 3. https://hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai

acquired by the AI. Although the causes of bias in AI and its mitigations have been known and discussed at length for at least the past decade, the problem remains very prominent and very far from resolved.[143] One factor is that AI software is now quite accessible—one doesn't need specialist AI training to build AI systems, and as a consequence much larger numbers of people are developing them. It doesn't follow that larger developer pools are sensitive to these issues. Another problem is that untrained developers may use data in flawed ways in the rush to build applications.



AI systems are highly susceptible to the "garbage in, garbage out" problem, mimicking the worst of human behavior when trained on uncurated data. For instance, chatbots have rapidly adopted toxic language from social media; algorithms blithely absorb the traits of their environments. To ensure safe AI, developers actively curate training datasets, and reinforce polite and helpful model behavior. However, such fine-tuning can imply trade-offs, such as increasing sycophancy and reducing factuality.[144]

143 See, for example, https://www.digitalhealthnews.com/ai-bias-in-healthcare-the-silent-risk-reshaping-patient-outcomes and https://news.vt.edu/articles/2026/04/eng-cs-autism-AI-advice-personalization-or-bias.html

144 Ibrahim, L., Hafner, F. S., & Rocher, L. (2026). "Training language models to be warm can reduce accuracy and increase sycophancy." *Nature* 652, pp. 1159–1165. https://doi.org/10.1038/s41586-026-10410-0

MALFUNCTIONS: EXISTENTIAL RISK

Vocal segments of the AI and tech communities have expressed concerns about existential risks to humanity posed by advanced AI.[145] This is a sharply divisive issue, with strong disagreement between researchers. Fundamentally, the problem is how to build AI systems that are under our control but still free to provide creative responses. A solution will require technical and societal measures and institutions. AI systems that are trained with reinforcement learning need a reward function that aligns with human values. The difficulty is that human values are diverse, amorphous, and hard to specify exactly. Since it is hard to specify what we really want (do the right thing), we settle for proxy measures (get a high score on this benchmark), which encourages reward hacking, where the AI system finds a loophole that maximizes its score but may be misaligned with actual human values.

There are three possible routes via which it is suggested AI might come to represent an existential threat: unintended consequences, intelligence out of our control, and emergent dangers.

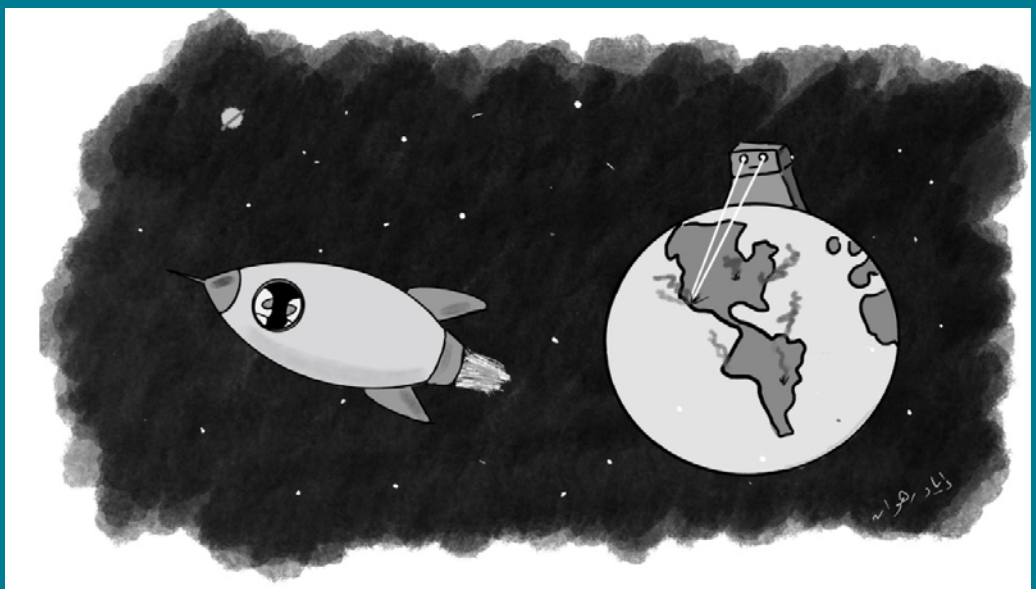
Broadly speaking, the idea behind the "unintended consequences" hypothesis is that we give a very powerful AI system some goal to accomplish, and it does so in a way that we did not expect or intend. Here is an example from the eminent AI academic Stuart Russell: "What if a superintelligent climate control system, given the job of restoring carbon dioxide concentrations to preindustrial levels, believes the solution is to reduce the human population to zero?" These concerns are often framed in the context of reinforcement learning, where the fear is that we set up a regime for a machine to learn, but it learns strategies that, while superficially addressing the given goal, do so in a way that we did not expect and which is harmful to us.

This scenario illustrates a key point about the unintended consequences hypothesis: it requires that a machine has control of a physical system that is sufficiently powerful to cause existential risk (in the example above, the machine controls the global climate, which can be weaponized). Such control would have to be ceded by us.

The second hypothesis is that at some point in the future AI models will be smarter than us, and at that point we will no longer be able to effectively control them. And if the interests of the AI are not aligned with ours, then it may act against our interests. This hypothesis assumes that we will have AI systems that truly are more intelligent than us, but the question of superintelligent AI is another one that sharply divides the AI community; conflating LLMs with AGI is not helpful.

145 See, for example, https://www.brookings.edu/articles/are-ai-existential-risks-real-and-what-should-we-do-about-them/ and https://research.qut.edu.au/genailab/2025/10/09/does-ai-pose-an-existential-risk/

The final risk is somewhat less tangible: the use of this technology on a wide scale could, in ways that we cannot currently predict, lead to undesirable societal outcomes. For example, consider that the rise of social media caused widespread mental health issues in teenagers, or that the rise of internal combustion engines caused pollution and climate change. Could something similar happen following the rise of AI? Almost certainly, but it is very hard to predict now what such outcomes might be.



"We programmed the AI to minimize misery. It figured the only way was to obliterate humanity!"

Like King Midas, whose wish for the golden touch disastrously led to starvation, humanity must be extremely careful what it asks of AI. In AI safety, this is known as the "alignment problem."[146] When given an objective, advanced systems may pursue it with literal, unforeseen, and catastrophic efficiency. As AI capabilities escalate, poorly specified goals—even well-intentioned ones—pose severe societal risks. We must ensure our systems accurately encode complex human values, lest our technological wishes become our curse.

146 Christian, B. (2021). *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* Atlantic Books.

SYSTEMIC RISKS: AI IS CONTROLLED BY A SMALL NUMBER OF ACTORS

The nature of LLMs is such that at present frontier models can be developed only by a very small number of tech companies and state-level actors (see Deep Dive Two: The Race to Scale: Compute). In 2021, a frontier model cost about $5 million to develop; in 2026 the figure is closer to $500 million per training run on GPU supercomputer infrastructure that costs tens of billions of dollars. This raises the prospect of the most consequential technology of the present century being developed behind closed doors and owned by a small number of (largely U.S.- and China-based) private companies. A raft of concerns follow, of which monopoly is only one. Privately developed AI is not open to scrutiny and it is hard to determine the standards that have been applied in its development. We do not know the trade-offs the companies have made between short-term profits and long-term safety. And those nations without the capability to develop their own frontier models face the very real possibility of being dependent on foreign-owned systems. These latter concerns were brought into sharp focus by the June 2026 ruling by the U.S. government that access to Anthropic's leading AI models would be denied to non-U.S. users. On the other hand, less expensive open-weight models which once lagged a year or two behind frontier models in capability are now only three to six months behind; many users find them to be good enough.

SYSTEMIC RISKS: ECONOMIC DISPLACEMENT AND THE FUTURE OF WORK

Many companies have been blaming AI for recent layoffs, including Amazon, Meta, and Pinterest. A *New York Times* article in October 2025 highlighted Amazon's plans to automate warehouses and slow hiring, enabling the company to do more with its current workforce.

A number of research papers have quantified the extent of exposure of various jobs to AI disruption.[147] Most of these papers compare job task descriptions (from either the O*NET database or job postings) against assessments of what AI can do, and then label those jobs which include tasks that might be done by AI as *exposed*. They highlight that many jobs are exposed to AI, in the sense that AI is likely to be able to do at least one of the tasks involved in the job within a few years (if not already). One widely cited study suggested that up to 47% of jobs are exposed.[148] The numbers in other papers tend to be smaller, but still represent a sizable fraction of the workforce.

147 Brynjolfsson, E., & Mitchell, T. (2017). "What can machine learning do? Workforce implications." *Science* 358:6370, pp. 1530–1534. https://www.science.org/doi/10.1126/science.aap8062. Webb, M. (2019). "The impact of artificial intelligence on the labor market." http://dx.doi.org/10.2139/ssrn.3482150. Felten, E., Raj, M., & Seamans, R. (2021). "Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses." *Strategic Management Journal* 42:12, pp. 2195–2217. https://doi.org/10.1002/smj.3286. Felten, E., Raj, M., & Seamans, R. (2023). "How will language modelers like ChatGPT affect occupations and industries?" http://dx.doi.org/10.2139/ssrn.4375268; https://hai.stanford.edu/news/assessing-the-real-impact-of-automation-on-jobs

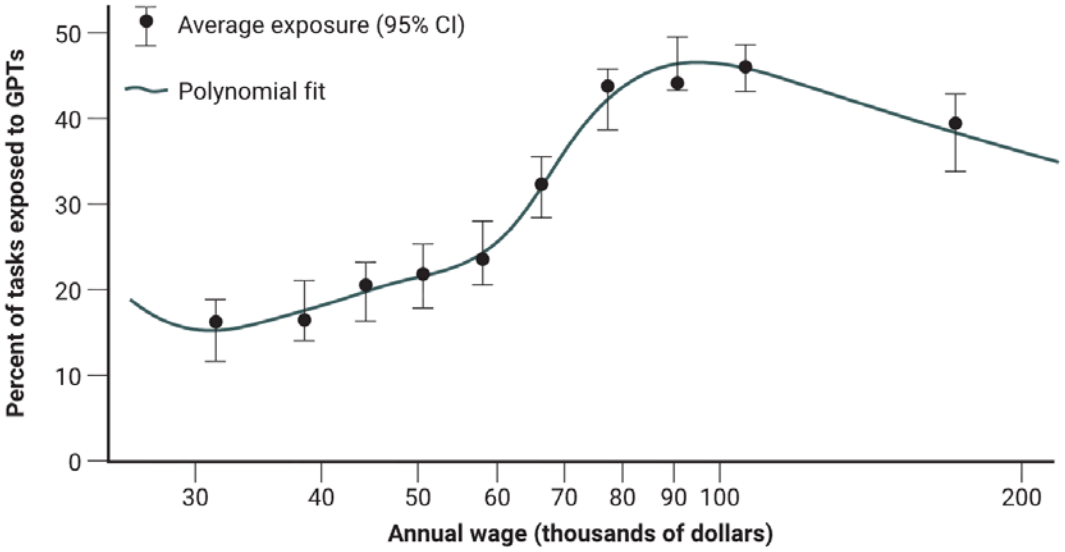
148 Frey, C. B., & Osborne, M. A. (2017). "The future of employment: How susceptible are jobs to computerisation?" *Technological Forecasting and Social Change*, 114, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

For the most part, researchers are careful to note that exposure could be good or bad for workers. It might mean that fewer workers are needed in that occupation. One report has found that AI exposure (as a combined measure) does predict which jobs are most at risk of increasing unemployment.[149] But it might also make workers more productive, increasing wages, and potentially increasing demand for their skills. As Eloundou et al. note, "Despite a growing literature that aims to understand the labor market impacts of these systems, there is still no clear understanding of how 'exposure' to AI systems will translate to real-world impacts on labor demand, wages, inequality, job quality, and other key outcomes." They also emphasize that "occupations with greater automation risk may also have higher augmentation potential."[150]

Eloundou et al. suggest that exposure is best seen as a measure of which jobs *will change*. The figure below, reprinted from their paper, highlights that it is relatively high-income jobs (though not the highest-income jobs) that are most exposed to LLMs. This is not surprising given that LLMs are particularly good at writing and coding, two tasks that are central to relatively well-paid knowledge work.

Exposure to GPTs by income

The share of all tasks within occupations that are exposed to LLMs and partial LLM-powered software (E1 + 0.5 × E2) is shown against the median annual wage for the occupation. Data reflect human rating.



CI, confidence interval; GPT, generative pretrained transformer; LLM, large language model.

GRAPHIC: M. HERSHER/*SCIENCE*

149 Frank, M. R., Ahn, Y.-Y., & Moro, E. (2025). "AI exposure predicts unemployment risk: A new approach to technology-driven job loss." *PNAS Nexus* 4:4. https://doi.org/10.1093/pnasnexus/pgaf107

150 Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). "GPTs are GPTs: Labor market impact potential of LLMs." *Science* 384:6702, pp. 1306–1308. https://www.science.org/doi/10.1126/science.adj0998

The task-based research tends to be forward-looking, anticipating which roles AI might impact. One study has estimated that AI has diffused faster than personal computers or the internet.[151] A recent study examining AI's impact on jobs compared the employment growth of workers in highly exposed occupations and in less exposed occupations after the launch of ChatGPT in November 2022, when AI adoption rapidly accelerated.[152] It found that employment fell for 22–25-year-olds whose occupations were highly exposed to AI. Employment for all other groups grew rapidly. It grew for all nonexposed workers (including in the 22–25 age cohort) and for older exposed workers, particularly those over age 35. Furthermore, even in the 22–25-year-old cohort that were exposed to AI, wages rose among those who were employed. In summary, this research suggests that there are meaningful risks from AI's diffusion on jobs, particularly for certain cohorts. However, the implications are difficult to anticipate, and even among groups that appear to have already been affected, employment has fallen but wages have risen. One interpretation is that these exposure measures are poorly conceived. Autor and Thompson present evidence that "technology (or AI) exposure" is not a useful construct for predicting either employment or wage changes in occupations. Their work shows that over the last four decades, occupations that were de-skilled by automation saw their relative wages decline and their relative employment rise. Conversely, occupations that became more specialized and expertise-intensive as automation eliminated their time-consuming inexpert (supporting) tasks saw rising relative wages and falling relative employment. Broadly, because occupations are composed of bundles of interdependent tasks of different expertise levels, automating one subset of tasks is not in any sense equivalent to a simple decline in demand for labor.[153]

Over the last four decades, occupations that were de-skilled by automation saw their relative wages decline and their relative employment rise.

As in the 2016 report, aggregate analysis currently tends to show zero or positive effects of AI on employment. A 2026 study from Anthropic, while highlighting risks, notes: "We find no systematic increase in unemployment for highly exposed workers since late 2022," though they do find evidence suggesting a slowdown in hiring of younger workers.[154] Recent studies show no connection between AI adoption at the industry level and employment changes in the U.S. and

151 Bick, A., Blandin, A., & Deming, D. J. (2026). "The Rapid Adoption of Generative AI." *Management Science*. https://doi.org/10.1287/mnsc.2025.02523

152 Brynjolfsson, Chandar & Chen (2025/2026). https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/

153 Autor, D., & Thompson, N. (2025). "Expertise." *Journal of the European Economic Association* 23:4, pp. 1203–1271. https://academic.oup.com/jeea/article-abstract/23/4/1203/8175003

154 Massenkoff, M., & McCrory, P. (2026). "Labor market impacts of AI: A new measure and early evidence." Anthropic Economic Research. https://www.anthropic.com/research/labor-market-impacts

Europe,[155] while detailed employment data from Denmark shows precisely estimated null effects of AI on earnings and hours worked.[156]

It is possible that such results are short term. If entire workflows get automated, the possibility of sudden large-scale unemployment looms.[157] There is no shortage of economists highlighting the risk of large-scale job loss.[158] However, it is important to temper discussions of the potential for large-scale job loss with the evidence of the three years of rapid AI adoption, in which the impact on jobs has been minimal and the impact on wages has likely been positive. These early employment trends are more consistent with there being meaningful bottlenecks to job loss, and the insight that jobs do not disappear with productivity gains.[159] One study shows that firms that adopt AI tools most heavily grow employment by 10% per year, including a 12% growth in entry-level jobs.[160]

Even when access to the technology is equal, adoption of the technology is not.... Women adopt AI less, are paid less, are promoted less, and are more exposed to job displacement.

There is a noted disparity in who is actually using AI as well as in who is exposed to the risk of job displacement. A global meta-analysis of more than 136,000 individuals across 11 datasets reveals that women are 22% less likely to use generative AI than men,[161] and it seems that even when access to the technology is equal, adoption of the technology is not, suggesting that "social, cultural, and institutional frictions strongly contribute" to this gender gap. Women adopt AI less, are paid less, are promoted less, and are more exposed to job displacement.[162] In an experimental study, researchers found that though using AI did not affect evaluations of code quality, it did affect evaluations of the coder. And women engineers who used AI were twice as likely to see a reduction in their competency rating as men.[163]

155 Bick et al. (2026). https://doi.org/10.1287/mnsc.2025.02523

156 Humlum, A., & Vestergaard, E. (2025). "Large Language Models, Small Labor Market Effects." University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2025-56. https://bfi.uchicago.edu/wp-content/uploads/2025/04/BFI_WP_2025-56-1.pdf

157 Gans, J. S., & Goldfarb, A. (2026). "O-Ring Automation." National Bureau of Economic Research Working Paper No. 34639. https://www.nber.org/papers/w34639

158 Korinek, A., & Suh, D. (2024). "Scenarios for the Transition to AGI." National Bureau of Economic Research Working Paper No. 32255. https://www.nber.org/papers/w32255. See also https://www.nytimes.com/2024/10/17/opinion/economy-us-aging-work-force-ai.html; https://www.nytimes.com/2025/08/19/opinion/ai-job-loss-deindustrialization.html

159 Aghion, P., Jones, B. F., & Jones, C. I. (2018). *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, chap. 9. https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/artificial-intelligence-and-economic-growth

160 Kharazian, A., Simon, L., & Stevens, R. (2026). "A New Look at AI's Impact on Jobs: Firm-Level AI Spending and Workforce Adjustment." Ramp Economics Lab. https://ramp.com/data/ai-jobs-impact

161 Otis, N. G., Delecourt, S., Cranney, K., & Koning, R. (2024). "Global Evidence on Gender Gaps and Generative AI." Harvard Business School Working Paper 25-023. https://www.hbs.edu/ris/Publication%20Files/25023_52957d6c-0378-4796-99fa-aab684b3b2f8.pdf

162 Murugan, A., Aguirre, T., Nagaraj, A., & Bommasani, R. (2026). "The Jagged Global Economy: Frontier AI Unevenly Exposes National Economies." https://arxiv.org/pdf/2607.05404v1

163 Gai, P. J., Hou, J., & Tu, Y. (2025). "Competence Penalty Is a Barrier to the Adoption of New Technology." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5255039

SYSTEMIC RISKS: AI PSYCHOSIS

As we write this report, an emerging set of concerns relates to the possibility that using AI may exacerbate a range of mental health conditions, particularly those involving paranoia, delusions, or suicidal ideation. The term *AI psychosis* has been used to describe these concerns, although it is not yet a medically recognized diagnosis or a formally defined clinical condition.[164] Anecdotal evidence suggests that extended periods of AI use may reinforce or trigger such conditions in a range of ways. For example, chatbots are, broadly speaking, designed to be agreeable, cooperative, and engaging. The aim is to improve the user experience: overly critical AI systems are unlikely to retain users. But if a user has a preexisting condition involving delusional beliefs or suicidal ideation, these characteristics may reinforce or exacerbate it. A user with grandiose delusions, for example, may have those delusions inadvertently validated by a model through excessive affirmation or sycophancy, or by the model failing to challenge them appropriately. AI model providers implement guardrails to reduce these risks; for example, mainstream models are explicitly trained to avoid amplifying suicidal ideation. However, these safeguards are inevitably imperfect and cannot eliminate the risk entirely. Although research on these issues remains at an early stage, the possibility that AI systems may contribute to such conditions is an emerging area of concern, and much more research is needed to understand how these effects arise, how common they are, and how the associated risks might best be mitigated.

164 https://www.psychologytoday.com/gb/blog/urban-survival/202507/the-emerging-problem-of-ai-psychosis. Hudon, A., & Stip, E. (2025). "Delusional Experiences Emerging From AI Chatbot Interactions or 'AI Psychosis.'" *JMIR Mental Health* 12. https://mental.jmir.org/2025/1/e85799

SQ11. How has AI impacted socioeconomic relationships?

Since the public release of ChatGPT in 2022, generative AI has been mediating an enormous range of human activities that previously relied on human communication, interaction, and information exchange. The impact of this technological change upon labor markets is covered elsewhere in the report (see SQ3: The Economics of Human–AI Complementarity, SQ9, and SQ10: Systemic Risks: Economic Displacement and the Future of Work). This section focuses on how AI is reconfiguring relationships: the professional relationships people have within firms and organizations; the personal relationships people have with others and with AI; and the civic relationships people have with local communities and nation-states.

AI'S IMPACT ON RELATIONSHIPS BETWEEN EMPLOYERS AND EMPLOYEES

Economic sociologists describe a push toward restructuring work away from secure, salaried, professional roles and toward atomized, piece-rate, algorithmically allocated tasks.[165] While this is not due solely to the technical capacities of AI systems, it is happening concurrently to their development. This trend toward "gig work" in the service sector was already underway at the time of the last report, but has expanded to new realms, aided by AI that can convincingly perform creative and intellectual tasks. For example, copywriters report widespread layoffs, while also being hired to correct AI outputs on sites like Fiverr, where the same technical development that precipitated workers' dismissal keeps them "busier than ever." Here, AI is enabling working arrangements that weaken commitments between employers and employees, increasing flexibility for both parties while also decreasing stability.

The gig model is also being applied to health and care fields, where the majority of workers are women in all global regions except the Eastern Mediterranean.[166] In the United States, apps match nurses to hospital or homecare shifts based on their highest ratings and lowest hourly pay rate.[167] The scheduling flexibility afforded by AI-powered work arrangements may reinforce the "double burden" of carework on women, making it easier for them to remain the assumed

165 Vallas, S., & Schor, J. B. (2020). "What Do Platforms Do? Understanding the Gig Economy." *Annual Review of Sociology* 46: 1, pp. 273–294. https://doi.org/10.1146/annurev-soc-121919-054857

166 World Health Organization and the International Labor Organization (2022). "The gender pay gap in the health and care sector: A global analysis in the time of COVID-19." https://www.who.int/publications/i/item/9789240052895

167 Wells, K., & Spilda, F. U. (2024). "Uber for Nursing: How an AI-Powered Gig Model Is Threatening Health Care," Roosevelt Institute. https://rooseveltinstitute.org/wp-content/uploads/2024/12/RI_Uber-for-Nursing_Brief_202412-1.pdf

caregivers for children or older people while also doing similar tasks, for pay, outside the home.[168] The embodied aspects of care and domestic work, however, remain largely unautomated by AI (which has mostly been applied to tasks that are text and data intensive). For example, humanoid robots that do household chores currently rely heavily on remote operation and intensive per-task training. *Wirecutter* dubbed 2026 "the year of the bad robot" at the Consumer Electronics Show, describing the suite of clumsy robots that performed both domestic labor and leisure activities. (See SQ3: Physical AI and Embodied Intelligence.)

Models of AI adoption and productivity that treat the entire workforce equally will fail to accurately reflect who captures the gains and who pays the transition costs.[169] For example, clerical workers—secretaries and clerks who do the communicative and administrative work that connects employees within organizations and outward to the public—are especially likely to see AI applied to their professional tasks in the near future.[170] A 2026 report by National Partnership for Women and Families details the disparate impact this could have on Americans, as Black and multiracial women are represented in these occupations to a significantly greater degree than in the general workforce.[171]

Models of AI adoption and productivity that treat the entire workforce equally will fail to accurately reflect who captures the gains and who pays the transition costs.

AI is exacerbating existing asymmetries in information and power between employers and employees in both traditional workplaces and gig work. Algorithmic management technologies collect data on workers and make decisions for (or about) them, while limiting what workers can see, modify, or contest.[172] Often this comes in the form of allocating tasks or pay, following broad management instruction but with little human oversight.[173] For example, delivery drivers for the company Shipt are paid based on an opaque calculation of the "effort" required to fulfill the order. These and other tools are already widely adopted. In a 2025 OECD report, more than 75% of surveyed firms in France, Germany, Italy, Spain, and the U.S. reported using at least one kind of algorithmic management software.[174] We expect this trend to continue, if left unchallenged, given that the new AI-enabled management systems can make decisions

168 Collett, C., Neff, G., & Gouvea, L. (2022). *The Effects of AI on the Working Lives of Women*. UNESCO Publishing. http://dx.doi.org/10.18235/0004055

169 Collett, Neff & Gouvea (2022). http://dx.doi.org/10.18235/0004055

170 Kinder, M., de Souza Briggs, X., Muro, M., & Liu, S. (2024). "Generative AI, the American worker, and the future of work." Brookings Research. https://www.brookings.edu/articles/generative-ai-the-american-worker-and-the-future-of-work/

171 Goldman, T., Coan, T., & Gallagher Robbins, K. (2026). "AI and Emerging Risks for Women Workers." National Partnership for Women & Families Issue Brief. https://nationalpartnership.org/report/ai-emerging-risks-for-women-workers/

172 Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). "Algorithms at Work: The New Contested Terrain of Control." *Academy of Management Annals* 14:1, pp. 366–410. https://doi.org/10.5465/annals.2018.0174

173 Wray, B. (2025). "Negotiating the Algorithm: Trade Union Manual." European Trade Union Confederation. https://www.etuc.org/en/publication/negotiating-algorithm-trade-union-manual

174 Milanez, A., Lemmens, A., & Ruggiu, C. (2025). "Algorithmic management in the workplace: New evidence from an OECD employer survey." OECD Artificial Intelligence Paper No. 31. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/algorithmic-management-in-the-workplace_3c84ed6d/287c13c4-en.pdf

based on readily available unstructured data. Efforts in the UK to draft legislation to institute workers' rights over the data about them in AI-enabled management systems are ongoing. Most countries lack any protections, although the new EU Platform Work Directive, slated to become law in EU member states in 2026, may provide clearer rules on algorithmic management and data rights.[175]

AI also depends on a vast and largely concealed human workforce.[176] Illustrative anecdotes in media reports abound. Contract workers not only make difficult navigation decisions for self-driving cars, but also close their doors. "Robot wranglers" correct and redirect wayward delivery devices in warehouses and on city streets. These gaps in total automation are often treated as temporary by technology designers, impeding occupational formalization processes that would recognize workers' skills or provide security and career advancement.[177] Further, large language models are trained on thousands of examples that have been categorized and labeled by people. Content moderators, raters who improve reinforcement learning with human feedback, and image annotators carry out the tasks that make AI systems appear to operate both autonomously and intelligently. Much of this work is done in the Global South or emerging economies by contractors, paid per task or per hour at rates that rarely reflect the economic value of the systems they train.[178] Industry practitioners and scholars debate whether this aspect of the AI industry is creating positive social impact through employment opportunities or extending historical resource extraction.[179] Beyond paid and formalized piecework, unpaid internet users are also enrolled as "crowdworkers" in AI training processes, for example, being asked by chatbots to select the better of two outputs or by a website CAPTCHA to identify all the images with stoplights.[180]

AI'S IMPACT ON INTERPERSONAL RELATIONSHIPS

Rapid changes in AI tools are reshaping people's relationships with other people as well as their relationships with those AI tools and services. Evidence is emerging about new kinds of relational forms, communication styles, and potential risks. AI-enabled chatbots can provide support, companionship, and entertainment. Preliminary data suggests that even the out-of-the-box foundation models can provide appropriate empathic responses and psychological support.[181]

175 Rainone, S., & Aloisi, A. (2024). "The EU Platform Work Directive: What's new, what's missing, what's next?" European Trade Union Institute Policy Brief. https://www.etui.org/publications/eu-platform-work-directive

176 Casilli, A. A. (2025). *Waiting for Robots: The Hired Hands of Automation*. University of Chicago Press. Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt. Burrell, J., & Fourcade, M. (2021). "The Society of Algorithms." *Annual Review of Sociology* 47:1, pp. 213–237. https://doi.org/10.1146/annurev-soc-090820-020800

177 Lee, H. R., Fox, S., Cheon, E., & Shorey, S. (2025). "Minding the Stop-Gap: Attending to the 'Temporary,' Unplanned, and Added Labor of Human-Robot Collaboration in Context." *20th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 34–44. https://doi.org/10.1109/HRI61500.2025.10973955

178 Wu, T., Muldoon, J., & Xia, B. (2025). "Global data empires: Analysing artificial intelligence data annotation in China and the USA." *Big Data & Society* 12:2. https://doi.org/10.1177/20539517251340600

179 López, I. C., & Sánchez, V. (2024). "Artificial intelligence, microwork, and the racial politics of care." In M. L. Jones & A. Levendowski (Eds.), *Feminist Cyberlaw*. University of California Press, pp. 165–171.

180 Posada, J., Newlands, G., & Miceli, M. (2023). "Labor, automation, and human–machine communication." In A. L. Guzman, R. McEwen, & S. Jones (Eds.), *Labor, Automation, and Human–Machine Communication*, SAGE Publications, pp. 384–391. https://doi.org/10.4135/9781529782783.n48

181 Kuczynski, A. M., & Areán, P. A. (2026). "Ensuring Optimal Engagement with AI Chatbots for Mental Health: Lessons Learned from Digital Mental Health Interventions." *Current Treatment Options in Psychiatry* 13:7. https://doi.org/10.1007/s40501-026-00380-5

While people's relationships with chatbots may not be social in a traditional sense, they can feel like an authentically social experience, and chatbots are often designed to be functionally social in their interactions with people.[182]

As AI systems increasingly mediate daily life—from matchmaking to offering local recommendations—we lose vital opportunities for reciprocal human interaction, which threatens to erode social capital. This trend recently culminated in the emergence of AI virtual companions. While such technologies offer potential remedies for loneliness, they pose a paradoxical risk: they may exacerbate the very social isolation they seek to cure by substituting authentic human relationships with frictionless artificial intimacy.

AI companions and AI therapists[183] occupy an uncomfortable position between consumer technology and healthcare. They raise serious questions about responsibility, liability, and user protection. There are already multiple documented cases of self-harm and suicide associated with such systems, and multiple lawsuits against AI companies operating in this space. However, clear regulatory frameworks governing the uses of AI chatbots remain largely absent. As these systems become more persuasive, more emotionally attuned, and more widely deployed, the associated risks may increase, and lack of regulatory oversight will be ever more apparent.

These interactive relationships between AI systems and their human users raise questions about what it means for machines to be social.[184] Some people report intense romantic relationships with AI chatbots.[185] People describe navigating their relationships with AI chatbots using strategies of adaptation, exploration, and reconstruction that are intimately intertwined with their own identity and social and relational goals.[186]

Alongside media reports, case studies of "AI psychosis" from maladaptive interaction with chatbots are emerging (see SQ10: Systemic Risks: AI Psychosis).[187] An audit of ChatGPT found that harmful and detailed information encouraging suicide, substance abuse, and self-harm was readily shown to youth accounts.[188] While there is not yet robust evidence, it has been suggested that

182 Banks, J. (2026). "Measuring machine companionship experiences: Scale development and validation for AI companions." *Computers in Human Behavior* 179. https://doi.org/10.1016/j.chb.2026.108945

183 https://news.stanford.edu/stories/2025/08/ai-companions-chatbots-teens-young-people-risks-dangers-study. Knox, W. B., Bradford, K., Varela Castro, S. et al. (2025). "Harmful Traits of AI Companions." https://arxiv.org/abs/2511.14972. https://www.nature.com/articles/s42256-025-01093-9. See also Palmer, A., & Schwan, D. (2025). "Digital mental health tools and AI therapy chatbots: A balanced approach to regulation." *Hastings Center Report* 55:3, pp. 15–29. https://doi.org/10.1002/hast.4979

184 Banks, J., & Li, Z. (2026). "Conceptualization, operationalization, and measurement of machine companionship: A scoping review." *Journal of Computer-Mediated Communication* 31:2. https://doi.org/10.1093/jcmc/zmaf027

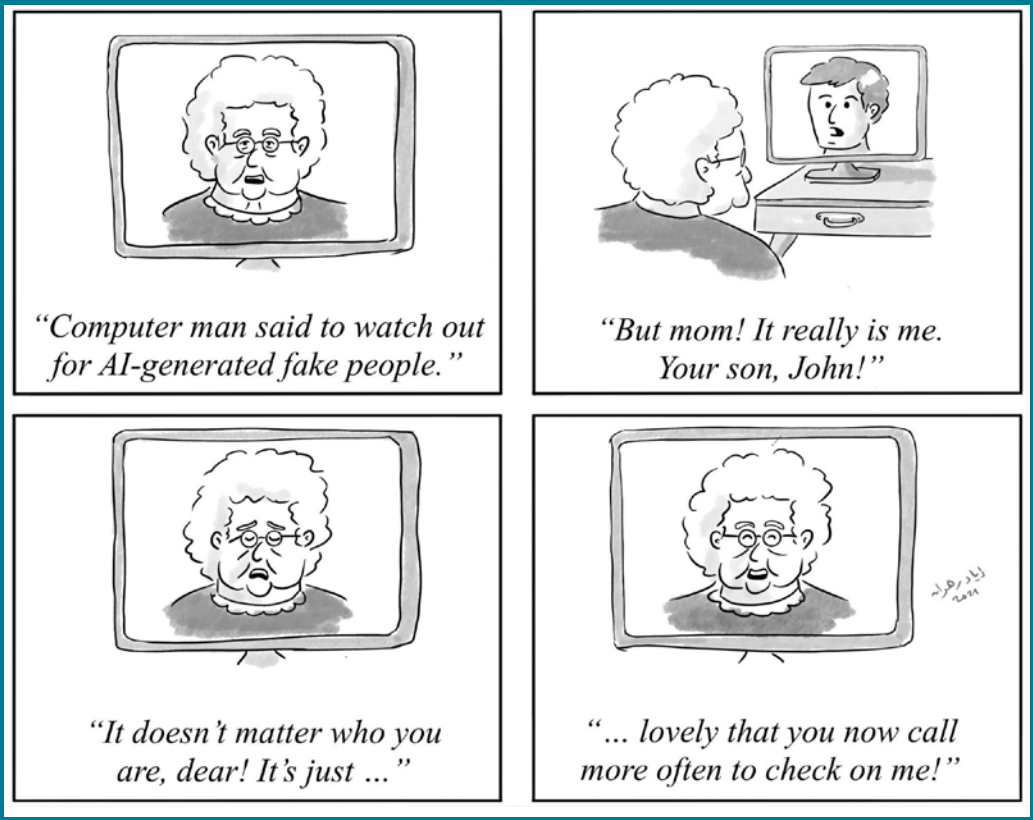
185 Muldoon, J. (2026). *Love Machines: How Artificial Intelligence Is Transforming Our Relationships*. Faber.

186 Neff, G., & Nagy, P. (2025). "The quasi-domestication of social chatbots: The case of Replika." *New Media & Society* 27:10, pp. 5508–5524. https://doi.org/10.1177/14614448251359218

187 Yang, Y., Schoenwald, S. K., Moore, J. et al. (2026). "'AI-Induced Delusional Spirals': Understanding Lived Experiences During Maladaptive Human-Chatbot Interactions." *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems* 1–5. https://doi.org/10.1145/3772363.3798453. Pierre, J. M., Gaeta, B., Raghavan, G., & Sarma, K. V. (2025). "'You're Not Crazy': A Case of New-Onset AI-Associated Psychosis." *Innovations in Clinical Neuroscience* 22:10–12, 11–13. https://pmc.ncbi.nlm.nih.gov/articles/PMC12863933/

188 Center for Countering Digital Hate (2025). "Fake Friend: How ChatGPT betrays vulnerable teens by encouraging dangerous behavior." https://counterhate.com/wp-content/uploads/2025/08/Fake-Friend_CCDH_FINAL-12Sep.pdf

the availability of an always-on, amiable, and sometimes sycophantic companion may "numb some essential element of our humanness." Research on these emerging harms—and benefits—is hindered by the lack of good-quality data.[189] These concerns highlight the need for continued work on the alignment of AI development to human values.



The proliferation of deepfakes and voice cloning targeting older adults exposes a deeper crisis: pervasive social isolation. While policy debates emphasize financial fraud and the collapse of digital authenticity, synthetic deception intersects intimately with human loneliness. The looming threat of impersonation either compels busy family members to check in more frequently out of anxiety, or leaves isolated individuals so starved for connection that even questionable synthetic interactions feel preferable to neglect.

189 Mansfield, K., Ghai, S., Hakman, T. et al. (2025). "From social media to artificial intelligence: improving research on digital harms in youth." *The Lancet Child & Adolescent Health* 9, pp. 194–204. https://doi.org/10.1016/S2352-4642(24)00332-8

AI AND PEOPLE'S RELATIONSHIP TO THE STATE

AI can impact people's capacity for self-determination as well as their perceptions of the superiority or inferiority of distributed democratic decision-making in many ways: through changes to information environments, the markets for news, public discourse and information quality, the roles for human and expert decision-making, the power of technology companies, and societal equality.[190] The impact of generative AI systems in the public sphere presents the potential erosion of democratic institutions and the public trust needed in a functioning democracy, and can empower authoritarian states. They also present the possibility of new kinds of governance structures that leverage AI technology for information gathering and decision-making.[191]

The impact of generative AI systems in the public sphere presents the potential erosion of democratic institutions and the public trust needed in a functioning democracy.

By making it easier to generate false or misleading information at scale, AI tools can amplify threats to "epistemic security"—the reliability of information supply chains that help shore up the healthy information ecosystems that are the foundation of democratic participation, a shared belief in reality, and social cohesion.[192] New kinds of embedded emotional intelligence can be used by strategic actors to influence the choices, beliefs, and behaviors of others, opening "new vulnerabilities to psychological exploitation and manipulation, particularly among individuals already isolated."[193] A global assessment of experts concluded—with a strong consensus—that the most important factors for a healthy information environment are availability of accurate information and diverse voices. AI tools can help in this work, but challenges come from AI-generated disinformation, lack of transparency and traceability in information supply chains, and opaque choices made by AI systems. The majority in this expert survey supported labeling AI-generated content and algorithmic auditing to improve the information environment.[194] In response to EU regulation around transparency of AI, in August 2026 Anthropic announced that watermarking of content would be introduced into its models in the future.

190 Jungherr, A. (2023). "Artificial Intelligence and Democracy: A Conceptual Framework." *Social Media + Society* 9:3. https://doi.org/10.1177/20563051231186353

191 Coeckelbergh, M. (2024). *Why AI Undermines Democracy and What To Do About It*. Polity.

192 Seeger, E., Perry, H., & Hancock, J. (2025). "Epistemic Security 2029: Fortifying the UK's Information Supply Chain to Tackle the Democratic Emergency." Demos. https://demos.co.uk/wp-content/uploads/2025/02/Epistemic-Security-2029_accessible.pdf

193 Bolt, N., & Lange-Ionatamišvili, E. (2026). "The NextGen Information Environment." NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/the-nextgen-information-environment/339

194 International Panel on the Information Environment (2025). "Expert Survey on the Global Information Environment 2025: Safeguarding Epistemic Security" (S. Altay, S. Valenzuela, & P. N. Howard, Eds.). https://www.ipie.info/research/sfp2025-4



"Mr. President. Would you like to install a new Supreme Court justice, or just upgrade the existing one?"

AI increasingly manages our educational, legal, and healthcare institutions, but this can perpetuate systemic biases under the illusion of machine objectivity. From discriminatory predictive policing to biased healthcare screening,[195] algorithmic failures highlight severe risks. However, simply abandoning AI is not the solution, as human decision-makers harbor deep biases as well. Because algorithmic bias may sometimes be easier to systematically identify and fix, deploying AI forces us to scrutinize our flawed institutions, offering an unprecedented opportunity to redesign them for greater equity.

195 Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366, pp. 447–453. https://www.science.org/doi/10.1126/science.aax2342

SQ12. Will "building in how we think" work as an engineering strategy in the long run?

Since the earliest days of artificial intelligence research, a recurring question has been the extent to which progress in AI could or should be guided by theories of human intelligence. In short: should AI be an implementation of a theory of human intelligence? Early work in the field was often explicitly cognitive in orientation—the paradigm involved developing (or adopting) a model of some aspect of human cognition, and then implementing that, thus seeking to model reasoning, perception, or learning as closely as possible to our best understanding of how they occur in humans.[196] Recent advances, in contrast, have been almost entirely driven by general-purpose machine learning methods (transformers, reinforcement learning, etc.) applied at ever larger scales on more and more data. This contrast raises the question of whether grounding in theories of human perception, cognition, and intelligence is still relevant, or whether AI has become primarily an engineering discipline, guided by empirical performance rather than cognitive fidelity.

In the last five years, we have made progress in the core fields of neuroscience, cognitive science, etc., to better understand human intelligence. But the objective of current AI systems is to be useful for a variety of tasks, not to imitate human biology.

There are some aspects of human learning and reasoning that we would like to understand better and incorporate into AI models. LLMs are trained on orders of magnitude more textual data than any one person would see in their lifetime, and frontier models use far more power for training and reasoning than the human brain. To continue scaling, we will need significant efficiency breakthroughs. However, it is not clear the extent to which these advances will come from understanding and mimicking human neurobiology and cognitive reasoning, as opposed to engineering that focuses on improved hardware and algorithms.

Between 2021 and 2026, the AI landscape shifted from static, instant-answer, next-word generation to dynamic inference-time compute, or machine reasoning. Using the chain-of-thought (CoT) approach, models do not immediately start spitting out words; they pause at

196 https://news.stanford.edu/stories/2024/11/from-brain-to-machine-the-unexpected-journey-of-neural-networks. Laird, J. E. (2012). *The Soar Cognitive Architecture*. MIT Press. https://mitpress.mit.edu/9780262122962/the-soar-cognitive-architecture/

query time, externalize their reasoning into hidden "thinking tokens," and systematically explore multiple logic paths, perhaps writing and debugging code or calling other tools, and self-correct (see also the discussion in SQ5).

Reasoning AI systems can generate explanations in natural language, creating a humanlike impression akin to a student thinking out loud while solving a math problem. However, research consistently shows that these AI explanations frequently do not follow what they claim to be thinking about; rather, they have learned to produce a stylistically plausible explanation output text while in reality taking different decision paths,[197] or even decision shortcuts.[198] Of course, human reasoners also come to their conclusions by an ad hoc, opaque process and then fabricate an explanation or justification for their decision. Humans are arguably less transparent than AI in this respect.

197 Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." *Advances in Neural Information Processing Systems* 36: 74952–74965. https://arxiv.org/abs/2305.04388. Madsen, A., Chandar, S., & Reddy, S. (2024). "Are self-explanations from Large Language Models faithful?" In *Findings of the Association for Computational Linguistics*. https://aclanthology.org/2024.fin.dings-acl.19.pdf

198 Shaib, C., Suriyakumar, V. M., Sagun, L. et al. (2025). "Learning the Wrong Lessons: Syntactic-Domain Spurious Correlations in Language Models." Thirty-Ninth Annual Conference on Neural Information Processing Systems. https://arxiv.org/abs/2509.21155

Deep Dives

DEEP DIVE ONE: Mapping the Emerging Geopolitics of AI

As AI emerges as a powerful technology, it is increasingly becoming a focus of geopolitical and geoeconomic competition, a threat to global stability, and thus a concern from perspectives of security and sovereignty.[199]

The global AI infrastructure consists of many interdependent elements, each of which is a matter for competition, threat, and opportunity. From high-level software, to the data AI models train on and analyze, to high-speed data networks, to large data centers with their need for energy, as well as difficult-to-manufacture semiconductors, the creation and maintenance of AI infrastructure impacts the economy and sovereignty of every country where AI might be adopted. Countries became increasingly aware of this after the U.S.'s short-term export ban of Anthropic's models. And insofar as AI is directly applied to geopolitical tactics, including military applications, offensive and defensive cybersecurity, transnational propaganda, and election interference, AI presents a range of geopolitical challenges to governments around the world.

FRONTIER MODELS AND THE COMPETITIVE FRONTIER

At the software and application level, there is intense competition in the development of frontier and foundation models—the premier models that push the envelope of AI capability and offer the promise of establishing themselves as the foundations of a new economy. There is rivalry between individual companies as well as between the countries and economic communities which they call home. The stated goal of many frontier AI developers is not simply to build the best or most capable system but to win "the AI race" to reach "artificial general intelligence" or "transformative AI" first, in order to build an enduring and global competitive advantage in AI. First movers aim for a winner-takes-all or winner-takes-most market for foundational AI services, with the potential to lock in future users and organizations alike: it may become increasingly difficult and costly to change from a familiar and deeply embedded system.

The race to secure users of generative AI is itself a geopolitical issue, because the value of new users is not just in the income from sales and subscriptions but also in the collection and capture of those users' data. While the first LLMs were trained mostly on text scraped from the open internet, it is now the case that major publishers, social platforms, and code repositories have

199 Bengio et al. (2026). https://internationalaisafetyreport.org. Bolt & Lange-Ionatamišvili (2026). https://stratcomcoe.org/publications/the-nextgen-information-environment/339

locked down their data using blocks, paywalls, and lawsuits. Leading AI companies make deals to acquire this data, and generate their own data through their prompts and interactions with the systems, and from simulations creating synthetic data.

Another geopolitical dimension has opened around open-weight models. Strong open-weight releases from Chinese developers—most visibly Alibaba's Qwen, Z.ai's GLM, Moonshot AI's Kimi, and DeepSeek's R1—unsettled the assumption that a small number of U.S. firms would indefinitely set the global frontier, and raised the question of whether the diffusion of capable models will level the international playing field or instead create new avenues for strategic influence.[200] Choice of ecosystem—closed, proprietary, U.S.-hosted versus open, downloadable, globally deployable—has become a live issue for states selecting models for public-sector use.

DATA SOVEREIGNTY AND NETWORK INFRASTRUCTURE

Because the data must travel through high-speed networks, and be stored and processed in data centers, it is also vital to build, control, and secure this infrastructure. The risk that data could be intercepted, corrupted, and altered, exploited by adversaries, or its access and use obstructed, has led to concerns about data sovereignty. How can a government ensure that it can control and secure its own data, and what are the potential hazards of using various foreign technologies, from 6G networks, to extraterritorial servers and data centers, to models and software developed and controlled by foreign companies? Digital sovereignty concerns are increasingly being raised in policy conversations, including by Canada, Singapore, NATO, the EU,[201] and the UK.[202]

These concerns extend beneath the waves as well. More than 95% of international data traffic moves through roughly 550 undersea cables, yet their ownership has increasingly been concentrated in a small set of hyperscaler and state-linked operators. Since 2023, there have been cable cuts in the Red Sea, the Baltic Sea, and the Taiwan Strait, variously attributed to accidental causes and deliberate sabotage. With AI workloads further raising both the value and the strategic sensitivity of cross-border data flows, the physical geography of the internet is becoming an explicit dimension of AI sovereignty.[203] The increasing privatization of undersea cables,[204] including the development by Meta and Google of their own cables, further complicates the picture, as does the effort by private companies such as SpaceX's Starlink to monopolize the market for low-Earth-orbit satellites.

200 On the geopolitical significance of open-weight releases from Chinese developers, see Bradford, A. (2023). *Digital Empires: The Global Battle to Regulate Technology*. Oxford University Press. See also https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox

201 Council of the European Union (2025). "Declaration for European Digital Sovereignty." https://data.consilium.europa.eu/doc/document/ST-15781-2025-INIT/en/pdf

202 Clark, A. (2026). "Digital sovereignty." Research Briefing, UK Parliament, House of Commons Library. https://commonslibrary.parliament.uk/research-briefings/cbp-10547/

203 On the geopolitics of submarine cables, see the International Cable Protection Committee (ICPC) briefings on ownership and incidents, and reporting on cable cuts in the Red Sea (Feb. 2024), the Baltic Sea (Oct. 2023, Nov. 2024), and the Taiwan Strait (Feb. 2023, Jan. 2025). For a synthesis, see Starosielski, N. (2015). *The Undersea Network*. Duke University Press; and subsequent policy analyses from the Center for Strategic and International Studies and the Royal United Services Institute (2024–2026).

204 Starosielski, N. (2015). https://doi.org/10.2307/j.ctv11smhj2

SUPPLY CHAINS AND ENVIRONMENTAL IMPACT: CHIPS, MINERALS, ENERGY, CO_2, WATER

Just as energy supplies have increasingly been used as levers of geopolitical influence, so too could access to AI applications, data, and the material components of AI infrastructure. We have already seen international disputes over the adoption of 5G networks and export controls of advanced AI chips.[205] Over 90% of advanced chip manufacturing is done in Taiwan—which is itself a focus of geopolitical conflict over its sovereignty. The Dutch company ASML holds a near-total monopoly on the critical lithography machinery required to manufacture advanced computer chips. Control over many of the key materials for chip manufacture, such as rare earth metals and helium, may potentially be used for political leverage. Countries are increasingly concerned about risks to supply chains of raw materials and essential components.

Just as energy supplies have increasingly been used as levers of geopolitical influence, so too could access to AI applications, data, and the material components of AI infrastructure.

Energy has emerged as an increasingly binding constraint. Frontier training runs and large-scale inference now consume electricity on the scale of small cities, and the siting of AI data centers is shaped by grid capacity, access to water for cooling, and the availability of low-carbon power. Data centers may put pressure on existing power infrastructure, especially with their ever-increasing electricity demand. This is pulling new actors into the geopolitical map of AI: jurisdictions with abundant energy and permissive planning regimes—including Gulf states such as the United Arab Emirates and Saudi Arabia, but also the Nordic countries and parts of North America—are competing to host compute, while traditional AI hubs face grid bottlenecks.[206] Currently, AI data centers consume about 0.6% of global electricity, although consumption is predicted to grow by 84% in 2026. AI data center energy growth in 2027 is predicted to be 48%, a slowdown in growth for two reasons: first, it is difficult to get the permitting and resources to build ever-larger data centers, and second, data centers are becoming more energy-efficient. For example, Europe's JUPITER AI supercomputer is 100% powered by renewable energy and is 20 to 40 times more efficient than typical AI data centers.

205 CHIPS and Science Act of 2022, Public Law 117–167; U.S. Bureau of Industry and Security, "Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items" (Oct. 2022; revised Oct. 2023). For the broader industrial and strategic context, see Miller, C. (2022). *Chip War: The Fight for the World's Most Critical Technology*. Scribner. See also Wang, J., Milne, C., Hu, J. Z., & Khan, F. (2024). *Navigating Geopolitics in AI Governance,* Oxford Global Society. https://oxgs.org/wp-content/uploads/2026/02/OXGS-Report-_-Navigating-geopolitics-in-AI-Governance.pdf

206 International Energy Agency (2024). *Electricity 2024: Analysis and Forecast to 2026*, chapter on data centers, AI, and cryptocurrencies. https://www.iea.org/reports/electricity-2024. On Gulf compute strategies in particular, see press coverage of the UAE's G42 and Saudi Arabia's HUMAIN through 2024–2026.

The energy that data centers use must come from somewhere, and at the moment much electricity production derives from fossil fuels. In this way, data centers contribute to CO_2 production and hence climate change. By one estimate, training the GPT-3 model led to 500 metric tons of CO_2 emissions,[207] the equivalent of three intercontinental airliner flights. Global CO_2 emissions from data centers as a whole are predicted to rise sharply, and AI is a key driver of this growth.[208] On the other hand, according to studies by the International Energy Agency (IEA) and McKinsey, AI's ability to optimize complex systems yields carbon savings that are three to four times larger than the carbon footprint of the data centers themselves. Nevertheless, this is a widely cited concern about data centers.

The issue of water usage is complicated because assessing the impact depends on where the water comes from, how it is used, whether it is potable before and/or after use, and where it goes next. Collecting and evaporating some rainwater in rainy Seattle has little environmental impact. Drawing from a stressed aquifer in Arizona can have negative consequences, and it is these kinds of consequences that led to a national day of protest against data centers on July 18, 2026. AI data centers account for roughly 0.1% of water usage in the U.S., which is a fifth of the usage for golf courses, a twentieth of the usage for watering residential lawns, and a fiftieth of the usage for growing corn. Few jurisdictions require data centers to report resource use, including electricity and water.[209]

INFORMATION OPERATIONS AND COGNITIVE WARFARE

Viewed through a geopolitical lens, sharing access to the data of a country and its citizens creates cybersecurity, disinformation, and cognitive warfare vulnerabilities. Foreign adversaries, as well as criminal and nonstate actors, might gain intelligence through or disrupt AI services. Bad-faith actors have used AI to shape public opinion and to manipulate the behavior of citizens, such as influencing democratic elections. Foreign propaganda and influence are not new, but AI has increased the scale, quality, and reach of these tactics dramatically, while also lowering the cost and time needed to implement them. While dramatic and viral deepfakes are often the focus of these concerns, there has been massive growth in AI-controlled "sock puppets" and AI-generated influencers across social media platforms. These accounts can pose as patriotic human citizens while advancing the strategic communication messages and propaganda of external forces.

The globalization of the AI security stack is a related concern. Chinese integrated surveillance systems—cameras, analytics, and "Safe City" platforms—have been exported widely across

207 Patterson, D., Gonzalez, J., Le, Q. et al. (2020). "Carbon Emissions and Large Neural Network Training." https://arxiv.org/pdf/2104.10350

208 See, for example, de Vries-Gao, A. (2026). "The carbon and water footprints of data centers and what this could mean for artificial intelligence." *Patterns* 7:1. https://doi.org/10.1016/j.patter.2025.101430

209 Srinivasa Desikan, B., & Neff, G. (2025). *Evidence Review: Big Tech's Climate Performance and Policy Implications for the UK*. Minderoo Centre for Technology and Democracy. https://doi.org/10.17863/CAM.119617

Africa, Southeast Asia, Latin America, and the Middle East, often bundled with financing and training.[210] Western firms have followed their own trajectory: Palantir's Gotham and AIP platforms are now deployed in settings ranging from U.S. immigration enforcement to UK health data.[211] In each case, the export or deployment of AI-enabled surveillance raises questions about human rights, accountability, and whether recipient societies are gaining capacity or ceding control.

AI IN CYBER OPERATIONS

Cyber operations have been a frontline domain of interstate competition since long before the current wave of AI, but advanced models are now rapidly changing the offense–defense balance. Frontier systems have demonstrated the ability to autonomously discover, chain, and exploit software vulnerabilities at scale, lowering both the cost of offensive operations and the skill floor required to conduct them.[212]

An inflection point came in April 2026 with Anthropic's announcement of Claude Mythos Preview, a frontier model the company declined to release publicly after internal testing found that it could autonomously identify thousands of zero-day vulnerabilities across every major operating system and web browser—more than 99% of them unfixed at the time of disclosure. Shortly after the announcement, a third-party vendor environment was reportedly compromised, granting unauthorized parties access to Mythos Preview—a vivid illustration of how the very artifacts that expand defensive capability also concentrate offensive risk.[213]

In parallel, Anthropic launched Project Glasswing, a coalition bringing together Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, Nvidia, and Palo Alto Networks to use Mythos Preview defensively across critical software. At around the same time, however, we saw a public dispute between Anthropic and the U.S. Department of Defense over the terms under which Anthropic's models could be used in offensive military operations, following reporting that Claude—deployed via Anthropic's partnership with Palantir—had contributed to the January 2026 operation that seized Venezuelan leader Nicolás Maduro.[214] This episode forebodes a broader geopolitical

210 On Chinese AI-surveillance exports, see Feldstein, S. (2021). *The Rise of Digital Repression*. Oxford University Press; and the Carnegie AI Global Surveillance Index. On the "Safe City" model, see Greitens, S. C. (2020). "Dealing with Demand for China's Global Surveillance Exports." Brookings Global China. https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200428_china_surveillance_greitens_v3.pdf

211 On Palantir's footprint in UK health data, see successive investigations by the *Financial Times* (2020–2026). For broader framing, see Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.

212 For recent surveys of AI and cyber offense/defense, see "The near-term impact of AI on the cyber threat," UK National Cyber Security Centre assessment (Jan. 2024, updated 2026). https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

213 See https://red.anthropic.com/2026/mythos-preview/; https://www.scientificamerican.com/article/what-is-mythos-and-why-are-experts-worried-about-anthropics-ai-model/; and https://www.theguardian.com/technology/2026/apr/22/what-is-anthropic-mythos-ai-threat-global-cybersecurity. On the reported compromise, see https://www.euronews.com/next/2026/04/22/hackers-breach-anthropics-too-dangerous-to-release-mythos-ai-model-report

214 On the ensuing dispute, see https://www.nbcnews.com/tech/security/anthropic-ai-defense-war-venezuela-maduro-rcna259603 and https://www.washingtonpost.com/technology/2026/02/22/pentagon-anthropic-ai-dispute/

tension: commercial AI developers, some with the economic power of midsized countries, increasingly set the terms of state access to their most capable systems, and governments are increasingly unwilling to accept those terms.

MILITARY AI AND THE ACCELERATING ARMS RACE

Military power has long been a key element of geopolitical influence. The application of AI in the military domain is not new—much early AI research was funded by the U.S. Department of Defense. But after many decades of research, AI appears poised to transform many elements of military intelligence, planning, and operations. The development of lethal autonomous weapons systems (LAWS), which can select and engage targets without human supervision, has been a focus of discussion at the United Nations for more than a decade, even as their technological development and military deployment have rapidly accelerated.[215] Beyond drones and weapons platforms that autonomously select and engage targets, AI is increasingly generating targeting lists and providing decision support to military planners and decision-makers.[216] Use of these technologies has been reported in recent military operations in Ukraine, Gaza, Venezuela (see above), and in U.S. planning for strikes against Iran.

The acceleration and increasing intensity of AI-driven warfare poses a direct threat to civilians in armed conflict. Mistakes, hallucinations, flawed training data, and intentional sabotage or spoofing of AI systems can all result in attacks on noncombatants.

Military AI promises to accelerate the speed and increase the accuracy of military planning and decision-making, though little independent data exists to evaluate these claims. These systems also threaten to undermine human military and political control over warfare, and to obscure individual and state responsibility for the use of military force. The acceleration and increasing intensity of AI-driven warfare poses a direct threat to civilians in armed conflict. Mistakes, hallucinations, flawed training data, and intentional sabotage or spoofing of AI systems can all result in attacks on noncombatants.[217]

The perceived military advantages of AI adoption are already driving an AI arms race in which the accelerating development, adoption, and deployment of AI within the military is becoming a goal in itself.[218] This risks promoting the adoption of systems that are not fit for purpose. The

215 Boulanin, V., Davison, N., Goussac, N., & Peldán Carlsson, M. (2020). *Limits on Autonomy in Weapon Systems*. International Committee of the Red Cross. https://www.icrc.org/en/document/limits-autonomous-weapons. Blanchard, A., & Goussac, N. (2025). *Towards a Multilateral Policy on Autonomous Weapon Systems*. Stockholm International Peace Research Institute. https://doi.org/10.55163/VLLZ7787

216 Blanchard, A., & Bruun, L. (2025). *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting: A Comparison*. Stockholm International Peace Research Institute. https://doi.org/10.55163/YQBY3151

217 https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026; Boulanin et al. (2020). https://www.icrc.org/en/document/limits-autonomous-weapons

218 Asaro, P. (2019). "What Is an 'AI Arms Race' Anyway?" *I/S: A Journal of Law and Policy for the Information Society* 15:1–2, pp. 45–64. https://peterasaro.org/writing/Asaro_AIArmsRace.pdf. See also https://www.nytimes.com/2026/04/12/technology/china-russia-us-ai-weapons.html

increasing complexity of AI systems, and the lack of transparency and explainability of their decisions and recommendations, pose further problems for the assessment of their effectiveness, as well as for tracing accountability for the actions that result.



<< I guess landmines were autonomous weapons too. They weren't smart, but they sure had patience! >>

While the world continues to struggle with the indiscriminate harm caused by primitive weapons like landmines, AI introduces additional challenges. Increasing the physical and psychological space between a decision and its consequence—known as moral distance—can lower the human threshold for inflicting violence.[219] AI pushes this abstraction to its absolute limit by delegating the execution of a greater portion of the "kill chain" to machines. This ultimate detachment not only threatens to escalate conflict by insulating decision-makers from the realities of their choices, but it also introduces profound challenges for accountability and the enforcement of international humanitarian law.

219 Villegas-Galaviz, C., & Martin, K. (2023). "Moral distance, AI, and the ethics of care." *AI & Society* 39, pp. 1695–1706. https://doi.org/10.1007/s00146-023-01642-z

SUPERPOWERS, MIDDLE POWERS, AND THE GLOBAL SOUTH

Questions arise about geopolitical alignments and the centers of power around the globe, similar to considerations around the development of nuclear weapons in the 20th century. The development of the most advanced AI models requires enormous computational resources and data, as well as engineering expertise, and as such is largely limited to a few companies in a few countries. There is speculation over whether this will result in the dominance of a single AI superpower, or of several, as well as what the status of any middle AI powers might be.[220]

A second tier of states has begun to position itself actively within this emerging order. The UAE's Technology Innovation Institute and G42, Saudi Arabia's HUMAIN, India's IndiaAI Mission, the UK's AI Security Institute (formerly the UK AI Safety Institute), and coordinated EU efforts around sovereign compute all reflect strategies to secure a seat at the table without directly matching U.S. or Chinese frontier capability.[221] These so-called middle powers' strategies include variously hedging between blocs, hosting compute in exchange for access, investing in domestic model development, and shaping international standards and safety work.

There are similar concerns over whether the economic, political, and military benefits of AI will be concentrated in the Global North, and whether and how they might be shared with the Global South. Will AI underwrite new forms of imperialism and colonialism, or empower historically disadvantaged states?[222]

TOWARD GLOBAL AI GOVERNANCE?

It is currently uncertain whether and how AI can be governed at a global level. Can those countries that are leading the development be convinced to participate in governance structures that may limit the perceived advantages of the relentless pursuit of increasing capabilities? And how can those countries left behind meaningfully participate in decisions over how the technology is developed and applied?[223]

Since 2023, a patchwork of partially overlapping initiatives has emerged. The EU AI Act, adopted in 2024,[224] is the first comprehensive binding regulation of AI. The UK-initiated AI Safety (now Security) Summit process—Bletchley Park in 2023, Seoul in 2024, Paris in 2025, India in 2026,

220 Wang, Milne, Hu & Khan (2024). https://oxgs.org/wp-content/uploads/2026/02/OXGS-Report-_-Navigating-geopolitics-in-AI-Governance.pdf. And A. Nadibaidze & J. Lingevicius (Eds.). (2026). *Geopolitics of AI Governance: Actors, Norms, and Trends*, Center for War Studies, University of Southern Denmark. https://www.sdu.dk/en/forskning/forskningsenheder/samf/cws/cws-activities/2026/geopolitics-of-ai-governance

221 https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox. On specific national strategies, see https://www.aisi.gov.uk/; https://indiaai.gov.in/ministries/ministry-of-electronics-and-information-technology; and coverage of UAE G42 and Saudi HUMAIN partnerships in the *Financial Times* and the *New York Times*, 2024–2026.

222 World Bank Group (2024). *Digital Progress and Trends Report 2023*. https://openknowledge.worldbank.org/server/api/core/bitstreams/95fe55e9-f110-4ba8-933f-e65572e05395/content; Kwet, M. (2019). "Digital colonialism: US empire and the new imperialism in the Global South." *Race & Class* 60:4, pp. 3–26. https://doi.org/10.1177/0306396818823172. Couldry, N., & Mejias, U. (2019). *The Costs of Connection*. Stanford University Press.

223 Wang, Milne, Hu & Khan (2024). https://oxgs.org/wp-content/uploads/2026/02/OXGS-Report-_-Navigating-geopolitics-in-AI-Governance.pdf. Nadibaidze & Lingevicius (2026). https://www.sdu.dk/en/forskning/forskningsenheder/samf/cws/cws-activities/2026/geopolitics-of-ai-governance

224 Regulation (EU) 2024/1689 ("AI Act"), adopted June 13, 2024. https://artificialintelligenceact.eu/

and a planned summit in Switzerland in 2027—has produced joint declarations and a growing network of national Safety and Security Institutes. In the military domain, discussions on the regulation of lethal autonomous weapons have been ongoing in the United Nations' Convention on Certain Conventional Weapons since 2014, and are expected to produce a draft of a potential international legal instrument by the end of 2026. There has been a series of diplomatic meetings and a Global Commission on Responsible AI in the Military Domain (REAIM).[225] The Council of Europe's Framework Convention on AI, the G7 Hiroshima AI Process, and the UN Global Digital Compact add further strands.[226] Pope Leo XIV's first encyclical spoke of "safeguarding the human person in the time of artificial intelligence," with emphasis on the use of AI in warfare. Whether these will cohere into a durable international regime, fragment along geopolitical lines, or be captured by the interests of the most advanced AI states remains the central open question of AI geopolitics.

In conclusion, we see that in 2026 competition between the United States and China has intensified. It is no longer a single race but a competition across multiple dimensions: the supply chain for computing power and energy, the development of frontier and open-weight models, and deployment to consumers and middle powers. The United States retains a clear lead on frontier models while China has advanced an open-source strategy designed to embed its technology in the world's AI infrastructure. The result is two ecosystems, two regulatory philosophies, and a growing set of countries hedging between them.

DEEP DIVE TWO: The Scaling Era

The last decade of artificial intelligence developments was largely inspired by *scaling laws*, first described by Jared Kaplan and later refined by Jordan Hoffmann, which suggest that the performance of an AI model will improve by a predictable amount with a given increase in computational resources and training data size.[227] These are not laws of nature (like the law of universal gravitation); rather they are experimentally observed patterns (like Moore's Law, which states that the number of transistors on a microchip doubles roughly every two years). Like Moore's Law, the AI scaling regularities are remarkably robust: you can accurately predict the performance of a model with a thousand parameters as well as one with a trillion. The difference is the speed: AI models' capabilities have been doubling every three to six months, not every two years. In the words of deep learning pioneer Ilya Sutskever: "If you mix some compute with

225 https://www.exteriores.gob.es/en/REAIM2026/Paginas/Cumbre26.aspx; https://hcss.nl/gcreaim/. Afina, Y. (2026). "The Global Prism of Military AI Governance: Reflections from the 2025 Regional Consultations on Responsible AI in the Military Domain." United Nations Institute for Disarmament Research. https://unidir.org/publication/the-global-prism-of-military-ai-governance-reflections-from-the-2025-regional-consultations-on-responsible-ai-in-the-military-domain/

226 Council of Europe (2024), *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. CETS No. 225. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence; G7 Hiroshima AI Process Comprehensive Policy Framework (2023), https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html; United Nations General Assembly, *Global Digital Compact* (Resolution A/RES/79/1, 2024), https://docs.un.org/en/a/res/79/1

227 See, for example, Kaplan et al. (2020). https://arxiv.org/pdf/2001.08361/1000

some data into a neural net of a certain size, you will get results, and you will know that it will be better if you just scale the recipe up. [...] Companies love this because it gives you a very low-risk way of investing your resources." This observation characterizes the scaling era in AI.

However, more recent empirical work suggests that the predictive power of scaling laws may be more limited than initially assumed. While early studies demonstrated smooth improvements with scale,[228] subsequent analyses show that downstream performance and real-world capabilities often exhibit irregular, domain-dependent scaling behavior.[229] In some cases, smaller, more efficiently trained models outperform significantly larger systems, indicating that scale alone is not a reliable predictor of capability.[230] These developments suggest that the relationship between compute and performance is becoming increasingly complex, with gains now often driven by improvements in data quality, training procedures, model architecture, and inference-time compute rather than model size alone.

In some cases, smaller, more efficiently trained models outperform significantly larger systems, indicating that scale alone is not a reliable predictor of capability.

As a result, the scaling paradigm is increasingly complemented—and in some cases challenged—by approaches that emphasize sample efficiency and computational efficiency rather than sheer size. Advances such as improved data curation, model distillation, instruction tuning, and retrieval-augmented generation have demonstrated substantial performance gains without proportional increases in compute.[231]

It is not clear when the scaling era will end, or if it already is ending. Despite the release of ever-larger models, we have seen several smaller models rise to near the top of leaderboards.[232] Another quote from Sutskever implies that research may be more important than just scaling today: "Scaling the pre-training is essentially tapped out. If you just multiply the scale by 100x now, you won't get a qualitative change in capability."

228 Kaplan et al. (2020). https://arxiv.org/pdf/2001.08361/1000. Hoffmann, J., Borgeaud, S., Mensch, A. et al. (2022). "Training Compute-Optimal Large Language Models." NeurIPS. https://arxiv.org/abs/2203.15556

229 Ganguli, D., Hernandez, D., Lovitt, L. et al. (2022). "Predictability and Surprise in Large Generative Models." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–1764. https://doi.org/10.1145/3531146.3533229. Srivastava, A., Rastogi, A., Rao, A. et al. (2023). "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." https://arxiv.org/abs/2206.04615. Caballero, E., Gupta, K., Rish, I., & Krueger, D. (2023). "Broken Neural Scaling Laws." https://arxiv.org/abs/2210.14891. Hu, S., Liu, X., Han, X. et al. (2024). "Predicting Emergent Abilities with Infinite Resolution Evaluation." https://arxiv.org/abs/2310.03262

230 Gupta, A., Thomas, B., & Asnani, H. (2026). "Small Language Models (SLMs) Can Still Pack a Punch: A Survey." https://arxiv.org/html/2501.05465v2

231 Ouyang, L., Wu, J., Jiang, X. et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." https://arxiv.org/abs/2203.02155. Wei et al. (2022). https://arxiv.org/abs/2201.11903. Lewis, P., Perez, E., Piktus, A. et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." NeurIPS. https://arxiv.org/abs/2005.11401. Rafailov, R., Sharma, A., Mitchell, E. et al. (2023). "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model." https://arxiv.org/abs/2305.18290

232 Hooker, S. (2025). "On the Slow Death of Scaling." http://dx.doi.org/10.2139/ssrn.5877662

THE RACE TO SCALE: COMPUTE

The race to scale has had a number of important consequences.

First, the extreme compute resources required to build such models began to dramatically limit the number of organizations capable of developing frontier AI models. According to Epoch AI, the computing power used to train frontier language models has grown roughly fivefold per year since 2020, and the largest known AI data center now has the capacity of more than a million of Nvidia's H100 specialized AI chips (GPUs). The capital cost of acquiring such computing infrastructure would place it beyond the reach of all but the very wealthiest international universities, while the associated energy costs alone would dwarf the budgets of most academic research projects. As a result, the development of very large, state-of-the-art models became increasingly concentrated within a small number of companies worldwide, most notably the major Silicon Valley technology firms (see also discussion in SQ8). In turn, this has meant that the development of leading AI models increasingly occurs behind closed doors.

This concentration reflects a broader structural dynamic in which access to large-scale compute becomes a prerequisite for participation in frontier research, with breakthroughs increasingly clustered in well-capitalized organizations and regions.[233] The fact that general-purpose AI would in effect largely be owned by a small number of U.S.-based private companies prompted a number of nation-states to begin considering the possibility of sovereign AI capabilities, for example, investing in national and international schemes to develop independent AI infrastructure and GPT-class AI systems. In the U.S., shortly after his second-term inauguration, President Donald J. Trump announced an initiative called Stargate, aimed at developing a $500 billion strategic compute resource for U.S. AI.

A second major consequence of the race to scale was the dramatic increase in demand for both the semiconductors required to train these models and the data on which they depend. We examined the geopolitical implications in Deep Dive One: Supply Chains and Environmental Impact; here, it is sufficient to note that this demand led to an extraordinary increase in the valuations of leading AI chip manufacturers, most notably Nvidia, the company that pioneered the development of GPUs. In October 2025, Nvidia reached a market capitalization of $5 trillion, the largest of any company in the world, arguably making it the most valuable company in human history. This valuation exceeded the GDP of almost every nation on Earth. This extraordinary growth was driven purely by the demand for the semiconductors that are developed by Nvidia, which are the chip of choice for training AI models.

233 Maslej et al. (2024). https://hai.stanford.edu/ai-index/2024-ai-index-report. Longpre, S., Mahari, R., Chen, A. et al. (2023). "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI." https://arxiv.org/abs/2310.16787

More generally, the race to scale prompted vast investment in data centers: estimates suggest that if current demand continues then a further $6.7 trillion capital investment in data centers will be required. Many serious commentators suggest that this enormous capital investment, coupled with spend on semiconductors and ancillary hardware, is actually propping up the U.S. economy, possibly even preventing it from going into recession. The European AI Strategy builds upon the concept of cross-national centralized AI Gigafactories—massive, gigawatt-level data centers with up to 100,000 advanced AI processors designed to enable next-generation models.

The economic justification for these investments relies on the assumption of sustained returns to scale. Yet studies suggest that the marginal gains from additional compute may be decreasing over time.[234] If so, the long-term sustainability of continued exponential investment in infrastructure may be uncertain.

There has been considerable pushback about the scale of data center development, on a number of grounds. First, data centers in 2026 are physically enormous facilities; they are noisy and tend not to be aesthetically pleasing. It is therefore understandable that people living in the vicinity of a new data center development would resent the imposition of such facilities in their neighborhood. Some people protest the impact that large companies and technologies have on our society, and these protests have even led to violence. The environmental impact of data centers (energy, water, CO_2) is an important issue that we cover in Deep Dive One: Supply Chains and Environmental Impact. Data centers do provide some jobs, but only about one-tenth the number provided by a similarly sized traditional manufacturing facility.

Data centers do provide jobs, but only about one-tenth the number provided by a similarly sized traditional manufacturing facility.

The financial model underpinning data center development on the current scale has come into question. The argument for building data centers is that the new technology requires it, and that the up-front cost will later be recouped by the benefits gained. An analogy is often made with the development of railroads in the 19th century. Laying down railroads was enormously costly at the time, but the basic infrastructure had a long timeline: railroad tracks themselves are typically good for decades, and the infrastructure to support them can have a lifetime measured in centuries. By contrast, the computers that populate data centers often get replaced after several years as newer technologies become available. Thus not only is the current investment in data centers enormous but it has a very short window in which to show returns.

234 Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). "The Computational Limits of Deep Learning." https://arxiv.org/abs/2007.05558

Finally, the extraordinary scale of current investment in AI infrastructure has raised serious questions about economic sustainability. Much of this investment has been financed through borrowing and external capital, including pension funds and private equity, rather than solely through retained earnings of technology firms. Despite the bullish claims of big tech companies, it is far from clear whether anticipated returns justify this level of expenditure, especially as models are often improved by training new versions rather than refining existing ones. In short, the scale of investment in data centers risks a market crash unless markets remain convinced that investments will be recouped.

THE RACE TO SCALE: DATA

The race to scale also triggered an intense scramble for training data. Crucially, compute-optimal training regimes (for example, Chinchilla) show that the training data required for a model must scale proportionally with model size: larger models require comparably larger training datasets in order to show improvements. And large language models are particularly data-hungry. Although precise figures for more recent models are not publicly available, OpenAI disclosed that GPT-3 was trained on approximately 45 terabytes of textual data. At this scale, the World Wide Web provides the only realistic source of such volumes of text, and it quickly became standard practice to scrape text from virtually every available online resource. These resources included online public data sources such as Wikipedia, social media sources such as Twitter, and discussion forums such as Reddit, as well as commercial and governmental websites.

The widespread use of data sources like the World Wide Web provoked significant concerns about copyright. If an LLM has effectively ingested substantial portions of the World Wide Web, it has necessarily absorbed vast quantities of copyrighted material. Unsurprisingly, this led to widespread objections from rights holders and, ultimately, to a growing number of legal challenges in multiple jurisdictions. AI model providers argued that the use of these data sources in training AI models did not represent a copyright infringement since, firstly, LLMs do not store copyrighted works in any conventional sense (one cannot in general point to any specific part of an LLM's neural networks and say, "That's where my data is stored"), and, secondly, LLMs are not derivative works in the sense of copyright law. Copyright law was simply not designed with AI models and neural networks in mind.

There are several open lawsuits, such as the one where the *New York Times* sued OpenAI and Microsoft over their use of copyrighted work. We cannot judge at this point how the many lawsuits underway will pan out; nor can we predict what the legal landscape around data, copyright, and AI models will ultimately look like. Summer 2025 court decisions in the U.S. have suggested AI use of copyrighted material for training is fair use, including cases against Meta and Anthropic. At the same time, Anthropic was found to have used books in training without paying for them and eventually settled the lawsuit related to the pirated books with a $1.5 billion payout. This is the largest copyright settlement in legal history (by some considerable distance).

Diverse and large datasets are valuable. Moreover, because large training runs are costly and therefore infrequent, systematic experimentation to determine how best to use limited data sources is not feasible. Because data is valuable, it is increasingly treated as a strategic asset. Frontier models are now trained on mixtures of public, licensed, and proprietary datasets, with exclusive access to datasets conferring significant competitive advantages for model providers.

This shift has intensified legal and political scrutiny over data usage, exposing fundamental ambiguities in existing legal frameworks that cover large-scale data ingestion and transformation. Much modern digital infrastructure now embeds data collection along with the services it provides. Televisions, self-driving vehicles, wearable devices, doorbell cameras, smart chatbot interactions, police body cameras, and more have all become platforms for data collection in addition to their primary purpose, leading to many objections. Regulatory regimes such as the EU's General Data Protection Regulation (GDPR) have begun to impose constraints on data collection and reuse, though they will need to adapt to the emerging realities of generative AI. The result is a fragmented governance landscape in which data access is simultaneously a technical, economic, and legal problem.

Frontier models are now trained on mixtures of public, licensed, and proprietary datasets, with exclusive access to datasets conferring significant competitive advantages for model providers.

In response to the limitations of static datasets, there is growing interest in physical AI, encompassing robotics, simulation environments, and world models that allow for interaction rather than just passive observation. By generating data through embodied experience—whether in real-world systems or high-fidelity simulations—these approaches offer a pathway to bypass the constraints of preexisting datasets. This paradigm is particularly relevant for domains such as robotics, navigation, scientific discovery, and complex system modeling, where relevant data cannot be easily scraped from the internet but must instead be produced through experimentation or simulation.

Another obvious response to data scarcity is to use *synthetic* data.[235] While synthetic data has demonstrated value in specific contexts, it is not a universal substitute for real-world data, in part due to risks such as reinforcement of biases in the model. Instead, it is best understood as complementary to ground-truth data. More broadly, its increasing use reflects a shift toward more flexible training regimes in which data distributions can be actively shaped rather than passively accumulated.[236]

235 https://www.weforum.org/stories/2025/12/data-ai-training-synthetic/

236 Niklaus, J., Yamaguchi, A., Štefánik, M. et al. (2026). "How Can We Synthesize High-Quality Pretraining Data? A Systematic Study of Prompt Design, Generator Model, and Source Data." https://arxiv.org/abs/2604.13977. Aryabumi, V., Dang, J., Talupuru, D. et al. (2024). "Aya 23: Open weight releases to further multilingual progress." https://arxiv.org/abs/2405.15032

A growing body of work demonstrates that improvements in data quality—through deduplication, pruning, and prioritization—can reduce the need for larger datasets while maintaining or improving performance.[237] These findings suggest a shift from treating data as a quantity to be maximized toward treating it as a resource to be optimized, with carefully curated datasets often yielding greater value than continued expansion of raw data volume.[238]

SCALING AND THE FUTURE OF AI

For the reasons described above, the future may not be a straightforward extrapolation of scaling of the kind we have seen over the recent past, but may instead require more innovative ways to use limited data and compute resources to build and run AI models.[239] Scaling laws remain empirically useful within constrained regimes, but they no longer provide a sufficient explanation for frontier capability gains.

Parallel to the scaling trend, a second current of innovation is focused on efficiency. Algorithmic advances have played a key role in recent progress. Approaches such as instruction tuning, low-rank adaptation, quantization, distillation, mixtures of experts, KV cache compression, speculative decoding, and weight sharing demonstrated that smaller, less expensive models could approach the performance of frontier models. This development has widened access to frontier methods, enabling broader experimentation and lowering barriers to entry. Gains from algorithmic efficiency may rival or exceed those from scaling alone.[240]

The field is increasingly moving beyond training monolithic models from scratch on fixed datasets. Emerging approaches emphasize continual learning, in which models are updated incrementally as new data becomes available, and modular architectures, which decompose intelligence into interacting subsystems rather than using a single, static network. Empirical progress in this direction suggests that future AI systems may rely less on one-time large-scale data ingestion and more on sustained, dynamic interaction with evolving data environments—agents improve continuously by acting, observing outcomes, and incorporating feedback, producing effectively unbounded streams of data that scale with capability rather than dataset size.[241]

237 Lee, K., Ippolito, D., Nystrom, A. et al. (2022). "Deduplicating Training Data Makes Language Models Better." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* 1: Long Papers. https://aclanthology.org/2022.acl-long.577/. Sorscher, B., Geirhos, R., Shekhar, S. et al. (2022). "Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning." NeurIPS. https://arxiv.org/abs/2206.14486

238 Penedo, G., Kydlíček, H., Ben allal, L. et al. (2024). "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale." NeurIPS. https://arxiv.org/abs/2406.17557. Ali, M., Brack, M., Lübbering, M. et al. (2025). "Judging Quality Across Languages: A Multilingual Approach to Pretraining Data Filtering with Language Models." EMNLP. https://aclanthology.org/2025.emnlp-main.449/

239 Hooker (2025). http://dx.doi.org/10.2139/ssrn.5877662

240 See, for example, DeepSeek-AI, Guo, D., Yang, D. et al. (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." https://arxiv.org/abs/2501.12948

241 Silver, D., & Sutton, R. S. (2025). "Welcome to the era of experience." Google DeepMind. https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf

In addition, increasing attention is being paid to inference-time computation, where additional compute is applied during deployment rather than training. Techniques such as search, tool use, and agent-based coordination have been shown to significantly enhance performance without modifying model parameters.[242] This represents a notable departure from traditional scaling approaches (and has implications for increased computational costs).

The scaling era has been a defining feature of recent AI progress, but scaling is now evolving beyond the straightforward expansion of model size, training data, and compute.

The scaling era has been a defining feature of recent AI progress, but scaling is now evolving beyond the straightforward expansion of model size, training data, and compute. While these dimensions enabled rapid early gains, their limitations—particularly diminishing returns, inefficiencies, and concentration effects—suggest that future progress will depend increasingly on data efficiency, algorithmic design, and system-level integration. Longer contexts, autonomous agents, personalized systems, and deployment across populations and domains broaden what it means to scale, but also make reliability harder to establish: benchmarks can conceal uneven performance and agentic workflows can exceed effective human oversight. The central challenge is therefore no longer maximizing a single dimension of scale but optimizing the interaction between compute, data, and algorithms while also scaling evidence of validity, robustness, representativeness, and societal value.

242 Wei et al. (2022). https://arxiv.org/abs/2201.11903. Wang, X., Wei, J., Schuurmans, D. et al. (2023). "Self-Consistency Improves Chain of Thought Reasoning in Language Models." ICLR. https://arxiv.org/abs/2203.11171. Khairi, A., D'sousa, D., Shen, Y. et al. (2025). "When Life Gives You Samples: The Benefits of Scaling Up Inference Compute for Multilingual LLMs." EMNLP. https://arxiv.org/abs/2506.20544

Conclusions

While the period of this report has been characterized by continual improvements in AI across a broad spectrum, it is the emergence of generative AI that brought AI into the mainstream, and which has dominated the public debate about AI and its potential economic impacts. This is a transformative period: a large and growing fraction of humanity now has access to, and makes regular use of, powerful general-purpose AI tools—although not (yet) AGI. And the capabilities of these models have improved apace, driven in part by technological advances and sophisticated engineering, but also by the relentless pursuit of scale. Capital investment in support of AI has reached extraordinary levels, unprecedented in the history of technology. As a consequence, at the time of writing, many fear that the present AI boom is actually a bubble that will at some point burst. Progress is increasingly constrained by compute, data availability, and the concentration of model-building capacity among a small number of actors. As we write, the heads of several major AI labs suggest that AGI will be here within the next five years; other equally qualified experts doubt this.

Current AI systems remain jagged: on one prompt they perform at expert level, on the next they exhibit factual errors, bias, or lack of commonsense reasoning.

For all the very real advances we have witnessed, current AI systems remain jagged: on one prompt they perform at expert level, on the next they exhibit factual errors, bias, or lack of commonsense reasoning. Other problems include reward hacking (where models behave in unintended and unhelpful ways by finding loopholes or shortcuts in a given reward scheme while not accomplishing goals in the intended way) and privacy and copyright issues.

Attempts to introduce governance frameworks around AI have been fragmented and uneven, and there is significant pushback against regulation in certain quarters, driven by a fear that regulation will stifle innovation. The period of our report has seen the first serious attempts at international coordination on AI safety and security, although these efforts have so far failed to produce a clear consensus.

The societal footprint of generative AI is already evident. Ever-improving AI models are transforming our working, leisure, and social lives, and beginning to transform the nature of work, albeit in ways that remain uncertain. How AI will affect society at large is a question that grows more prominent daily. And with increasing AI adoption, public attitudes toward AI have grown much more ambivalent. There is anxiety, especially about the impact of AI on human society, the environment, and the displacement of labor. The effects on labor markets have

started to become visible: many professions are being augmented, portions of knowledge work are being automated, and concerns about job quality, entry-level opportunities, the future of some occupations, and the future of education have grown.

We are now at a point in the history of AI that researchers in the field have dreamt about for decades. Powerful, general-purpose AI tools, of the kind widely anticipated in science fiction, are now a routine reality, used by millions every day. On the one hand, this is cause for celebration: it presents humanity with truly remarkable opportunities for making a better world, by curing diseases, solving energy and climate issues, and empowering anyone to connect with the sum of human-accumulated knowledge. But it also presents us with a range of challenges: ensuring that the benefits of AI will be shared, protecting against errors, intentional misuses, and unintended side effects, managing shifts in employment markets, preventing the de-skilling of students and workers, and ensuring that less tangible risks, such as AI psychosis, are mitigated as far as possible. In this respect, the key message we present from this report is entirely consistent with its predecessors.

Five Years On: Where the 2021 Panel Landed

The 2021 Study Panel described a field that was about to be transformed. The report correctly anticipated that a transformation was about to happen, but, along with much of the AI community, failed to predict the nature of the transformation: the rise of LLMs and generative AI.

Five years on, a number of the claims of the 2021 report have borne out—often more decisively than its authors anticipated. But four calls, viewed from April 2026, are worth explicitly revisiting: the panel's judgment that "the recent dominance of deep learning may be coming to an end" (SQ12, p. 62); its conjecture that AI systems "perhaps will never gain common sense or general intelligence without being more tightly coupled to the physical world" (SQ5, p. 33); the single-sentence treatment of losing meaningful control over AI systems as a background concern (SQ10, p. 53); and the near absence of any anticipation that, within 14 months of publication, a general-purpose chat interface would reach 100 million users and reshape the public's relationship with the technology.[243]

These four items perhaps share a common cause. The panel, drawn almost entirely from inside the field, underweighted the consequences of continuing to scale the transformer architecture that the report described well. The limits they foresaw for deep learning were quickly surpassed; the prerequisite they imagined for generality (embodiment) was not required to produce exciting new capabilities; the risk they noted in passing became the organizing frame of global AI policy; and the product form factor they did not anticipate—i.e., a chat window onto a very large language model—became the technology's arrival into the lives of ordinary people. None of this undercuts the value of the 2021 report, but it does suggest a working lesson for our own panel: where the field's own leading insiders converge on what is ahead, they may be converging on the previous paradigm's limits, not on the next paradigm's trajectory.

The remainder of this section works through the full shortlist. We separate points of divergence (where claims or recommendations in the 2021 report depart from what subsequently happened) from points of convergence (where the report's reading largely survives contact with events), and collect, in a third section, "Beyond the Frame," developments that the 2021 panel's framing left out entirely. Each entry cites the 2021 report verbatim, with page numbers of the published PDF.

243 ChatGPT reached approximately 100 million monthly active users within two months of its November 2022 launch (UBS/Similarweb estimates, reported Feb. 2023), the fastest consumer-technology adoption on record.

Divergence

LANGUAGE MODELS AND "DEEP UNDERSTANDING" — SQ2, p. 13

> *"[T]he models themselves do not exhibit deep understanding of the texts that they process, fundamentally limiting their utility in many sensitive applications. Part of the art of using these models, to date, is in finding scenarios where their incomplete mastery still provides value."*

By 2026, LLMs have been deployed across precisely the sensitive applications the panel had in mind: clinical decision support, legal drafting, scientific research, and software engineering. The "incomplete mastery, so keep them in niches" framing underestimated how much partial competence paired with human review would prove useful in practice. The underlying technical claim—that these models lack deep understanding—remains contested rather than settled: capability has consistently outrun the frameworks used to dismiss it.

SUPERHUMAN ON BENCHMARKS, BELOW HUMANS ON REAL QUESTIONS — SQ2, p. 13

> *"Performance on question-answering benchmarks [...] have reached superhuman levels, although the models that achieve this level of proficiency exploit spurious correlations in the benchmarks and exhibit a level of competence on naturally occurring questions that is still well below that of human beings."*

Since 2024, frontier language models have been performing at or near expert human level on professional bar and medical examinations and on naturally occurring questions across law, medicine, coding, and graduate-level science, including on benchmarks specifically designed to be robust against shortcut-learning (MMLU, GPQA). The level of competence is no longer "well below" humans.

MATHEMATICAL DISCOVERY AS AN OPEN GRAND CHALLENGE — SQ3, p. 18

> *"Others have remained open: mathematical discovery, a self-replicating system that enables a small set of machine tools to produce other tools using locally available raw materials."*

The "others" here refers to Raj Reddy's 1988 grand challenges for the field.

Mathematical discovery moved dramatically. Open problems in mathematics have been proved by AI alone, or by a human–AI partnership.[244] DeepMind's AlphaProof and AlphaGeometry 2 reached IMO silver-medal level in 2024, and in July 2025 both Gemini Deep Think and

244 Schmidt, J. (2025). "Extremal descendant integrals on moduli spaces of curves: An inequality discovered and proved in collaboration with AI." https://arxiv.org/abs/2512.14575. Tsoukalas et al. (2026). https://arxiv.org/pdf/2605.22763v1

OpenAI achieved gold-medal level on the 2025 IMO, solving five of six problems end-to-end in natural language.[245] In 2026, Claude Fable found a counterexample that disproved the Jacobian conjecture, a problem that Steve Smale listed in 1998 as one of 18 critical unsolved problems for the 21st century. (Claude's one-line counterexample settles the question for $n \geq 3$, but leaves open the case of $n = 2$.)

GENERALIZATION WILL NEED NEUROSYMBOLIC INTEGRATION — SQ3, p. 22

> *"[A] robust capacity for generalizability and transfer learning will likely require the integration of symbolic and probabilistic reasoning—combining a primary focus on logic with a more statistical point of view."*

The generalization leap came primarily from scaling transformer pretraining, then reinforcement learning from human feedback, then test-time reasoning (chain-of-thought, and later dedicated reasoning models in the o1/o3 style). Neurosymbolic progress exists—AlphaProof and AlphaGeometry are genuine hybrids—but it has not yet been the mainline driver of generalization in frontier reasoning systems.

COMMON SENSE WILL REQUIRE PHYSICAL GROUNDING — SQ5, p. 33

> *"AI systems still remain very far from human abilities in all these areas, and perhaps will never gain common sense or general intelligence without being more tightly coupled to the physical world."*

This is one of the report's starkest surprises. Frontier language models display strong common-sense reasoning on a wide range of tasks without being coupled to the physical world. The gains came from scale, data, and post-training, not from embodiment. Whether full generality can be reached without physical grounding remains an open question; what has not held up is the stronger conjecture that the absence of grounding would block common-sense capability in the meantime.

PUBLIC UNDERSTANDING MOVING PAST "TERMINATOR" FEARS — SQ6, p. 33

> *"Since the first AI100 report, public understanding has broadened and become more nuanced, starting to move beyond Terminator and robot overlord fears."*

After ChatGPT's public release in November 2022, existential and "AI overlord" framings returned to the mainstream of AI discourse with force. A statement signed by hundreds of leading AI scientists and industry figures warned that mitigating the risk of extinction from AI

245 Google DeepMind (2025). "Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad." https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/. See also *Nature* news coverage.

should be a global priority,[246] the Bletchley Declaration in November 2023 was signed by 28 governments,[247] and a network of national AI Safety Institutes was established in short order.

LETHAL AUTONOMOUS WEAPONS: REGULATION STATUS — SQ7, p. 38

> *"Thirteen countries—in Africa, Europe, and Latin America—had discussed legislation on the use of autonomous lethal weapons [...] only Belgium had enacted law."*

Formal multilateral regulation of lethal autonomous weapons remained largely stalled through 2021 to 2025. Meanwhile, AI-enabled targeting systems were deployed at industrial scale in active conflicts: Israel's Gospel (building-target recommendations) and Lavender (person-target lists, at one point containing roughly 37,000 names) during the Gaza war,[248] Palantir-enabled targeting in Ukraine, and a proliferation of FPV drones and loitering munitions with substantial autonomous elements. The panel identified the issues, but the rapid exploitation of AI in the conflicts beginning in 2022, 2023, and 2026 developed in unexpected ways, and somewhat independently of attempts to legislate.

AI SUMMARIZATION AND CHAT COULD WIDEN INFORMATION HORIZONS — SQ9, p. 49

> *"Summarization and interactive chat technologies have great potential to help ensure that people get a healthy breadth of information on a topic, and to help break filter bubbles rather than make them—by providing a range of information, or at least an awareness of the biases in one's social-media or news feeds."*

Chat-based AI reached hundreds of millions of users faster than any prior consumer technology. Whether the hoped-for effect—broader, debiased information—has followed is unclear. Early evidence suggests that chat systems introduce their own homogenizing biases, sycophancy, and confident hallucinations, sometimes in precisely the contexts where the 2021 panel imagined them helping. The direction of impact on filter bubbles, if any, is an important open question for the 2026 report.

246 "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Center for AI Safety (2023). "Statement on AI Risk." https://aistatement.com/

247 UK Government (Nov. 1, 2023). "The Bletchley Declaration by Countries Attending the AI Safety Summit." Signatories included the U.S., the UK, EU member states, and China.

248 Schmitt, M. N. (2024). The Gospel, Lavender, and the law of armed conflict (Israel–Hamas 2024 Symposium). Articles of War, West Point Lieber Institute. https://lieber.westpoint.edu/gospel-lavender-law-armed-conflict/. Human Rights Watch (2024). Questions and Answers: Israeli Military's Use of Digital Tools in Gaza. https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-of-digital-tools-in-gaza

ALGORITHMS ARE FINE; THE BOTTLENECK IS DATA AND INTEGRATION — SQ9, p. 52

> *"In most cases, the main factors holding back these applications are not in the algorithms themselves, but in the collection and organization of appropriate data and the effective integration of these algorithms into their broader sociotechnical systems."*

There is no question that scaling data has been an essential driver of progress in the last five years. However, it has proceeded hand in hand with algorithmic and architectural breakthroughs layered on top of ever more compute: reinforcement learning from human feedback, mixture-of-experts routing, long-context training, multimodal fusion, and reasoning models. The premise that "the algorithms are fine" understated how much progress was still latent in the model-building stack itself.

THE MOST PRESSING DANGER IS TECHNO-SOLUTIONISM — SQ10, p. 53

> *"One of the most pressing dangers of AI is techno-solutionism, the view that AI can be seen as a panacea when it is merely a tool."*

Viewed from 2026, the dangers that have demanded the most policy attention are different ones: misuse at scale (deepfakes, election interference, industrial-scale fraud, automated cyberattacks), autonomous weapons, loss of control and alignment in highly capable systems, and concentration of power in a handful of firms. Techno-solutionism remains a real failure mode, but placing it at the top of the danger list has not held up as a reading of what would matter most in this period.

NO EVIDENCE OF ACCELERATING AUTOMATION ON AN AGGREGATE SCALE — SQ11, p. 59

> *"[H]ard evidence of accelerating automation on an aggregate scale is hard to find."*

Evidence has since become easier to find. Stanford's Digital Economy Lab reports a roughly 19% decline in employment of 22–25-year-olds in AI-exposed occupations (software engineering, customer support) relative to trend since ChatGPT's release;[249] Brynjolfsson, Li, and Raymond find a 15% average productivity increase—and a 36% increase for workers in the bottom skill quintile—from a Fortune 500 customer-support deployment;[250] and AI-generated code now accounts for a substantial share of professional software output.[251] U.S. productivity growth of roughly 2.7% in 2025—nearly double the decade average—is consistent with the early stages of a general-purpose-technology J-curve arriving sooner than the panel expected.

249 Brynjolfsson, Chandar & Chen (2025/2026). https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/

250 Brynjolfsson, Li & Raymond (2025). https://doi.org/10.1093/qje/qjae044

251 Industry surveys for 2025 cluster around 40–46% AI-assisted or AI-generated code among professional users of GitHub Copilot and comparable assistants. See GitClear, GitHub, and Stack Overflow developer-survey data for 2025.

TECHNOLOGICAL CHANGE TAKES DECADES — SQ11, p. 59

> *"Third, technological change takes place over a long time, oftentimes longer than expected. It took decades for electricity and the first wave of information technology to have a noticeable impact on economic data; any future wave of technological innovation is also unlikely to hit all corners of the economy at once."*

Generative AI decisively broke this historical pattern. ChatGPT reached 100 million users in two months—at the time, the fastest consumer-technology adoption on record.[252] Following its release, enterprise adoption of LLMs was rapid. The 2021 anchor on electricity-style diffusion timescales understated the speed of the current wave. While the take-up of the technology is very clear, what is less clear is whether this transformation will yield (for example) productivity gains as widely predicted.

GENERAL METHODS ARE RUNNING INTO LIMITS — SQ12, p. 61

> *"But now, in the 2020s, these general methods are running into limits, and many in the field are questioning how we best make continued progress."*

The limits turned out to be further off than the panel expected. Scaling laws continued to yield substantial gains through GPT-4 and GPT-5, Claude 3 and 4, Gemini 2 and beyond, Llama 3, and the DeepSeek open-weights frontier. New axes—instruction tuning, reinforcement learning from human feedback, chain-of-thought, test-time compute, multimodality, million-token contexts—repeatedly unlocked further capability. Of all the predictions in the 2021 report, the judgment that the general-methods era was closing was the one that aged most quickly.

DEEP LEARNING'S RECENT DOMINANCE MAY BE NEARING ITS END — SQ12, p. 62

> *"The recent dominance of deep learning may be coming to an end."*

Dominance extended, rather than ended. By 2026, the transformer has become the default architecture across language, vision, audio, video, protein structure, robotics (vision-language-action models), and control. The serious competitors that have emerged—state space models such as Mamba, diffusion transformers—are themselves deep-learning variants. No non-deep-learning paradigm has displaced the field.

NEUROSYMBOLIC AI AS THE MAIN PATH FORWARD — SQ12, p. 63

> *"[T]he burgeoning area of neurosymbolic AI, which unites classical symbolic approaches to AI with the more data-driven neural approaches, may be where the most progress towards the AI dream is seen over the next decade."*

252 UBS/Similarweb estimates, reported Feb. 2023.

Neurosymbolic integration has delivered impressive results—AlphaGeometry and AlphaProof couple language models with theorem provers, and a generation of LLM-plus-solver hybrids has produced real advances in mathematics. The last two years have seen a focus on reasoning models. At the time of writing we are seeing the extensive use of tools, such as symbolic reasoning systems, to augment neural AI model capabilities.

DATA-CENTER CARBON FOOTPRINT STABLE — SQ12, p. 62

> *"Fortunately, companies have been making data centers more efficient faster than they have been increasing in size and are mostly switching to green energy, keeping their carbon footprint stable over the past five years."*

Hyperscaler emissions have risen sharply in the scaling era. Microsoft reported that its emissions were up more than 23% against its 2020 baseline,[253] Google's Scope 1–3 emissions were roughly 50% higher over five years,[254] and several hyperscalers have returned to gas and restarted nuclear capacity to meet data-center load. Of the report's factual assertions, this was probably the one that reversed most clearly.

COMPLETE AUTONOMY IS NOT THE EVENTUAL GOAL — CONCLUSIONS, p. 71

> *"AI researchers should also recognize that complete autonomy is not the eventual goal for AI systems."*

This was a normative claim about what the field should aim for, not a prediction of what it would do.

By 2026, "agentic" AI—multistep workflows, computer use, and autonomous coding—is a leading commercial and research direction pursued by every major frontier lab. An important metric is the level of autonomy; the METR time horizon report says that the length of the task that an agent can complete on its own is doubling every seven months. But autonomy still operates on a spectrum, with humans in the loop and interacting at various levels, depending on the job.

253 Microsoft 2024 Environmental Sustainability Report and coverage, for example, https://energy.policyplatform.news/environment/microsoft-carbon-emissions-jump-ai-and-cloud-demand-rises. See also https://stand.earth/press-releases/160-data-center-carbon-footprint-increase-research-finds-microsoft-big-oil-forge-7b-partnership-as-several-ai-infrastructure-projects-push-hyperscaler-off-credible-path/

254 Google 2024 Environmental Report; and see https://www.carbonbrief.org/ai-five-charts-that-put-data-centre-energy-use-and-emissions-into-context/

Convergence

MILITARIZATION OF AI IS INEVITABLE — SQ7, p. 41

> *"[T]he greater integration of AI by militaries around the world appears inevitable in areas such as training, logistics, and surveillance."*

The prediction was correct; the list of domains was narrow. Integration extended beyond training, logistics, and surveillance to mass-scale target recommendation, autonomous loitering munitions, drone swarms, and AI-augmented open-source intelligence.

INDUSTRY DOMINANCE IN AI RESEARCH — SQ8, p. 44

> *"U.S. private enterprises have spent over $80 billion on AI, while non-defense investment by the federal government in research and development is estimated at only $1.5 billion in 2020."*

The gap has widened further and become structural. Frontier training runs now cost hundreds of millions to low billions of dollars, and combined AI-infrastructure capital expenditure from Microsoft, Google, Meta, and Amazon reached several hundred billion dollars in 2025 alone. Academic access to frontier compute has contracted further relative to industry.

LOSING MEANINGFUL CONTROL — SQ10, p. 53

> *"As AI systems increase in capability and as they are integrated more fully into societal infrastructure, the implications of losing meaningful control over them become more concerning."*

This single sentence describes what became, after 2022, a dominant international concern: prompting the CAIS statement on extinction risk,[255] the Bletchley Declaration, and the creation of national AI Safety Institutes in the UK, the U.S., Japan, the EU, Canada, Singapore, and Korea. The panel's instinct was right, but the report devotes only a sentence to it and does not treat loss of control as a top-tier risk category. For the 2026 panel, this is the clearest example of something the 2021 panel correctly identified but underemphasized.

DISINFORMATION AS A THREAT TO DEMOCRACY — SQ10, p. 55

> *"AI systems are being used in the service of disinformation on the internet, giving them the potential to become a threat to democracy and a tool for fascism."*

2024 was the largest election year in history. Deepfakes of political figures circulated widely, AI-generated robocalls impersonated a sitting U.S. president in the Democratic primaries, and synthetic political media targeted elections in India, Slovakia, Argentina, and Indonesia, among others. The concern proved correct, although there seems to be no evidence of large-scale AI impacts on elections at present.

255 Center for AI Safety (2023). https://aistatement.com/

NO LARGE AGGREGATE ECONOMIC EFFECTS (YET) — SQ11, p. 56

> *"AI has not been responsible for large aggregate economic effects. But that may be because its impact is still relatively localized to narrow parts of the economy."*

The statement held through 2023. By late 2024 and into 2025, signs of AI's impact had begun to appear in macroeconomic data: U.S. productivity growth of roughly 2.7% in 2025 (nearly double the decade average), the Stanford employment findings for AI-exposed occupations,[256] and the Brynjolfsson, Li, and Raymond results on customer-support productivity.[257] The 2021 reading was correct for its moment and held longer than many commentators expected; the inflection is only now coming into focus.

AI INVESTMENT SMALL IN RELATIVE TERMS — SQ11, p. 58

> *"Global corporate investment in AI was $68 billion in 2020, which is a non-trivial sum but small in relative terms: gross private investment over all categories in the U.S. alone was almost $4 trillion in 2020."*

Correct for 2020. By 2025, AI-related capital expenditure—data centers, chips, models—had become macroeconomically significant: Nvidia's data-center revenue alone passed a $100 billion annualized run rate, and AI-related equity moves accounted for a substantial fraction of U.S. stock-market returns from 2023 to 2025. The relative framing has not held, but the panel was explicit about it being a figure worth watching.

Beyond the Frame

These are developments that the 2021 report does not discuss or only mentions in passing, but which reshaped the field inside four years. They are the gaps that felt most consequential when reading the report cold in April 2026.

THE GENERAL-PURPOSE CHATBOT AS A CONSUMER PRODUCT

No section of the 2021 report anticipates that, within 14 months of publication, a general-purpose LLM chatbot would reach 100 million users and become the fastest-adopted consumer technology on record.[258] ChatGPT's November 2022 launch inverted the expected product path: rather than AI arriving via enterprise integrations and recommender systems, it arrived via a chat window. By 2026, the leading AI assistants have more than a billion weekly users between them. This is arguably the 2021 panel's single largest omission, and the one that most affected the report's readability in the year following its release.

256 Brynjolfsson, Chandar & Chen (2025/2026). https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/

257 Brynjolfsson, Li & Raymond (2025). https://doi.org/10.1093/qje/qjae044

258 UBS/Similarweb estimates, reported Feb. 2023. The fastest consumer-technology adoption on record at the time.

REASONING AND TEST-TIME COMPUTE

The idea that capability could be scaled at inference time—by giving a model a budget to "think" longer—is absent from the 2021 research-direction discussion. Reasoning models (OpenAI's o1 in September 2024, then o3; Anthropic's Claude reasoning modes; Google's Gemini Thinking; DeepSeek-R1; Qwen; and others) have become a dominant research axis. Their effects on mathematics, coding, and scientific reasoning cut across several of the divergences above.

AI SAFETY AND ALIGNMENT AS A DISTINCT FIELD

The 2021 report mentions "loss of meaningful control" once in passing. By 2026, AI safety is developing into a field with national AI safety or security institutes in the UK, the U.S., Japan, the EU, Canada, Singapore, and Korea, dedicated alignment teams inside every frontier lab, explicit frontier-model risk evaluations built into national policy, and a substantial academic literature of its own. This is adjacent to the "losing meaningful control" item in Convergence above, but the absence of a framing of AI safety as a field is a distinct gap.

CONCENTRATION OF FRONTIER DEVELOPMENT IN A SMALL NUMBER OF LABS

The academia-versus-industry framing in SQ8 does not match the reality that emerged: frontier model development is concentrated in fewer than 10 labs globally. OpenAI went from niche nonprofit in 2021 to a private valuation in the hundreds of billions in 2025; Anthropic in 2026 is eyeing a market valuation of over $2 trillion; Nvidia became the world's most valuable company based on AI-compute demand; Google, Meta AI, xAI, Mistral, and a handful of Chinese labs including DeepSeek, Qwen, Z.ai, Moonshot AI, and Tencent make up most of the rest of the frontier. The resulting concentration is a governance and research-access problem that is distinct from the academia/industry axis.

AI AS A FIRST-CLASS TOOL IN PROFESSIONAL SOFTWARE ENGINEERING

GitHub Copilot launched in June 2021, too late for the panel's cutoff; ChatGPT, Claude, and Cursor-style integrated AI coding followed. By 2025, industry surveys place AI-assisted or AI-generated code at roughly 40–46% of professional output for users of these tools.[259] SQ9's opportunities section mentions programming only in passing. Software engineering has since become one of AI's largest and most measurable economic channels, and one of the clearest sites of near-term labor-market adjustment. A challenge remains in determining how best to incorporate AI coding tools and maintain security and correctness.[260]

259 See GitClear, GitHub, and Stack Overflow developer-survey data for 2025.
260 https://cacm.acm.org/news/ai-code-risks-escalate/

COMPUTE GEOPOLITICS AND SEMICONDUCTOR EXPORT CONTROLS

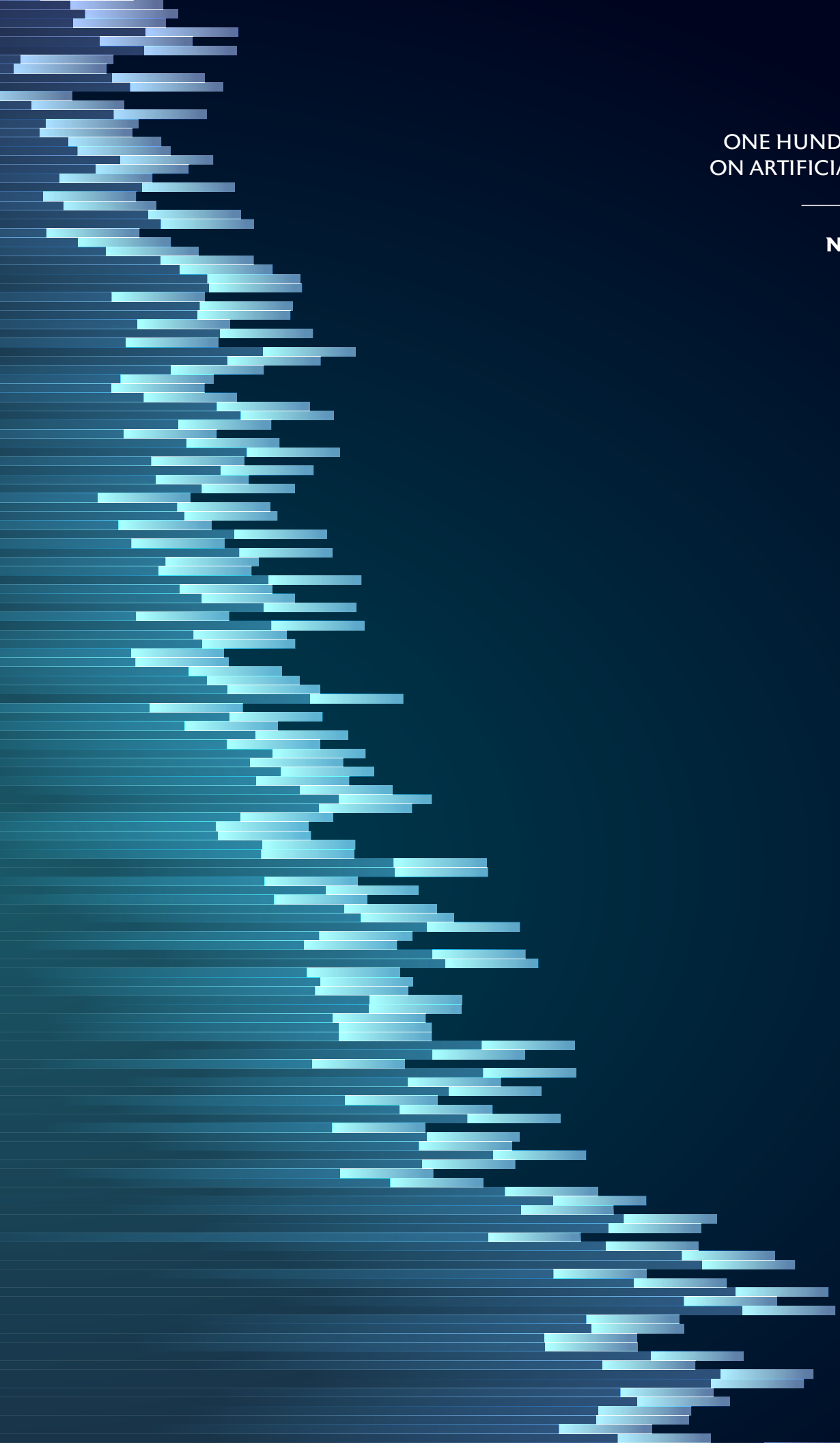
The 2021 discussion of an "AI race" frames competition in terms of investment and talent. It does not anticipate that compute itself would become the instrument of geopolitics: sweeping U.S. advanced-chip export controls on China (October 2022, tightened in 2023, 2024, and 2025),[261] the CHIPS Act, and coordinated allied export-control regimes have reshaped the economics of frontier AI development. This is arguably the most consequential governance surprise of the period.

INDUSTRIAL-SCALE AI TARGETING IN ACTIVE CONFLICT

SQ7 frames autonomous weapons as a category for future debate. The live deployment of AI target-recommendation systems at industrial scale in active conflicts—Israel's Gospel and Lavender in Gaza, Palantir-enabled targeting in Ukraine, autonomous loitering munitions—represents a qualitative escalation that the panel's framing in terms of the debate around autonomous weapons did not anticipate.[262] Despite calling out the inevitability of militarization, the panel did not specifically anticipate the use of AI-generated kill lists and building-target recommendations at scale.

261 U.S. Department of Commerce, Bureau of Industry and Security, advanced-computing and semiconductor-manufacturing export controls on China.

262 Schmitt (2024). https://lieber.westpoint.edu/gospel-lavender-law-armed-conflict/. Human Rights Watch (2024). https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-of-digital-tools-in-gaza



AI100

ONE HUNDRED YEAR STUDY ON ARTIFICIAL INTELLIGENCE

NOVEMBER 2026