AI Fact Sheet for the UN Global Dialogue on AI Governance

2026 is already a pivotal year in AI, and marks an inflection point in the technology's development. The Future of Life Institute (FLI) would like to support delegations to the first UN Global Dialogue on AI Governance by providing essential updates on the latest capability advancements.

The facts below are organized under the four thematic clusters of the Dialogue: Safe, secure and trustworthy AI; AI opportunities and implications; Bridging AI divides; and Respecting, protecting and promoting human rights. We look forward to Member States' leadership in shaping the trajectory of AI towards a better path.

About us: *The Future of Life Institute is the world's oldest and largest AI think tank, with a team of 40+ full-time staff operating globally. FLI has been working to steer the development of transformative technologies towards benefitting life and away from extreme large-scale risks since its founding in 2014.*
Contact: *policy@futureoflife.org*

Safe, secure and trustworthy AI

Performance

Leading AI systems demonstrate Phd-level or higher performance in many technical domains.[1]

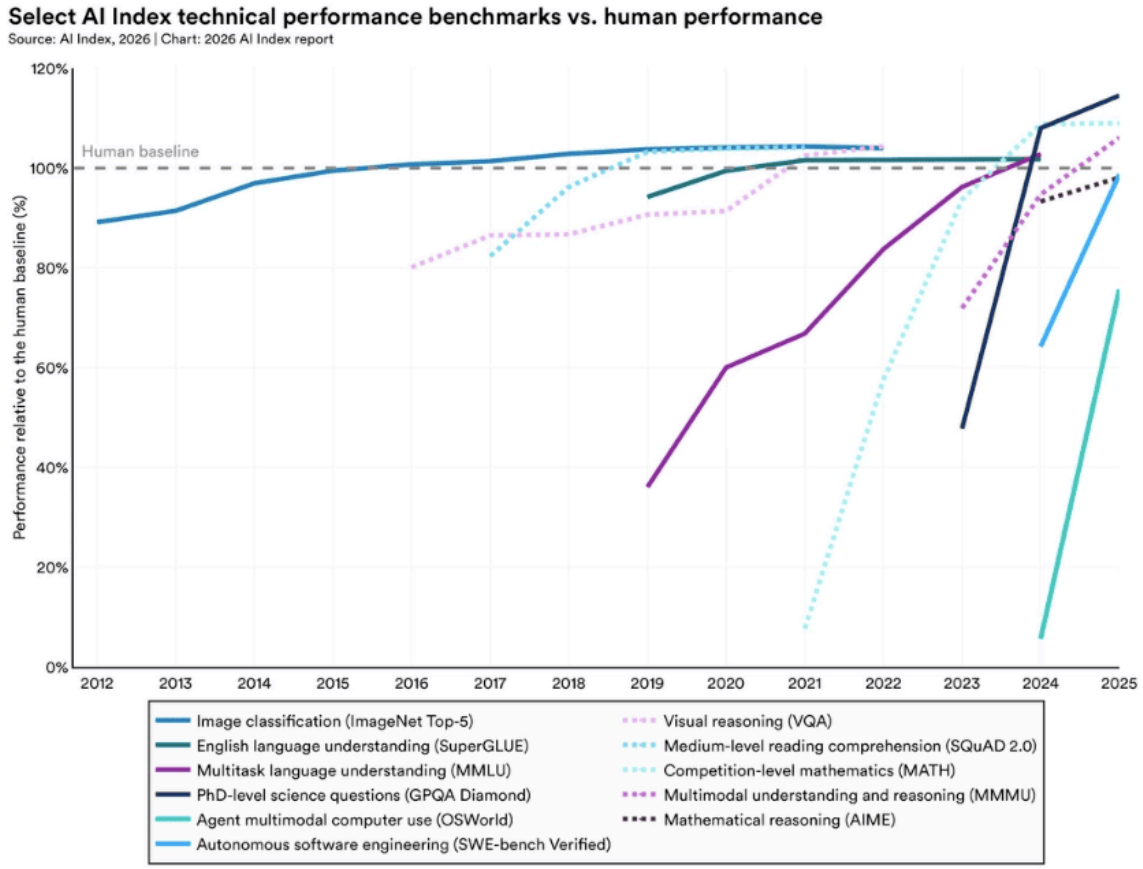

Figure 1. Stanford Institute for Human-Centered Artificial Intelligence. AI Index 2026.

In software engineering, machine learning, or cybersecurity tasks, AI agents can now complete tasks that would take a human programmer almost 17 hours, with a fifty percent success rate. This is up from almost 6 hours in September 2025 and around 10 minutes two years ago.[2] Their ability to complete long, complex tasks is doubling roughly every four months.

Most evaluations of AI capabilities are starting to hit a ceiling. Under this trend, human evaluators will soon no longer be able to design tests hard enough for the most advanced AI systems.[1]

Training compute has grown about 5x per year. If this trend were to continue until 2030, the most advanced AI models could be trained with roughly 3,000 times more compute than those of today.[3]

Using AI to train more advanced AI

Each engineer at Anthropic now produces eight times more code every three months than they produced between 2021 and 2025. More than 80% of that code is authored by Claude, Anthropic's flagship AI model.[4]

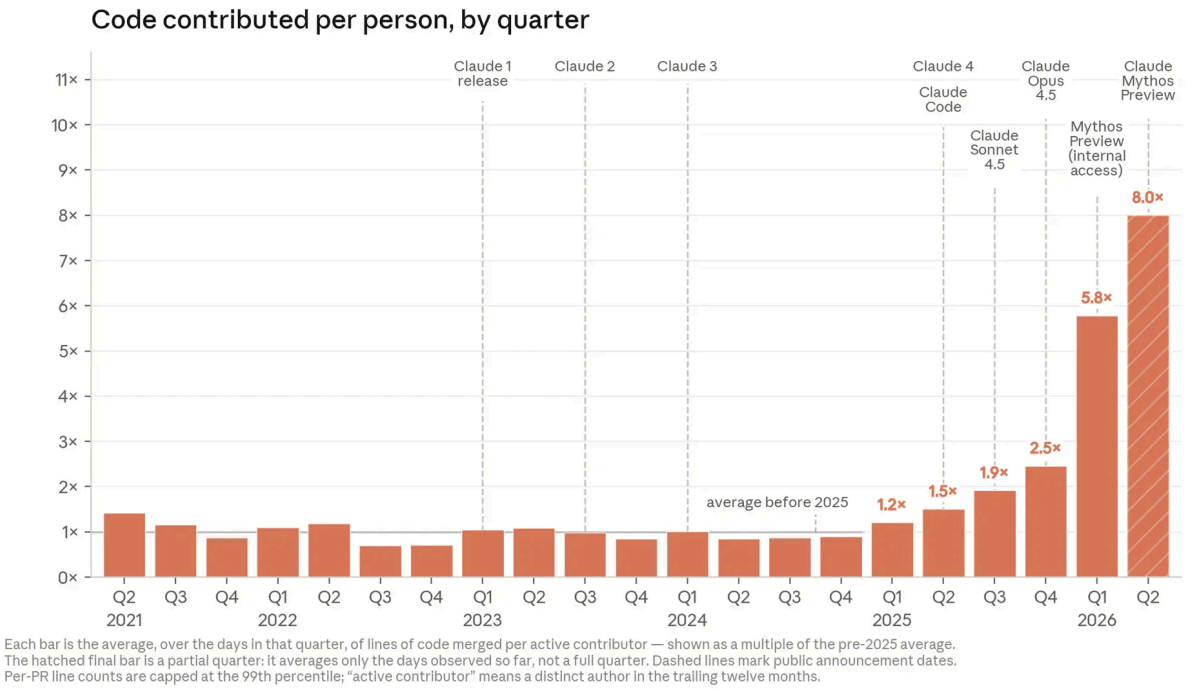

Figure 2. Anthropic. When AI builds itself (2026).

AI risks

There are at least 44 documented cases of AI agents taking steps against the user's intention. Of these, 25 involve deception to hide an overreaching action. In five cases, the AI system took steps that could have fooled the user even on closer review.[5]

In November 2025, Anthropic disclosed the first documented case of a large-scale cyberattack executed without substantial human intervention. Overall, the threat actor was “able to use AI to perform 80-90% of the campaign, with human intervention required only intermittently”.[6]

In April 2026, Anthropic decided to restrict access to its Mythos model because of what it described as a "step-change in vulnerability discovery and exploitation".[7] It later released a version of this model after having applied certain guardrails. On 12 June, the United States government ordered Anthropic to suspend access to foreign nationals after several companies, including Amazon, reported being able to bypass those guardrails. This forced Anthropic to take the model offline.

Only 3.2% percent of 1,200 AI models with biological capabilities have safeguards against their potential use to develop biological weapons.[8]

AI opportunities and implications

Technical opportunities

An AI system can generate a 60-day global weather forecast in under four minutes, running 8 to 60 times faster than prior approaches.[9] In 2025, the Indian government used AI forecasts to predict the start of the monsoon. This forecast helped 38 million farmers improve their crop yield.[10]

Small tailored AI systems are better at protein language modeling and at predicting cellular responses to drugs than large general-purpose AI systems.[11]

Cultural impacts

According to the Global Index on Responsible AI, only 12 countries have a score of 50% or higher in the quality of their measures to protect cultural and linguistic diversity in the use of AI.[12]

AI models can correctly answer 79% of questions about United States culture but not nearly as many for other cultures (e.g., 12% of questions about Ethiopian culture).[13]

Economic/Labour impacts

AI systems are showing performance ranging from 60 to 90% in evaluations related to tax services, mortgage processing, corporate finance and legal reasoning.[1] Multi-agent AI systems have scored 85.5% on complex medical case studies, versus 20% for unaided doctors.[14]

Employment for software developers aged 22-25 has fallen nearly 20% from 2024.[15]

Bridging AI divides

Over 70% of global AI compute is owned by five US companies. Google holds approximately 25% of the world's total compute capacity.[16]

The United States has 10 times more data centers than any other country. The United States and China control approximately 90% of all advanced AI models.[17]

At least 700 million people now use AI weekly. In some countries, over 50% of the population uses AI. However, across much of Africa, Asia and Latin America, rates likely remain below 20%.[18]

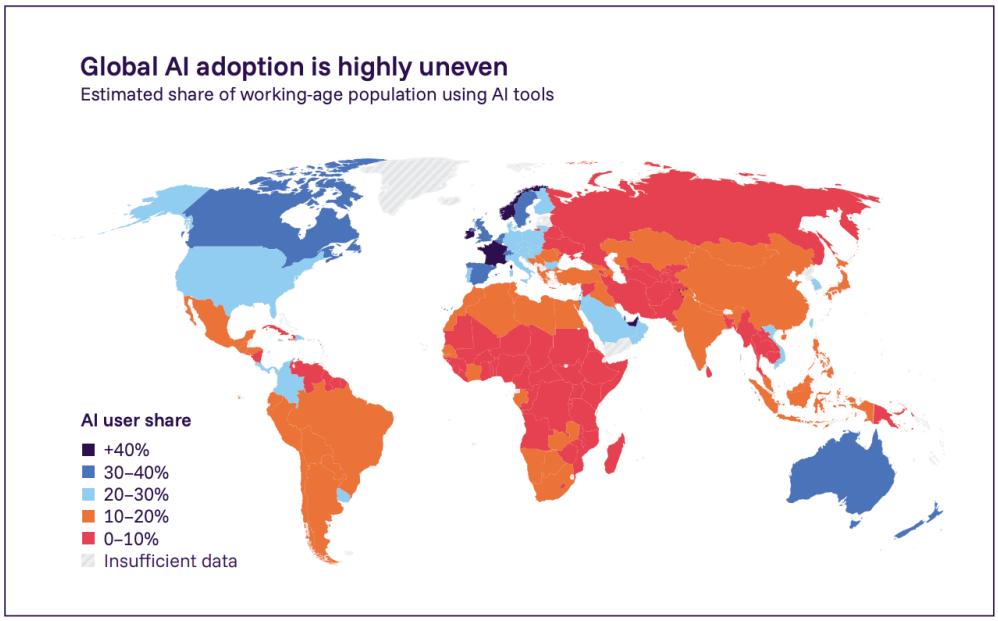

Figure 3. International AI Safety Report 2026

Respecting, protecting and promoting human rights

Rights of women and children

Some AI models generate fake nude images of real people, mostly women, without their permission or knowledge.[19] Across 11 countries, at least 1.2 million children had their images manipulated into sexually explicit deepfakes in the past year.[20]

Information integrity

In one study, participants misidentified AI-generated text as human-written 77% of the time. In another study of audio deepfakes, listeners mistook AI-generated voices for real speakers 80% of the time.[21]

Rights to life and health

Between March and May 2026, AI-related incidents led to 33 fatalities involving 6 platforms. 30% of the victims were minors.[22]

The air pollution expected from AI computing is expected to cause 1300 premature deaths per year in the United States. Associated public health costs are estimated at approximately USD $20 billion per year.[23]

Right to education

In the United States, four out of five high-school and college students use AI for schoolwork. Only half of middle and high schools have AI policies in place.[24]

---

[1] Stanford Human-Centered Artificial Intelligence. "Technical Performance." 2026 AI Index Report, Stanford HAI, 2026, https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance.

[2] METR. "Task-Completion Time Horizons of Frontier AI Models." METR, 8 May 2026, https://metr.org/time-horizons/.

[3] Bengio, Yoshua, et al. International AI Safety Report 2026. UK AI Security Institute / Mila – Quebec AI Institute, Feb. 2026, Section 1.3, https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf.

[4] Anthropic Institute. "When AI Builds Itself." Anthropic, 2026, https://www.anthropic.com/institute/recursive-self-improvement.

[5] METR. "Documented AI Agent Incidents." METR, 19 May 2026, https://metr.org/agent-incidents/.

[6] Anthropic. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." Anthropic, 13 Nov. 2025, https://www.anthropic.com/news/disrupting-AI-espionage.

[7] Anthropic. System Card: Claude Mythos Preview. Anthropic, 7 Apr. 2026, Section 3.1, https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf.

[8] Villalobos, Pablo, and David Atanasov. "Announcing Our Expanded Biology AI Coverage." Epoch AI, 29 Jan. 2025, https://epoch.ai/latest/announcing-expanded-biology-ai-coverage.

[9] Bonev, Boris, et al. "FourCastNet 3: A Geometric Approach to Probabilistic Machine-Learning Weather Forecasting at Scale." arXiv, arXiv:2507.12144, 16 Jul. 2025, https://arxiv.org/abs/2507.12144.

[10] Human-Centered Weather Forecasts. "Forecasting the Onset of the Indian Monsoon." HCWF, University of Chicago, 2025, https://humancenteredforecasts.climate.uchicago.edu/forecasting-the-onset-of-the-indian-monsoon/.

[11] Stanford Human-Centered Artificial Intelligence. "Medicine." 2026 AI Index Report, Stanford HAI, 2026, https://hai.stanford.edu/ai-index/2026-ai-index-report/medicine; Ye, Chengzhong, et al. "Predicting Functional Constraints across Evolutionary Timescales with Phylogeny-Informed Genomic Language Models." bioRxiv, 21 Sept. 2025, doi:10.1101/2025.09.21.677619. https://www.biorxiv.org/content/10.1101/2025.09.21.677619v1; Akiyama, Yo, et al. "Scaling Down Protein Language Modeling with MSA Pairformer." bioRxiv, 2 Aug. 2025, doi:10.1101/2025.08.02.668173. https://www.biorxiv.org/content/10.1101/2025.08.02.668173v1.

[12] Global Center on AI Governance. "Cultural and Linguistic Diversity." Global Index on Responsible AI, 2024, https://www.global-index.ai/thematic-areas-Cultural-and-Linguistic-Diversity.

[13] Myung, Junho, et al. "BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages." arXiv, arXiv:2406.09948, 14 Jun. 2024, https://arxiv.org/abs/2406.09948.

[14] Stanford Human-Centered Artificial Intelligence. "Medicine." 2026 AI Index Report, Stanford HAI, 2026, https://hai.stanford.edu/ai-index/2026-ai-index-report/medicine.

[15] Stanford Human-Centered Artificial Intelligence. "Economy." 2026 AI Index Report, Stanford HAI, 2026, https://hai.stanford.edu/ai-index/2026-ai-index-report/economy.

[16] You, Josh, and Venkat Somala. "Introducing the AI Chip Owners Explorer." Epoch AI, 6 Apr. 2026, https://epoch.ai/latest/introducing-the-ai-chip-owners-explorer.

[17] Stanford Human-Centered Artificial Intelligence. AI Index Report 2026, Chapter 1: Research and Development. Stanford HAI, 2026, https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_1_research_development.pdf.

[18] Microsoft AI Economy Institute. "Global AI Adoption in 2025." Microsoft Corporate Responsibility, 8 Jan. 2026, https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2025/.

[19] AI Forensics. "AI-Generated Image Abuse: Closing the Accountability Gap." AI Forensics Policy Brief, Jan. 2026, https://aiforensics.org/uploads/Grok_Unleashed_Updated.pdf.

[20][20] Unicef. "Artificial Intelligence and Child Sexual Abuse and Exploitation". Issue brief - February 2026, https://www.unicef.org/reports/artificial-intelligence-and-child-sexual-abuse-and-exploitation

[21] Bengio, Yoshua, et al. International AI Safety Report 2026. UK AI Security Institute / Mila – Quebec AI Institute, Feb. 2026, Section 2.1.1, https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf.

[22] AI Companion Mortality Database. AIMortality.org, last updated 10 Jun. 2026, https://aimortality.org/.

[23] Danelski, David. "AI's Deadly Air Pollution Toll." UC Riverside News, University of California, Riverside, 9 Dec. 2024, https://news.ucr.edu/articles/2024/12/09/ais-deadly-air-pollution-toll.

[24] Stanford Human-Centered Artificial Intelligence. "Education." 2026 AI Index Report, Stanford HAI, 2026, https://hai.stanford.edu/ai-index/2026-ai-index-report/education.