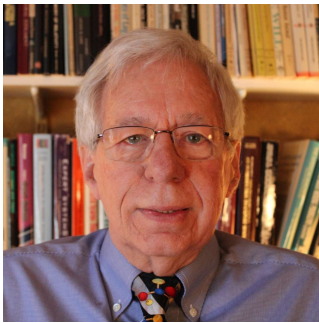


## AI Ethics & Policy Column

### Top AI Ethics and Policy Issues of 2025 and What to Expect in 2026

DOI: 10.1145/3795125.3795132



Larry R. Medsker is Co-Editor-in-Chief of the Springer journal *AI and Ethics* and editor of the forthcoming *AI and Ethics Handbook*. His work bridges AI research and technology policy, informed by leadership roles within the ACM, including as Chair of the ACM U.S. Technology Policy Committee and as Policy Officer for SIGAI. He is a Research Professor at the University of Vermont and Director of the Science & Technology Ethics Policy (STEP) Collaborative and Affiliate Faculty member at George Mason University.

#### Abstract

2025 marked a pivotal shift in AI - from testing to deployment. This happened as generative and agentic systems became essential in key sectors worldwide. This feature highlights the major AI ethics and policy developments of 2025, and concludes with a forward-looking perspective on the ethical and policy challenges likely to shape 2026.

#### Global AI Regulation Enters Its Enforcement Phase

One significant change in 2025 was the enforcement of AI principles and regulations laid out over the past few years. In the US, Trump overturned Biden's 2023 Executive Order on Safe, Secure, and Trustworthy AI, which originally expanded safety requirements for models and increased reporting duties for developers. This is a real shift in attitude towards AI, as the US prioritises deregulation and fast innovation over responsible AI. (Squire Patton Boggs, 2025)

This may lead to friction with the EU, which is taking a more stringent approach. In June, the EU AI Act began requiring organizations to categorize systems by risk level, prepare oversight plans, conduct red-team tests, and publish transparency information (European Commission, 2024). High-risk systems, such as those used in employment, credit, education, and public services, now face strict conformity assessments and ongoing post-market monitoring.

The ACM U.S. Technology Policy Committee (ACM USTPC, 2025) issued several official recommendations advocating for clear documentation practices, rigorous evaluation standards, and effective oversight to ensure public accountability.



Image credits: Pink Office - Jamillah Knowles & Digit / <https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>

## AI Safety, Alignment, and Evaluation Standards Mature

In 2025, AI safety infrastructure experienced rapid growth. Safety - once primarily discussed in conceptual terms - has evolved into a structured engineering discipline. The rise of third-party evaluation centers and independent auditing processes highlights a growing understanding that safety assessments need to go beyond static benchmarks. In this vein, the UK has proposed the AI Growth Lab. This will be a sandbox where new AI models can be tested in real-world conditions, with temporary regulatory modifications to enable effective research. (DSIT, 2025) Benchmarks for assessing deception, persuasion, and long-term planning were widely adopted by leading laboratories, such as OpenAI, Google, Anthropic, Moonshot AI, and AliBaba (Stanford CRFM, 2025).

The ACM USTPC emphasized explainability as essential for fairness, arguing that black-box systems undermine both scientific integrity and democratic oversight. Their guidance influenced policy discussions across healthcare, finance, and critical infrastructure, where transparency became a necessary condition for deployment. These developments reflect themes from the upcoming AI and Ethics Handbook (Medsker 2026), which argues that safety must consider the socio-technical context, not just model-level testing.

## Backlash

One of the most notable changes in 2025 is the growing recognition that refusing to deploy GenAI can be ethically justified. Ethical deployment is now seen as relying not only on regulations but also on essential AI literacy: understanding system limits, social context, and human judgment. This perspective places the primary responsibility on institutions, not

individual users, to establish clear governance, provide proper oversight, and determine when AI should not be used at all. This view challenges ideas of technological inevitability and highlights the importance of human judgment in AI-driven systems. (Duarte et al., 2026).

### Agentic AI and Autonomous Decision Systems

Agentic AI systems achieved significant advances in 2025. For example, Cognitive Automation Agent is a new platform that independently manages clinical workflows in the US. This is one of the first of its kind, as it can initiate actions independently of clinicians. (Epic, 2025)

The increasing capabilities of agentic AI raise critical questions about oversight, predictability, and moral responsibility. Its deployment means that AI ethics evaluations will focus on the actions of such systems, rather than their predictions. To address these concerns, the USTPC emphasized the importance of clear responsibility, strong monitoring systems, and transparent governance structures (ACM USTPC, 2025).

### Vibe Coding

A look back at 2025 wouldn't be complete without discussing the new practice of "vibe coding", where developers generate, refine, and debug code through iterative interaction with LLMs. While often framed as a productivity boost, this approach effectively delegates significant design and implementation decisions to AI agents, raising familiar questions about accountability, security, and oversight in automated decision-making. These unanswered questions haven't halted its popularity, and it has become so ubiquitous that Collins Dictionary named it Word of the Year for 2025.

### Misinformation, Deepfakes, and Digital Trust

The global election cycles of 2024–2025 have created significant pressure on information ecosystems. Highly convincing deepfakes, synthetic news, fraudulent political ads, and automated persuasion tools are spreading widely, with AI-based scams on the rise.

For example, in July, an unknown actor used Marco Rubio's voice and writing style to contact 5 senior US government officials. (Gedeon, 2025). And AI impersonations of pop stars have reportedly scammed fans out of \$5.3bn for concert tickets and VIP experiences that do not exist. (Dilts Marshall, 2025). These incidents have sparked widespread calls for regulation and mark a decline in digital trust.

Governments, platforms, and researchers have adopted provenance metadata, watermarking, and digital signature technologies to verify content (C2PA, 2024). However, the effectiveness of these protections varies by platform and situation, as watermarks can be altered. The World Economic Forum is calling for robust security protocols to combat deepfakes. (Colman, 2025).

## Children and AI

A significant digital trust challenge emerging around 2025 involves the intersection of children's rights and democratic integrity, as generative AI chatbots increasingly mediate social, educational, and political information for young users. The USTPC warned about the manipulative potential of chatbots interacting with minors and highlighted broader democratic risks posed by generative misinformation, given that chatbots can facilitate manipulative tactics and misinformation at scale (ACM USTPC 2025). UNICEF shares this concern, emphasizing that AI systems impacting children must be designed with explicit safeguards, transparency, and accountability to prevent exploitation and undue influence. (ACM USTPC, 2025; UNICE, 2024).

## Data Rights and Training Data Governance

In 2025, debates over the ethics and legality of AI training data grew more intense. We saw many more lawsuits related to large-scale web scraping, unauthorized use of copyrighted materials, and biometric data collection. In June, Reddit and the BBC took legal action against Perplexity AI. (McMahon, 2025). (Reuters, 2025). Several governments moved to require greater transparency about the sources, composition, and legal basis of training datasets. Courts and regulators made uneven but necessary progress on whether training generative AI models on copyrighted works qualifies as fair use. Late-year disputes between major publishers and AI developers highlighted unresolved questions about who is entitled to compensation. (Nawotka, E. 2025).

Depending on the outcome of these cases, we may see the only legal generative AI systems in the U.S. being those trained on public-domain works or under licenses. (Samuelson, 2024). Meanwhile, the EU and the UK have moved toward obligations for developers to document training data sources and justify the inclusion of copyrighted or sensitive material (UK IPO, 2024), with these requirements expected to expand significantly during 2026–2027.

## Systemic Risk and Critical Infrastructure

AI has become more deeply integrated into critical infrastructure, but the systems in place are fragile. An AWS service disruption acted as a reminder of this. On October 20th, millions of services that people rely on were shut down, from payment platforms like PayPal to consumer technology like smart beds (Newman, 2025). This highlighted the fragility of a system controlled by only a few actors. (Gkritsi & Haeck, 2025)

## Workforce Transformation

Copilots and agentic assistants altered workflows in administrative work, accounting, journalism, law, and STEM research. Productivity increased rapidly but unevenly. Some industries experienced major gains, while others faced risks of displacement or widening skill

gaps. For example, in March, JP Morgan rolled out its own coding assistant for its software engineers, increasing their productivity by 10-20%. (Reuters 2025) Meanwhile, the jobs market remains uncertain, with workers scrambling to remain employable as they now need to work with, or compete with, AI amid over 200K layoffs in the global tech sector (Mzekandaba, 2026). The USTPC's recommendations on AI literacy, reskilling, and fair workforce development emphasize that increased national investment in education and training is vital for reducing disparities (ACM USTPC, 2025).

### Environmental Impact and the Ethics of Scale

Google released a report on the energy use of its Gemini model, estimating that a single prompt emits 0.03g of carbon dioxide, consumes 0.26 ml (5 drops) of water, and has an energy impact equivalent to watching TV for 9 seconds. (Vahdat & Dean, 2025). These numbers scale up fast, given that Gemini powered the 5.9 trillion Google searches in 2025. (Kumar, 2025) However, this is supposedly a 44-fold decrease in carbon footprint per prompt, due to efficiency increases at data centers.

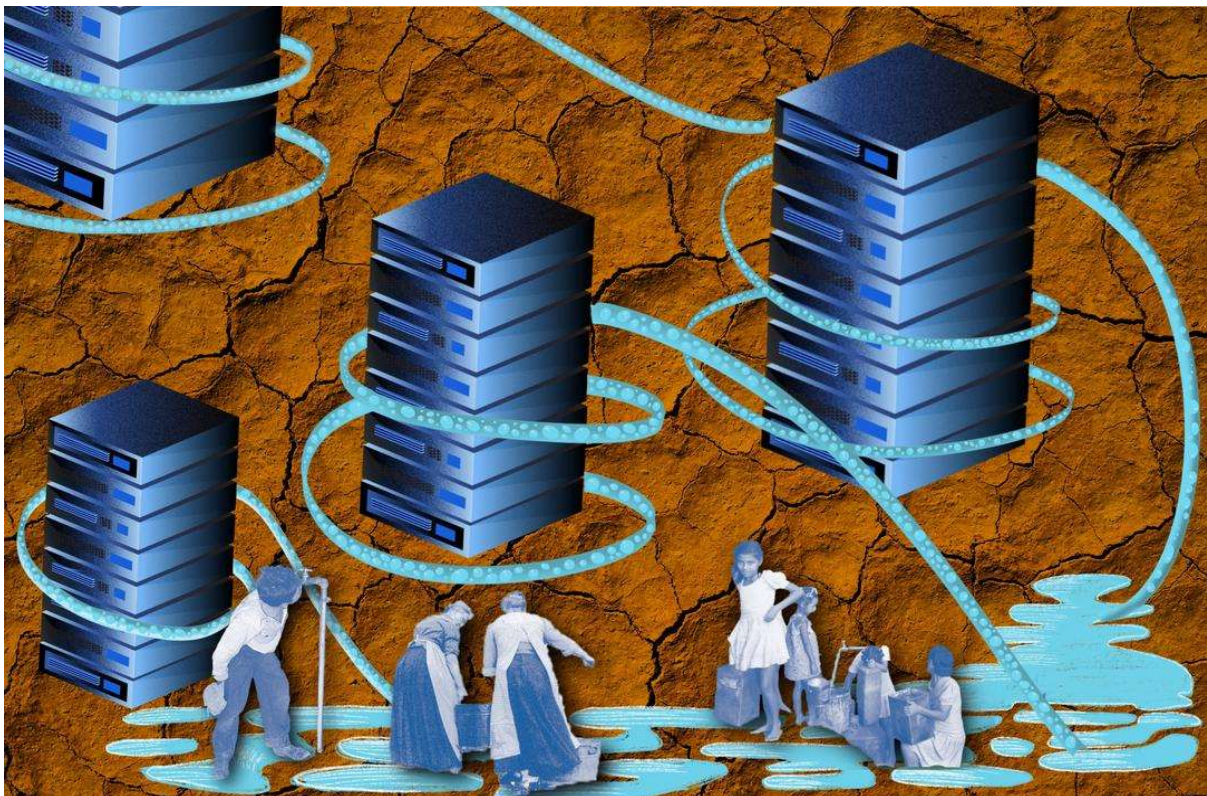


Image Credits: The Environmental Impact of Data Centers in Vulnerable Ecosystems - Gloria Mendoza / <https://betterimagesofai.org> / <https://creativecommons.org/licenses/by/4.0/>

Data centers currently account for 4.4% of the USA's energy demand, but projections estimate that this will triple by 2028. (O'Donnell & Crownhart, 2025). Other studies documented significant energy and water use associated with training and running large-scale models (Morrison et al., 2025), and policymakers and industry leaders responded by calling for sustainable AI practices, increased efficiency incentives, and broader environmental disclosure requirements (Conference Board, 2025).

### Bias, Fairness, and Non-Discrimination

Despite advances in dataset auditing and fairness tools, algorithmic discrimination remained a significant challenge in 2025. Models used for hiring, credit, public benefits, and education continue to mirror historical inequalities (Buolamwini and Gebru, 2018; Barocas, Hardt, and Narayanan, 2019). The USTPC reaffirmed its long-standing stance, calling for a pause on facial recognition deployments in high-risk settings where civil rights impacts are foreseeable (ACM USTPC, 2025). Policymakers are increasingly adopting risk-based frameworks that combine socio-technical analysis with technical audits.

### Looking Ahead to 2026

If 2025 marked the year AI regulation became operational, 2026 could be when autonomy, sovereignty, and sustainability take center stage. Three pressures are likely to dominate: governing increasingly automated systems, managing economic and workforce disruption, and confronting the environmental and infrastructural limits of large-scale AI. As AI becomes deeply integrated into economic and social systems, collaboration among governments, research institutions, civil society, and organizations such as ACM USTPC and SIGAI will be essential to developing trustworthy and fair systems.

Economic concerns are increasingly prominent in public discourse and policy-making, with widespread changes taking place as AI displaces many white-collar tasks. Growing questions about whether expectations for AI have peaked or whether parts of the sector are experiencing a speculative bubble have raised the possibility of a market correction with broader economic consequences (Georgieva, 2025; International Monetary Fund, 2025). Such a shift could affect not only the AI technology sector itself, but also investment patterns, labor markets, and innovation strategies across the wider economy.

The ethics landscape is also evolving. By 2026, "AI-generated" labels may give way to verifiable provenance signals that can be shared across platforms (Hancock and Bailenson 2024). Regulators could strengthen enforceable duties around verification, rapid takedown, and auditing - especially as deepfakes increasingly affect high-stakes domains such as health, finance, and education. Agentic systems, automated decision-making tools, and practices such as AI-assisted "vibe coding" are likely to become more prominent as organizations move quickly from pilot projects to real-world deployment. Data from this rapid adoption of agentic AI and automated development workflows may spur debates on accountability, security, and testing practices, particularly in high-risk environments where iterative human oversight is vital

but inconsistently applied. International action on likeness and consent rights may accelerate, making unauthorized use of synthetic identities a distinct civil offense.

We might see 2026 be the year when restraint becomes a strategic necessity rather than just a moral choice. Over the past year, AI capabilities have advanced faster than institutions, labor markets, and governance frameworks could adapt, exposing a growing gap between reassuring narratives and real-world results. While many providers emphasized responsibility through process and compliance, only a few leaders publicly acknowledged systemic risks, from workforce disruption to misuse at scale, and called for explicit limits on deployment. Among the most notable was Dario Amodei, whose call for clear rules and responsible scaling framed ethics as a core engineering constraint rather than an after-the-fact safeguard (Amodei 2024). His contribution has shaped subsequent discourse.

Major questions include when and where generative AI should be used, who benefits from its use, and under which conditions it should be restricted. A widely articulated perspective emphasizes that GenAI should assist, not replace, human judgment, with accountability firmly placed on institutions rather than automated systems. Concerns remain that generative AI can reinforce biases, exacerbate epistemic injustice, and centralize power, underscoring the need for transparency, contestability, and data rights. Ongoing debates about training data reveal unresolved issues related to consent, labor, and compensation. Furthermore, environmental impacts and dual-use risks such as misinformation, surveillance, and security threats have become key ethical challenges. (Radanliev, 2025; Janssen, 2025).

Societal responses to these questions are becoming more visible. Communities may express a “digital backlash” against algorithmic technologies, as seen in protests over data center projects, student-led petitions, app deletions, industry open letters, and academic position papers (Guest et al. 2025). Educators, technologists, policymakers, artists, labor unions, and community groups are increasingly opposing AI systems perceived as harmful, exploitative, flawed in education, environmentally damaging, or socially unjust. Supporting resistance, refusal, reclamation, and reimagining AI remains an essential ethical goal, even as some “responsible” AI narratives suggest opposition is futile (Duarte et al., 2025).

The coming year will test whether global AI governance can keep pace with innovation while protecting democratic values, social trust, and human well-being. Ultimately, 2026 should reveal whether adherence to emerging frontier and general-purpose AI standards effectively influences real-world behavior or if it merely becomes a box-checking exercise (Clark, 2024). We will also observe whether trust frameworks are unified globally or divided into regional systems with incompatible rules and technologies. Addressing these challenges will require active participation from the AI community through groups like ACM USTPC and SIGAI, as well as engagement from the broader research ecosystem in venues such as *AI and Ethics* and *AI Matters*. Together, these efforts are essential for guiding responsible innovation and ensuring AI advancements benefit society.

*AI policy is a dynamic, ever-evolving landscape. We welcome your ideas and perspectives - email us at [aimatters@sigai.acm](mailto:aimatters@sigai.acm) to discuss topics for the AI Matters Ethics & Policy column.*

*Special thanks to Ella Scallan for editing this piece.*

## References

ACM USTPC (Association for Computing Machinery US Technology Policy Committee). 2025. "USTPC Policy Products." Accessed December 6, 2025. <https://www.acm.org/public-policy/public-policy-statements>

Amodei, Dario. 2024. "AI Is Getting More Powerful. We Need Clear Rules." Time, April 2024.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning: Limitations and Opportunities. Online book. Retrieved from <https://fairmlbook.org/>

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81: 77–91.

C2PA. 2024. Content Provenance and Authenticity Standard. Coalition for Content Provenance and Authenticity. <https://spec.c2pa.org/specifications/specifications/2.3/index.html>

CAC (Cyberspace Administration of China) et al. 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. [http://www.cac.gov.cn/2023-07/13/c\\_1690898327029107.htm](http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm)

CISA. (Cybersecurity & Infrastructure Security Agency). 2024. "Widespread IT Outage Due to CrowdStrike Update" (Alert AA24-201A), last modified August 6, 2024, accessed December 6, 2025, [www.cisa.gov](http://www.cisa.gov).

Clark, Jack, Gillian Hadfield, Percy Liang, and others. 2024. "Frontier AI Regulation: Managing Emerging Capabilities and Systemic Risks." Center for AI Safety White Paper. Conference Board. 2025.

Colman, B. (2025) *Why detecting dangerous AI is key to keeping trust alive in the deepfake era*. World Economic Forum, 7 July. Accessed January 21, 2026. Available at: <https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/>

The Conference Board. AI and Sustainability 2025: Corporate Perspectives on Environmental Impact. <https://www.conference-board.org/press/AI-and-sustainability-2025>

Dilts Marshall, E. (2025) *Taylor Swift, Sabrina Carpenter impersonators scam fans out of \$5.3 billion in 2025: Report*. Billboard Pro. Accessed January 21, 2026. Available at: <https://www.billboard.com/pro/taylor-swift-sabrina-carpenter-impersonators-scam-fans-2025/>

DSIT and Kendall, L. (2025) *New blueprint for AI regulation could speed up planning approvals, slash NHS waiting times, and drive growth and public trust*. Press release, 21 October. Available at: <https://www.gov.uk/government/news/new-blueprint-for-ai-regulation-could-speed-up-planning-approvals-slash-nhs-waiting-times-and-drive-growth-and-public-trust>

Duarte, Tania, Ismael Kherroubi Garcia, Anshur, et al. 2025. Resisting, Refusing, Reclaiming, Reimagining: Charting Challenges to Narratives of AI Inevitability” We and AI, DOI: 10.5281/zenodo.17343830. <https://weandai.org/resisting-refusing-reclaiming-reimagining-ai/>

Duarte, Tania, Kathryn Conrad, Ismael Kherroubi Garcia. 2026. “The Emergence of Critical AI Literacy.” Forthcoming. AI and Ethics Handbook. Cham, Switzerland: Springer Nature. <https://www.amazon.de/dp/3032122007>

Epic Systems Corp. 2025. “Epic Building Out Agentic AI, Broadens Focus Beyond EHRs.” Epic News, March 7, 2025. <https://www.epic.com/epic/post/epic-building-out-agentic-ai-broadens-focus-beyond-ehrs/>

European Commission. 2024, June 13. Artificial Intelligence Act: Regulatory Framework for Trustworthy AI (Regulation (EU) 2024/1689). Official Journal of the European Union L, 2024, 1689. <eur-lex.europa.eu>

Georgieva, Kristalina. 2025. “IMF and Bank of England Warn AI-Led Market Bubble Could Burst.” Financial Times, October 8, 2025. <https://www.ft.com/content/fe474cff-564c-41d2-aaf7-313636a83e5b>

Gedeon, J. (2025) *AI scammer posing as Marco Rubio targets officials in growing threat. The Guardian*, 8 July. Accessed January 21 2026. Available at: <https://www.theguardian.com/us-news/2025/jul/08/marco-rubio-ai-impostor>

Guest, Olivia et al. 2025. “Against the Uncritical Adoption of 'AI' Technologies in Academia.” DOI [10.5281/zenodo.17065098](https://doi.org/10.5281/zenodo.17065098).

Hancock, Jeffrey T., and Jeremy N. Bailenson. 2024. “The Future of Digital Trust: Deepfakes, Authenticity, and the Crisis of Verifiable Media.” *Journal of Online Trust and Safety* 3 (1): 1–22.

International Monetary Fund. 2025. *World Economic Outlook: Navigating Global Divergence*. Washington, DC: International Monetary Fund. <https://www.imf.org/en/Publications/WEO>

Janssen, M. 2025. “Responsible Governance of Generative AI: Conceptualizing AI as a Complex Adaptive System.” *Policy & Society* 44 (1): 38–5

Kumar, N. 2025. How many Google searches per day [New 2026 Data]. DemandSage, 8 December. Accessed January 19th 2026. Available at: <https://www.demandsage.com/google-search-statistics/>

MacIntyre, John, and Larry R. Medsker, eds. 2020–. *AI and Ethics* (journal), Springer Nature. <https://link.springer.com/journal/43681>

McMahon, L. 2025. *BBC threatens AI firm with legal action over unauthorised content use*. BBC News. Accessed January 21 2026. Available from: <https://www.bbc.co.uk/news/articles/cy7ndgylzzmo>

Medsker, Larry R., ed. Forthcoming. AI and Ethics Handbook. Cham, Switzerland: Springer Nature. <https://www.amazon.de/dp/3032122007>

METI (Ministry of Economy, Trade and Industry, Japan). 2023. G7 Hiroshima Process: Guiding Principles for Advanced AI.

Morrison, Jacob, Clara Na, Jared Fernandez, Tim Dettmers, Emma Strubell. 2025. "Holistically Evaluating the Environmental Impact of Creating Language Models." arXiv, March 3, 2025. <https://arxiv.org/abs/2503.05804>

Mzekandaba, S. (2026, January 9). 2025 global tech sector layoffs surpass 200k. ITWeb. <https://www.itweb.co.za/article/2025-global-tech-sector-layoffs-surpass-200k/G98YdqLGPb4MX2PD>

Newman, Lily Hay. 2025. "The Long Tail of the AWS Outage." WIRED, October 22, 2025. <https://www.wired.com/story/aws-cloud-outage-long-tail>

NIST (National Institute of Standards and Technology). 2023. AI Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

O'Donnell, J. and Crownhart, C. 2025. We did the math on AI's energy footprint. Here's the story you haven't heard. MIT Technology Review, 20 May. Available at: <https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech>

OECD. 2023. OECD Framework for the Classification of AI Systems. Organisation for Economic Co-operation and Development.

Patterson, David, et al. 2023. "The Carbon Footprint of Large AI Models." Communications of the ACM 66(10): 58–71.

Nawotka, E. 2025. New lawsuit against AI companies seeking more money. Publishers Weekly. Accessed January 21, 2026. Available at: <https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/99347-authors-file-new-lawsuit-against-ai-companies-seeking-more-money.html>

Gkritsi, E. and Haeck, P. 2025. Amazon cloud outage fuels call for Europe to limit reliance on US tech. Politico. Accessed January 18, 2026. Available at: <https://www.politico.eu/article/aws-amazon-web-services-outage-europe-limit-reliance-us-tech/>

Radanliev, P., et al. 2025. "AI Ethics: Integrating Transparency, Fairness, and Privacy." Journal of Information, Communication and Ethics in Society 23 (2): 1–18.

Reuters. 2025. *JPMorgan Engineers' Efficiency Jumps as Much as 20% from Using Coding Assistant*. Reuters, March 13, 2025. <https://www.reuters.com/technology/artificial->

[intelligence/jpmorgan-engineers-efficiency-jumps-much-20-using-coding-assistant-2025-03-13/](#).

Reuters. 2025. *Reddit sues Perplexity for scraping data to train AI system*. 22 October. Available at: <https://www.reuters.com/world/reddit-sues-perplexity-scraping-data-train-ai-system-2025-10-22/>

Samuelson, Pamela. 2024. "Generative AI Meets Copyright Law." *Communications of the ACM* 67 (11): 20–23.

Squire Patton Boggs (2025) Key insights on President Trump's new AI Executive Order and policy & regulatory implications. February. Accessed January 21, 2026. <https://www.squirepattonboggs.com/insights/publications/key-insights-on-president-trumps-new-ai-executive-order-and-policy-regulatory-implications/>

Stanford CRFM. 2025. Stanford Center for Research on Foundation Models. (2025). Holistic Evaluation of Language Models (HELM) (Version 2.0). [Stanford University](#). Retrieved December 6, 2025, from <https://crfm.stanford.edu/helm/>

UK Intellectual Property Office. 2024. *Guidance on Copyright and AI Model Training*.  
UNICEF. 2024. *Policy Guidance on AI for Children*. Florence: UNICEF Office of Research – Innocenti. Accessed 2024.  
<https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>

Vahdat, A. and Dean, J. 2025. *How much energy does Google's AI use? We did the math*. Google Cloud Blog, 21 August. Accessed January 18, 2026. Available at: <https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference>

White House. 2024. *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*. <https://www.whitehouse.gov/>.

Wikipedia contributors. 2024. "Delta Air Lines v. CrowdStrike." Wikipedia, July 19, 2024. Accessed December 16, 2025. [https://en.wikipedia.org/wiki/Delta\\_Air\\_Lines\\_v.\\_Crowdstrike](https://en.wikipedia.org/wiki/Delta_Air_Lines_v._Crowdstrike)

Williams, Hannah Murphy. 2024. "JPMorgan Tests AI Code-Writing Agents to Automate Software Development Tasks." *Financial Times*, February 20, 2024.

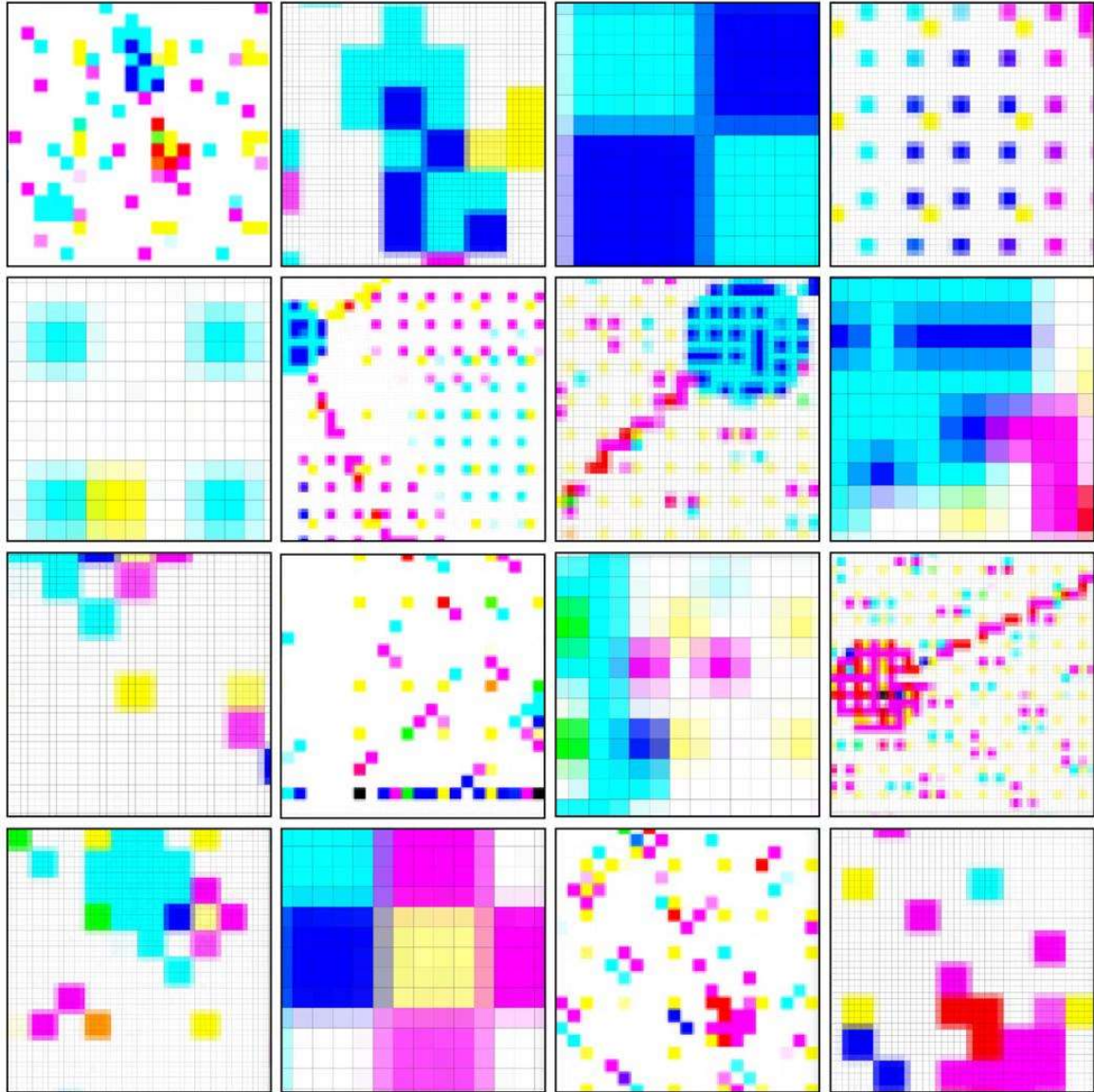


Image credits: Game of Pixels x Toy Models - Elise Racine / <https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>