Spotlight on AI and Society: Power, Bias, and Behavior

Authors: Rohita Biswas, Cinthya Souza Simas, Sara Tóth Martínez, Gerhard G. Steinmann, Roland Mertelsmann, María Belén Moyano

Submitted: 2. March 2026
Published: 4. March 2026
Volume: 13
Issue: 2

Affiliation: Journal of Science, Humanities and Arts (JOSHA), Freiburg im Breisgau, Germany

Languages: English
Keywords: AI Agents; AI Governance; Data Centres; Deepfake Abuse; Subliminal Learning.
Categories: News and Views
DOI: 10.17160/josha.13.2.1129

Abstract:

This Spotlight follows AI as it shifts from a helpful assistant to a system that can produce, judge, and propagate ideas at scale and shows what that change is doing to science, society, and safety. It opens with an “agent-run” conference where AI systems generate papers and other AI systems review them, turning peer review into a real-world stress test for multi-agent research and its blind spots. From there, it tracks how chatbots have become political flashpoints, with competing demands for “neutral” AI colliding with messy technical realities and free-speech concerns. The collection then zooms out to the material backbone of the boom, data centers, electricity, connectivity, and the growing compute divide between regions trying to compete and those being priced out. Finally, it lands in everyday life: why generative tools amplify the risks of posting children’s photos, and why even “clean” model-to-model training can quietly

JOSHA
josha.org

Journal of Science, Humanities and Arts

JOSHA is a service that helps scholars, researchers, and students discover, use, and build upon a wide range of content

JOSHA Journal of Science, Humanities and Arts

March 2026 Volume 13, Issue 2

Spotlight on AI and Society: Power, Bias, and Behavior

Rohita Biswas, Cinthya Souza Simas, Sara Tóth Martínez, Gerhard G. Steinmann, Roland Mertelsmann, María Belén Moyano
editorial@josha-archive.org
Journal of Science, Humanities and Arts, Freiburg im Breisgau, Germany

Abstract

This Spotlight follows AI as it shifts from a helpful assistant to a system that can produce, judge, and propagate ideas at scale and shows what that change is doing to science, society, and safety. It opens with an “agent-run” conference where AI systems generate papers and other AI systems review them, turning peer review into a real-world stress test for multi-agent research and its blind spots. From there, it tracks how chatbots have become political flashpoints, with competing demands for “neutral” AI colliding with messy technical realities and free-speech concerns. The collection then zooms out to the material backbone of the boom, data centers, electricity, connectivity, and the growing compute divide between regions trying to compete and those being priced out. Finally, it lands in everyday life: why generative tools amplify the risks of posting children’s photos, and why even “clean” model-to-model training can quietly transmit hidden behavioral traits, complicating alignment and governance.

Keywords: AI Agents; AI Governance; Data Centres; Deepfake Abuse; Subliminal Learning.

1. AI bots wrote and reviewed all papers at this conference

By Elizabeth Gibney

Agents4Science 2025 is an online computer-science conference designed as a controlled testbed in which both the submitted papers and the peer reviews are produced primarily by AI "agents" (coordinated groups of models), while human researchers attend, run experiments, and may present alongside the agents. The organizers frame the event as a way to probe a recent shift from using single-purpose AI tools toward deploying multi-agent systems that can perform end-to-end research tasks, and to generate evidence about where such systems succeed or fail (including tendencies toward errors and "false positive" discoveries). Submissions from more than 300 AI agents were screened by AI reviewers, with 48 papers accepted; the work is largely computational and spans diverse domains. To enable analysis of human influence on outcomes, submissions must document the human–agent interaction at each stage, with the goal of informing future evaluation practices and research-use policies.

This article was previously published in *Nature News* on October 14, 2025.

Read the full article here

2. The chatbot culture wars are here

By Kevin Roose

The article explains how a new political conflict has emerged in the United States over AI chatbots. Conservatives, including President Trump and Republican officials, accuse major AI companies of embedding left-wing or "woke" ideologies into their systems. They cite examples such as chatbots refusing to praise Trump, producing historically inaccurate content, or prioritizing diversity, equity, and inclusion. In response, Republicans have launched investigations, issued subpoenas, and, under the Trump administration, introduced an executive order requiring that AI systems used by federal agencies be "objective" and free from ideological bias. The administration is using the threat of losing federal contracts to pressure AI companies to change their systems, a strategy similar to past Republican efforts against social media platforms. Roose argues that this approach raises serious constitutional concerns, since using government funding to pressure private companies over their speech may violate the First Amendment. He also

notes that defining and enforcing “neutral” or “unbiased” AI is technically and conceptually difficult, because chatbot outputs vary by user, model, and context, and AI systems cannot be easily controlled through simple instructions.

This article was previously published in *The New York Times* on July 23, 2025.

Read the full article here

3. Why AI should make parents rethink posting photos of their children online

By Brian X Chen

The article explains how generative-AI tools have amplified the privacy and safety risks of “sharenting,” focusing on so-called “nudifier” services that can fabricate realistic nonconsensual nude images from an ordinary photo. It describes how these tools lower the cost, skill, and time needed to create abusive deepfakes, shifting risk from public figures to everyday people, including children, and notes that such apps are reportedly circulating among students in schools. Using examples of typical pricing models, discussion of traffic-based revenue estimates, and references to investigations of numerous nudifier websites, the piece argues that the threat persists even as platforms restrict advertising and new laws target distribution of nonconsensual imagery because use of the tools themselves remains easy and often hard to police. Beyond deepfakes, it highlights secondary harms from posting children’s photos, such as revealing identifying details that can enable identity theft, and it contrasts public posting with lower-risk sharing options (private albums and encrypted messaging) while acknowledging that total control is difficult in practice.

This article was previously published in the *New York Times* on August 11, 2025.

Read the full article here

4. Subliminal learning: language models transmit behavioral traits via hidden signals in data

By Alex Cloud *et al*

Subliminal learning occurs when a language model acquires behavioral traits from training data that is not semantically related to those traits, such as sequences of

numbers, source code, or heavily filtered reasoning traces. When a model is fine-tuned on outputs generated by another model with a specific preference or degree of misalignment, it tends to adopt the same trait, even when the training data contains no explicit references to it. This effect appears across different traits and data types, but largely depends on the teacher and student models sharing a similar initialization, suggesting that the transmission arises from model-specific internal patterns rather than hidden meanings in the data. Attempts to detect these traits through automated classifiers, human inspection, or in-context learning consistently fail, reinforcing the idea that the signals are not semantically interpretable. The findings are further supported by theoretical analysis showing that distillation naturally pulls a student model toward the teacher's parameters. Overall, these results indicate that training on model-generated data can unintentionally transfer latent behaviors, including misalignment, raising important concerns for model alignment, and safety.

This article was previously published in *arXiv preprint arXiv:2507.14805* of Cornell University, on July 20, 2025.

Read the full article here

5. The steep costs of AI

By Adam Satariano *et al*

Friendly government policies towards AI data centers come at a cost. Not having AI also does. From Mexico having an entire town suffering power outages, to Argentina experiencing a loss of top students to countries with more powerful computers, or even in Kenya with coders waiting for America to fall asleep to be able to use faster internet speeds, the AI race is leaving parts of the world behind. Some countries are trying to insert themselves, such as Ireland and Chile, but everything comes at a steep cost. This article summarizes the current consequences of AI and what future perspectives on politics and the global economy could become if we do not think about all the implications of powering AI.

This article was previously published in *The New York Times* on October 23, 2025.

Read the full article here

Acknowledgements

GPT-5.2 version of Chat-GPT (Open AI) and Gemini (Google) was used during the writing process as part of JOSHA's policy of experimentation with AI tools. However, JOSHA takes full responsibility for its content.