

REBUILDING TRUST IN GLOBAL GOVERNANCE



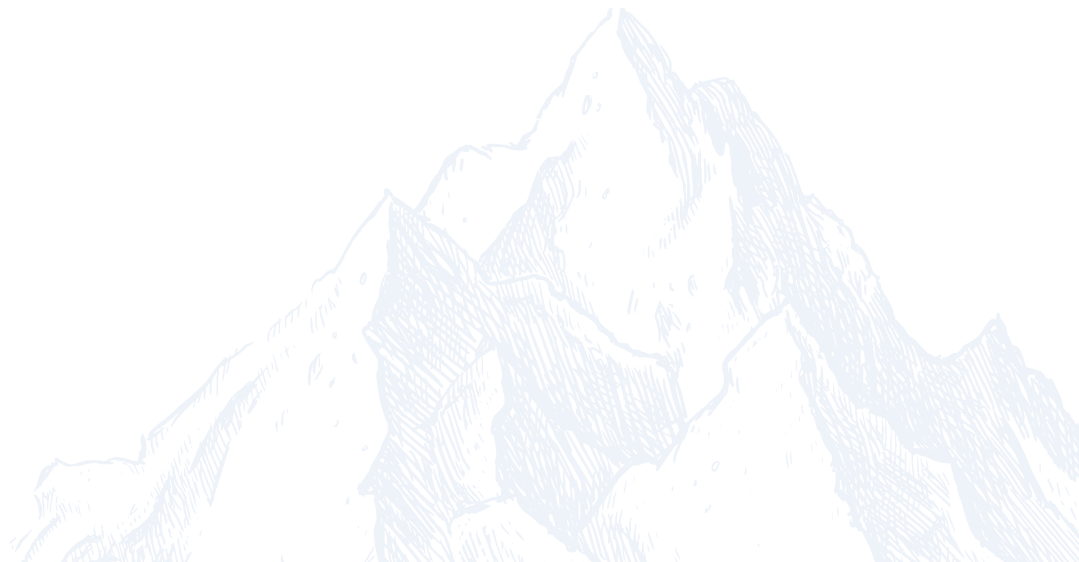
HOFFMANN CENTRE | TRUST SERIES

AI, Trust and Flooding the Zone

Beatrice Weder di Mauro

19 September 2026

Contribution to the HC/InTent Series on Rebuilding Trust



AI, Trust and Flooding the Zone

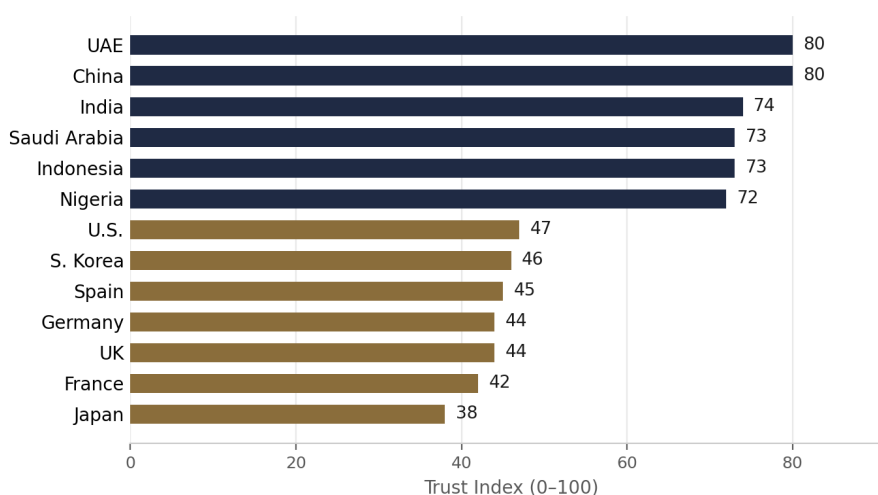
Beatrice Weder di Mauro

Society is losing its grip on how truth gets established.

The institutions that help establish truth like academia, the courts, the press or government agencies, depend on public trust. Yet that trust cannot be taken for granted. On Edelman's cross-country measure of trust in business, government, media and NGOs, several wealthy democracies sit near the bottom of the 2026 ranking, while China and the United Arab Emirates sit at the top (Edelman Trust Institute 2026). Neither high trust nor low trust tells us whether information is being properly checked.

Trust falls along the developed/developing divide

Average trust in business, government, media and NGOs, 2026



Trust Index (average trust in business, government, media and NGOs), 2026. Selected countries from the 28-market survey. Source: Edelman Trust Institute (2026, p. 5).

What puts warranted trust at risk? When producing credible-looking claims becomes cheaper faster than checking them does, verification becomes the bottleneck. The democratisation of media was supposed to liberate us from the old gatekeepers. As Paul Starr puts it:

“To the celebrants of digital democracy, the downfall of the public sphere’s gatekeepers counted as one of the chief benefits of the Internet. Speech would no longer need the permission of the great media corporations, their owners or publishers, editors or reporters, programming executives or producers. The online world has indeed afforded greater opportunities for the unfiltered expression of individual opinion and the unedited posting of images, videos, and documents. By the same token, however, the gates have swung wide open to rumors, lies, and increasingly sophisticated forms of propaganda, fraud, and deception.” (Starr 2020, p. 73)

Several mechanisms can weaken the link between trust and verification. One is the normalisation of official falsehood. Governments and organisations that once treated lying as a scandal increasingly treat it as a tactic, deployed openly and at little cost to those who use it, a shift traced in detail by scholars of the contemporary disinformation environment (Bennett and Livingston 2020). Once manipulation becomes a normal part of public communication, people stop assuming that official statements are true by default. That assumption, once lost, does not return on demand.

The second is noise, and it is worth being precise about how deliberate it can be. This is not only a by-product of a crowded information environment; it is a deliberate tactic used by foreign and domestic actors alike. Russia's version is particularly well documented: analysts describe a high-volume, multichannel propaganda model that disseminates messages rapidly, repetitively and continuously, without regard for truth or internal consistency. It was deployed in support of the 2014 annexation of Crimea and the conflicts that followed (Paul and Matthews 2016).

The third is the loss of a culture of correction. A functioning system for producing knowledge, science, journalism, expert commentary, depends on people admitting errors in public and revising their views in response. That is the mechanism by which truth is supposed to win out over time. What has become common instead is the opposite: when in doubt, deny and attack, and never concede a mistake. This encourages a retreat into cynicism. Once people conclude that truth is fundamentally unknowable, reasoned debate stops looking worthwhile: there is no agreed baseline of facts left to reason from. When denial replaces correction as the default response, the feedback loop that allows a system to correct its own errors breaks down, and falsehood persists.

The academic flood

The same dynamic is now visible inside the institution I know best. Peer review exists to certify which findings can be trusted, and it is buckling under a volume it was never built to handle. By May 2024, Wiley had retracted more than 11,000 papers over two years amid the Hindawi paper-mill crisis, which exposed industrial-scale manuscript fabrication and manipulation of peer review (Kincaid 2023; Subbaraman 2024). AI has also entered the reviewing side of the problem: an analysis by AI-detection company Pangram Labs classified roughly one in five reviews submitted to ICLR 2026, a major machine-learning conference, as fully AI-generated (Emi 2025). This is a detector-based estimate, not a verified count.

A column I wrote earlier this year set out the underlying economics for my own field: the supply of competent economics papers is rising fast as AI assists with drafting and methodology, while the number of top-five journal slots that confer career-making credibility remains at roughly 300–400 a year for a global pool of 8,000–11,000 professors at PhD-granting universities (Weder di Mauro 2026). Something has to give, and what tends to give first is the reliability of the signal a publication is supposed to send.

AI's dangers

AI reproduces this dynamic and adds two features that make it more consequential, not less. The first is that AI can sound confident and competent while being wrong, without bearing the social costs a person faces for the same mistake. The model has no reputation of its own at stake, no colleague whose confidence it loses, no next performance review: fluency is disconnected from personal accountability for whether the content is correct.

A recent field experiment offers some grounds for hope. Readers of *Süddeutsche Zeitung* were shown how difficult it is to distinguish real images from AI-generated ones. Their trust in news fell, including trust in the newspaper itself, but visits to its digital content and subscriber retention increased (Campante et al. 2025). This suggests that the relative value of reliable verification can rise, rather than disappear, once people recognise the threat. The risk is that people fail to notice the problem and simply grow accustomed to it.

The second feature is personalisation. Where social media capture attention at scale, AI can build something closer to a personal relationship, one exchange at a time. Users can shape these exchanges directly, telling the system what they like and dislike and adjusting its tone and personality until it feels as though they are talking to someone who knows them. That is a different, and in some ways more effective, form of capture than an engagement algorithm optimising across millions of anonymous users. People describing an AI as a companion, a confidant or even a partner are the visible edge of this: a system that can be endlessly agreeable and available at any hour, competing with human relationships that require reciprocity and carry real costs when handled badly.

These are not separate problems. They compound. A public that already has reason to distrust official statements, is exhausted by noise and has lost the habit of expecting anyone to admit error in public is now also drawing its information from systems that are fluent, lack personal accountability and can be optimised to be liked rather than to be right. The consequences extend beyond politics and media. Any institution that depends on people distinguishing reliable analysis from plausible-sounding analysis, be it a bank's research desk as much as an academic journal, is operating in the same degraded information environment.

But it is not inevitable

None of this is inevitable. The corrections that worked in earlier crises of trust relied on institutions, not technology alone: professional norms, editorial standards and mechanisms that made lying costly and detection likely. What is missing now is the equivalent for a world in which AI may produce most of the content and increasingly check it too: an accountable, auditable layer between fluent output and public trust. Someone identifiable must stand behind the judgement, and the record of what was checked, what was wrong and what was corrected must remain visible rather than disappear. Whether the same logic can be extended to journalism, public communication and the platforms themselves — before the next generation of AI systems becomes as embedded in daily life as social media already is — remains an open question.

Where this goes from here is unsettled, and what follows is speculation rather than a forecast. One path is bifurcation: a smaller audited sphere of verified journalism, refereed research, credentialed professional advice may be trusted precisely because its checks are effective, transparent and accountable; alongside a much larger sphere in which most people get their information and their sense of connection from AI systems optimised for engagement rather than correctness, much as information already flows through algorithmic feeds rather than edited publications.

A second path is that AI itself becomes part of the correction: systems trained and evaluated specifically to admit uncertainty, cite sources that can be checked and improve against a visible record of past mistakes, functioning less like an oracle and more like a well-supervised junior colleague. The aim is not slow verification, but verification that can scale without losing its independence or rigour.

Which path prevails probably depends less on what the technology can do than on whether anyone builds and pays for the institutions that would make the second path work. These are the same kinds of institutions that took decades to build for peer review, professional journalism and the courts, and that are now being asked to prove their worth again in far less time.

References

- Bennett, W. L., and S. Livingston, eds. (2020). *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States*. Cambridge University Press.
- Campante, F. R., R. Durante, F. Hagemeister and A. Sen (2025). *GenAI misinformation, trust, and news consumption: Evidence from a field experiment*. NBER Working Paper 34100; CEPR Discussion Paper 20526.
- Edelman Trust Institute (2026). *2026 Edelman Trust Barometer: Global Report*.
- Emi, B. (2025). *Pangram predicts 21% of ICLR reviews are AI-generated*. Pangram Labs, 18 November.
- Kincaid, E. (2023). *Hindawi reveals process for retracting more than 8,000 paper mill articles*. Retraction Watch, 19 December.
- Paul, C., and M. Matthews (2016). *The Russian “Firehose of Falsehood” Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation, Perspective PE-198.
- Starr, P. (2020). *The flooded zone: How we became more vulnerable to disinformation in the digital era*. In W. L. Bennett and S. Livingston, eds., *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States*. Cambridge University Press, pp. 67–92.
- Subbaraman, N. (2024). *Flood of fake science forces multiple journal closures*. The Wall Street Journal, 14 May.
- Weder di Mauro, B. (2026). *AI, the future of economic research and of the many-handed economist*. VoxEU/CEPR, 17 July.

CONTACT US

HOFFMANN CENTRE FOR GLOBAL
SUSTAINABILITY

Maison de la paix
Chemin Eugène-Rigot 2A
Petal 3, 4th floor
Case Postale 1672
CH-1211 Genève 1

hcgs@graduateinstitute.ch
www.graduateinstitute.ch/hcgs

