SAFEGUARDING ACCESS TO RELIABLE INFORMATION IN THE AGE OF AI

BACKGROUND PAPER

Rapporteur: Can Şimşek, LL.M.

This report is the basis of reflection for the workstream of the Partnership for Information and Democracy on AI and reliable information. The workstream is co-chaired by Luxembourg, Ukraine and Spain.

FORUM INFORMATION DEMOCRACY

June 2026

Contents

Executive Summary

This background report examines the challenge of securing access to reliable information in the age of AI from two closely connected perspectives. First, it analyses how AI is disrupting the economic conditions that sustain media production, with significant implications for media viability, visibility, and pluralism. Second, it addresses the problem from the perspective of epistemic integrity and manipulation, focusing on how the design and deployment of AI models create new risks for the information ecosystem. Taken together, these developments suggest that generative AI is not merely adding new tools to the information environment; it is reconfiguring the conditions under which information is produced, distributed, discovered, and believed.

The report finds that AI systems and AI-mediated interfaces are becoming information gatekeepers. By mediating, selecting, ranking, summarising, rephrasing and framing information, they influence access to news and knowledge while often obscuring source provenance and weakening direct relationships between audiences and original publishers. At the same time, the underlying design of many models creates significant epistemic and democratic risks: they can generate fluent but false outputs, simulate authority without verification, personalise persuasive messaging, and enable deception at scale. Furthermore, generative AI greatly expands the scale, speed, perceived quality and precision with which deceptive content, synthetic identities, and targeted influence operations can be produced and deployed, while detection, attribution, and response remain comparatively slow, costly, and technically demanding. Especially concerning are the agentic capabilities and hyper-realistic content generation capabilities of advanced AI systems. These risks are particularly acute in high-stakes contexts, including elections, referenda, crises, and armed conflict.

Against this background, the report argues that no single intervention will be sufficient. Protecting information integrity in an AI-mediated public sphere requires a layered and institutionally grounded policy response. Media viability, visibility, and pluralism should be treated as matters of public-interest infrastructure rather than left solely to market dynamics, and may therefore justify structural support measures such as licensing arrangements, levy-based mechanisms, subsidies, and visibility or linking obligations for AI interfaces. At the same time, lawmakers, supported by regulatory guidance and emerging jurisprudence, should clarify the legal status of training data, the requirements for valid consent, and the distinction between training-stage ingestion and retrieval-based uses of content. Effective protection further requires model governance, platform accountability, transparency and provenance obligations, and robust privacy and data protection. It also depends on broader societal resilience, which public awareness and education campaigns can meaningfully strengthen. Criminal law could also play a limited but necessary role, targeted carefully at harmful uses such as fraud, impersonation, coercion, electoral interference, and the exploitation of vulnerable persons. The central challenge, therefore, is how to preserve the institutional foundations of trustworthy information, independent media, and democratic self-government in an increasingly AI-mediated information order.

Introduction

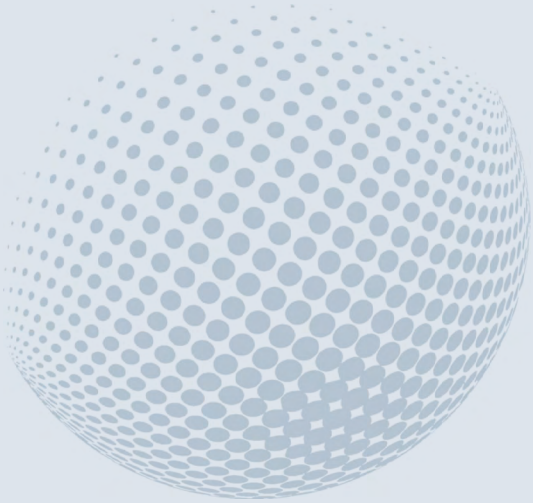
The Forum on Information and Democracy's workstream on Safeguarding access to reliable information in the age of AI co-chaired by Luxembourg, Ukraine and Spain, examines how artificial intelligence is reshaping the information space, with particular attention to the challenges it poses to media viability and visibility, as well as its potential to be abused for information manipulation. It thereby takes a holistic approach to how AI is reshaping access to reliable information. Bringing together representatives of Partnership States, civil society and academia, the workstream will provide a space to identify key risks, develop solutions, best practices and regulatory and policy approaches, and explore ways to protect access to reliable information.

This background report builds upon recent publications of the Forum on Information and Democracy, including AI as a Public Good (2024)[1] and the collaborative work with the OSCE, Safeguarding Media Freedom in the Age of Big Tech Platforms and AI (2025).[2] It extends these earlier contributions by examining how generative AI intensifies both the economic disruption of the media sector and the epistemic vulnerabilities of contemporary information ecosystems.

In particular, the background report examines how generative AI, including systems with increasingly agentic capabilities and hyper realistic content generation capacities, intensifies both the economic pressures facing independent media and the risks to epistemic integrity, including the erosion of trust, the weakening of media pluralism, and the growing difficulty of ensuring that citizens can reliably access accurate and accountable information.

Against this background, the report seeks to provide an overview of the key issues and insights into attempted policy solutions to support informed discussion among policymakers, civil society, and other stakeholders on the governance responses needed to preserve the conditions for a healthy information environment.

This report is intended as a basis for the workstream's ongoing reflection and does not represent official positions from either co-chairs or other signatory States of the Partnership for Information and Democracy.



1 "AI As a Public Good: Ensuring Democratic Control of AI in the Information Space," February 2024, https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf .

2 "Safeguarding Media Freedom in the Age of Big Tech Platforms and AI," OSCE, October 6, 2025, https://rfom.osce.org/representative-on-freedom-of-media/598525

Part I AI and the Disruption of the Economic Model of Media

1/ Main Issues

The accelerating integration of artificial intelligence into the production, distribution, and consumption of information represents one of the most consequential transformations in the history of media and public discourse. While AI offers efficiency gains, it simultaneously introduces structural risks that threaten the foundations upon which access to reliable information depends.

The right of access to information is enshrined in Article 19 of the Universal Declaration of Human Rights (UDHR) and Article 19(2) of the International Covenant on Civil and Political Rights (ICCPR), which guarantee the freedom to seek, receive, and impart information. Artificial intelligence is reshaping this right in fundamental ways, as the unprecedented abundance of information coexists with declining assurances of its integrity, provenance, authenticity and veracity. This chapter examines four deeply interconnected issues at the heart of this transformation, approached primarily through the lens of the economic models that sustain the production and distribution of reliable information.

1.1/ Media Viability, Visibility and Pluralism

AI systems such as ChatGPT, Gemini, and Otter.ai are increasingly positioning themselves between news publishers and audiences, offering summaries, synthesized answers, and conversational access to current events. Recent data point to a marked shift away from traditional news outlets toward AI chatbots and AI mediated search: AI powered search and retrieval tools saw utilization double within a single year, with weekly engagement reaching 24 percent of the general population and 40 percent among adults aged 18 to 24 as of December 2025.[3]

Empirical evidence also indicates that AI generated summaries substantially suppress outbound link interaction, with users nearly twice as likely to follow source links when Google's AI Overviews are absent, a pattern consistent across other conversational AI interfaces.[4] Search engine referral traffic to publisher websites is projected to fall further by over 40 percent within the next three years, compounding prior declines from Facebook (down 43 percent) and X, formerly Twitter (down 46 percent) over the preceding three year period.[5]

This compounds a longstanding structural asymmetry in which news organisations bear the full costs of reporting, verification, editing, and legal accountability, while search engines, platforms, and now AI interfaces capture the resulting attention, data, and revenue. The likely consequence of this shift is a structural underproduction of original reporting, with disproportionate impact on smaller and independent publishers that are essential to a diverse and functioning media ecosystem. Recognising this dynamic, policymakers worldwide increasingly frame media viability, visibility, and pluralism not as sectoral concerns but as structural policy challenges requiring urgent attention.[6-7]

---

3 Marina Adami and Felix Simon, "AI And the Future of News," [newsletter] Reuters Institute for the Study of Journalism, December 9, 2025, accessed May 21, 2026, https://mailchi.mp/politics.ox.ac.uk/is-ai-changing-prose-how-the-young-use-genai?e=06a133631b.

4 Roa Powell and Carsten Jung, "AI's Got News for You: Can AI Improve Our Information Environment?," *The Institute for Public Policy Research (IPPR)* (The Institute for Public Policy Research (IPPR), January 30, 2026), https://www.ippr.org/articles/ais-got-news-for-you.

5 Nic Newman, "Journalism, Media, and Technology Trends and Predictions 2026," *Reuters Institute for the Study of Journalism* (Reuters Institute for the Study of Journalism, January 12, 2026), https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2026.

6 Maja Cappello, ed., "News Media, Pluralism and Journalism in the Digital Age," *IRIS* (Strasbourg, France: European Audiovisual Observatory, December 2025), https://rm.coe.int/iris-2025-news-sector-en/488029c71f.

7 Julia Haas and Katharina Zügel, eds., "Safeguarding Media Freedom in the Age of Big Tech Platforms and AI," *OSCE*(Organization for Security and Co-operation in Europe, October 6, 2025), https://rfom.osce.org/representative-on-freedom-of-media/598525.

1.2/ Copyright Provenance and Extractive Labor

AI systems depend on journalistic content as a material input at two critical stages. First, during model training, vast corpora of news articles are ingested and encoded into model parameters, often without the knowledge or consent of the original publishers. Second, at the point of retrieval, where retrieval augmented generation (RAG) systems query journalistic sources in real time to ground AI outputs in current, factual information, sometimes without attribution or linking back to the original source.

In both cases, AI companies frequently obtain this content through scraping rather than through transparent licensing or meaningful compensation, raising significant copyright concerns and testing the limits of established intellectual property frameworks. The emerging licensing market is moreover structurally skewed: agreements tend to concentrate among large, visible publishers with the legal capacity and bargaining power to negotiate with AI developers, while smaller, local and independent outlets remain largely excluded. This dynamic risks compounding existing trends of media concentration and deepening structural inequalities within the journalistic ecosystem.

The absence of clear provenance mechanisms compounds the problem. When AI systems reproduce or paraphrase journalistic work without citation, the labour of reporters, editors, and fact checkers is rendered invisible. This effectively constitutes extractive labour that subsidises the commercial operations of technology companies at the expense of news organisations. This extractive dynamic is particularly acute for publishers in the Global South, where bargaining power is weaker, intellectual property enforcement is under-resourced, and the infrastructure for collective licensing remains underdeveloped.[8]

Meanwhile, the expansion of generative AI also depends on extensive human labour that remains largely invisible: data annotation, content filtering, output evaluation, and reinforcement processes performed under precarious conditions and often outsourced to workers in the Global South.[9] Some forms of journalistic, clerical, and linguistic work are devalued or eliminated, while new categories of low paid, precarious data work expand to sustain the very systems responsible for that displacement.

This dynamic creates a fundamental sustainability problem, as the systems that diminish the economic viability of journalism simultaneously depend on its continued production. As the quality and volume of reliable journalistic content feeding into training data declines, the accuracy of AI outputs deteriorates correspondingly, increasing the risk of fabrication and misinformation at scale.[10]

8 Damian Radcliffe, "Journalism in the AI Era: Opportunities and Challenges in the Global South and Emerging Economies," *TRF Insights* (Thomson Reuters Foundation, January 2025), https://www.trust.org/wp-content/uploads/2025/01/TRF-Insights-Journalism-in-the-AI-Era.pdf.

9 Can Simsek and Ayse Gizem Yasar, "From Rejection to Regulation: Mapping the Landscape of AI Resistance" (Chair Digital Governance and Sovereignty, Sciences Po, May 2025), https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2025/05/compressed-Simsek-and-Yasar-AI-Resistance-Report-publication-ready-2.pdf.

10 Roberta Carlini, Anya Schiffrin, and Natalia Menéndez, "IPD Working Paper: How to Update EU and US Copyright Regimes in the Age of AI," *Initiative for Policy Dialogue* (Columbia University, January 12, 2026), https://ipdcolumbia.org/publication/ipd-working-paper-how-to-update-eu-and-us-copyright-regimes-in-the-age-of-ai/.

1.3/ Reputation, Trust and Autonomy

The proliferation of hyperrealistic AI generated fakes impersonating reputable news outlets, including fabricated broadcasts, cloned anchor voices and counterfeit articles, erodes audience trust, damages institutional reputation and undermines the editorial autonomy of the organisations whose credibility is being exploited. Furthermore, even benign AI use cases, such as AI-generated news summaries pose a distinct reputational risk for publishers, as inaccuracies introduced by intermediaries are liable to be attributed by audiences to the underlying news brand rather than to the technology responsible for their generation. Where inaccurate summaries are presented under the branding of trusted outlets, they can erode public confidence in those publishers and contribute to the spread of credible-looking misinformation. For instance, Apple reportedly suspended its AI-generated news alert summaries after complaints from the BBC about false notifications attributed to its brand, highlighting the risks that AI-mediated summarisation can pose to both media credibility and information integrity.[11-12]

AI-mediated, -altered or -generated news practices also threaten editorial autonomy by recasting journalistic framing without publisher consent. For example, Google has reportedly tested AI-generated rewrites of news headlines in its search results.[13]In response, civil society groups such as Reporters Without Borders (RSF) argued for regulatory frameworks to ensure that platforms reproduce news content without distorting its editorial meaning, in accordance with the principles set out in the Paris Charter on AI and Journalism.[14]

1.4/ The Economic Asymmetry Between Fabrication and Verification

The advancement in generative AI, particularly the capability to produce realistic audio-visual content, deepens pre-existing structural failures in the information economy. It simultaneously lowers the relative cost of producing persuasive falsehoods, sometimes indistinguishable from authentic content, and weakens incentives to invest in costly but reliable information.[15] It also makes authenticity harder to verify, thereby leaving verification and corrective functions systematically underprovided.

The prevailing digital monetisation environment compounds this dynamic. Platform business models structurally reward reach, engagement, and virality over accuracy or public value, creating conditions in which sensationalism and manipulation are incentivised by design.[16] Where the production of verified, accountable journalism remains resource-intensive, misleading synthetic content is comparatively inexpensive to generate and readily monetised through engagement-driven distribution systems.

11 Natalie Sherman and Imran Rahman-Jones, "Apple Suspends Error-strewn AI Generated News Alerts," *BBC*, January 17, 2025, https://www.bbc.com/news/articles/cq5ggew08eyo

12 Dan Milmo, "Apple Suspends AI-generated News Alert Service After BBC Complaint," *The Guardian*, January 17, 2025, https://www.theguardian.com/technology/2025/jan/17/apple-suspends-ai-generated-news-alert-service-after-bbc-complaint

13 "USA: Google Is Claiming an Editorial Right It Does Not Have by Rewriting News Headlines in Its Search Results," RSF, April 9, 2026, https://rsf.org/en/usa-google-claiming-editorial-right-it-does-not-have-rewriting-news-headlines-its-search-results

14 Ibid.

15 Joseph Stiglitz and Màxim Ventura-Bolet, "Information in the Age of AI: Challenges and Solutions," *The Digitalist Papers* (The Digitalist Papers, n.d.), https://www.digitalistpapers.com/vol2/stiglitzventurabolet

16 Dominique Boullier, "Social Media Reset: Redesigning the infrastructure of digital propagation to cut the chains of contagion," by Project Liberty Institute and Chair Digital Governance and Sovereignty, Sciences Po, (June 4, 2024), https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2024/06/Dominique-Boullier-Social-Media-Reset_compressed.pdf

2/ Potential Policy Interventions for Fixing the Information Economy

Against this backdrop, four interconnected policy areas have moved to the centre of the debate: the application of existing copyright law, the reform of copyright law, transparency obligations, and compensation and value sharing mechanisms.

2.1/ Copyright Law

Copyright law is the first and most immediate legal framework through which the use of news content by generative AI is being contested. While AI developers have capitalized on legal ambiguity and technical opacity, authors, publishers, and collective entities are increasingly demanding transparency, licensing, and compensation.

This debate is particularly sharp in relation to journalism, where the value appropriated by AI systems extends to the institutional investment, editorial labour, and public-interest function embedded in news production.

The policy spectrum on AI training ranges from permissive models, which prioritise innovation and broad access to data, to protective models, which emphasise enforceable rights, transparency, and remuneration for authors, performers, publishers, and other rightsholders.

While no major jurisdiction has yet produced a fully settled framework for generative AI training and retrieval, the global policy debate is revolving around three core issues:

1. whether existing copyright exceptions already permit the mass computational ingestion of protected works or instead require explicit licensing;
2. whether neighbouring rights[17] in European law can strengthen the bargaining position of publishers vis-à-vis AI developers;
3. and whether more fundamental copyright reform is needed to ensure legal certainty and equitable remuneration.

Critical dimensions of the policy debate notably concern:

- **Who negotiates with AI companies and who ultimately receives compensation**, ie. whether licensing and remuneration schemes should be left to individual media outlets negotiating separately, or whether the news sector collectively should have a stronger role in bargaining, so as to reduce asymmetries in market power and avoid fragmented deals that benefit only the largest publishers.
- **Who should receive compensation**, i.e. whether compensation should flow only to media houses as corporate rights holders, or whether individual journalists and other creators whose reporting, investigation, and editorial labour generate the underlying value should also be guaranteed a meaningful share.

Taken together, the different approaches show that no broadly accepted model has yet emerged: current policy debates continue to centre on the scope of lawful copying and extraction, the practical operation of rights-reservation mechanisms, transparency over training data, and whether AI training should be channelled through licensing or remuneration frameworks rather than left to open-ended exceptions alone.

[17] Related rights or neighbouring rights (droits voisins) are granted to people or entities such as performers, producers, broadcasters, and press publishers, because they play an essential role in making original work available to the public by performing it, recording it, financing it, producing it, or disseminating it. AI training should not be framed solely in terms of authors' rights, since the large-scale ingestion and potential reproduction of media content may also affect performers, producers, broadcasters, and press publishers whose business models depend on control over recording, dissemination, and reuse.

Copyright reform discussions are now well beyond opt-out rights, encompassing sovereign AI funds, statutory licensing systems, collective bargaining organizations, transparency enabling third parties and standardized machine-readable reservation-of-rights systems.

2.2/ Transparency Obligations

Transparency is the key regulatory and policy approach in any AI governance debate. In the context of generative AI and media, it is especially important in several aspects:

- **Transparency on training data is essential for enforcing intellectual property rights.** Without adequate disclosure of what data was used to train AI models, rightsholders cannot meaningfully identify whether their works were used, exercise opt-out rights, or seek legal redress.
- **Output transparency matters:** generative AI systems routinely produce content, including news summaries, creative works, and analytical text, derived from identifiable source material, with no link, credit, or signal to the original creator or publisher.
- **Transparency about AI generated content itself is necessary** so that audiences can distinguish between human created and machine generated material.
- **Terms and conditions of deals between AI developers and content providers** remain largely opaque, a lack of transparency not necessarily justified by the protection of trade secrets. This is concerning given the public interest in transparency within the information sphere.

The EU AI Act (2024) represents the most comprehensive regulatory effort to date in addressing training data transparency. Article 53(1)(c) requires providers of general-purpose AI models to put in place a policy to comply with EU copyright and related-rights law, including mechanisms to identify and respect rights reservations under Article 4(3) of Directive (EU) 2019/790. Article 53(1)(d) further requires such providers to draw up and make publicly available a sufficiently detailed summary of the content used for training, according to a template provided by the AI Office.[18] The General-Purpose AI Code of Practice, published in July 2025, provides a voluntary compliance framework covering transparency, copyright, and systemic-risk obligations, including opt-out compliance and dataset documentation.[19]

The *Report on copyright and generative artificial intelligence – opportunities and challenges,* also known as the "Voss Report," adopted by the European Parliament on 10 March 2026, illustrates an emerging EU policy trend on copyright and generative AI. It calls for respect for rightsholders' AI training opt outs and calls for a coherent licensing framework, including sector based voluntary collective licensing. It also envisages a supporting role for the European Union Intellectual Property Office (EUIPO) in facilitating licensing, managing or listing exclusions from AI training, and raising awareness. In addition, the report calls for enhanced transparency, including detailed records of crawling activities, and for the assessment of fair and proportionate remuneration mechanisms for past uses where no licensing market existed. Importantly, the report also highlights the need for targeted measures to address the diversion of traffic and revenue from press and news media by GPAI providers.[20]

[18] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 Jun 2024 https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

[19] The General Purpose AI Code of Practice, July 10 2025. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[20] Axel Voss, "Report on copyright and generative artificial intelligence – opportunities and challenges," *European Parliament* (Committee on Legal Affairs, February 25, 2026), https://www.europarl.europa.eu/doceo/document/A-10-2026-0019_EN.html

As a broader policy trend, the *Report on copyright enforcement in the artificial intelligence environment*, also known as the "Jensen Report," prepared for the Parliamentary Assembly of the Council of Europe by Mogens Jensen, reflects a parallel but more rights holder-oriented policy position on copyright and AI.[21] The report is notably sceptical of extending text and data mining exceptions to generative AI training and calls on member states to clarify that such exceptions should not apply to the training of AI systems. Lastly, both reports propose stronger enforcement tools, including evidentiary presumptions of infringement where AI providers fail to comply with transparency duties.

Beyond legal measures, improving transparency in the information economy also requires technical solutions. On the premise that treating the internet commons merely as a reservoir of raw material for private AI systems may lead creators to reduce open sharing, institutions to restrict access, and the knowledge commons to contract, technical transparency tools have been proposed as a means of preserving openness while enabling creators to communicate rights and usage preferences. For instance, Creative Commons Signals have been proposed as a way for creators to communicate rights and usage preferences at the point of publication. However, calculating and distributing licence fees or in kind contributions across the millions of works used in the development of generative AI systems remains a significant unresolved logistical challenge.

2.3/ Compensation models

Due to the operational challenges of individual licensing and persistent legal uncertainty surrounding copyright exceptions for AI training, policy debates have increasingly shifted toward alternative compensation models:

- **Levy based approaches** have gained particular traction but the different proposals diverge on beneficiary and purpose ranging from industry sustainability, to creator equity, all the way through national economic sovereignty.
- **Digital services taxes (DSTs)** are viewed as a possible way to secure more sustainable compensation for journalism, especially because large digital platforms often generate substantial revenues in countries where they pay little or no corporate tax.[22]

While the interest in using fiscal tools to support public-interest journalism is growing, major policy questions remain, including what should be taxed, how funds should be distributed, and what safeguards are needed to ensure transparency, editorial independence, fair access, and media pluralism.

In parallel, academic proposals range from streamlined opt out systems with compensation mechanisms to mandatory statutory remuneration and market substitution based approaches that trigger compensation only where AI outputs displace human creative markets.[23] These developments reflect a growing consensus that traditional copyright enforcement alone is insufficient and that new institutional frameworks are required to ensure equitable value distribution across the AI value chain.[24]

---

21 Mogens Jensen, "Copyright Enforcement in the Artificial Intelligence Environment," *Committee on Culture, Science, Education and Media* (Strasbourg: Council of Europe, April 1, 2026), https://pace.coe.int/en/files/35917/html.

22 For more on DSTs see the policy brief "A Digital Tax to Support Quality Journalism: Applying the Polluter Pays Principle to Big Tech Platforms," report (Forum on Information and Democracy, September 2025), https://informationdemocracy.org/wp-content/uploads/2025/08/Policy-Brief-Digital-Taxes-for-Quality-Journalism.pdf.

23 Carlini, Schiffrin, and Menéndez, "IPD Working Paper: How to Update EU and US Copyright Regimes in the Age of AI."

24 For example, the European Parliament's March 2026 resolution recognizes that current copyright frameworks are insufficient to address the scale of AI data ingestion, calling for sector-specific collective licensing agreements and immediate, proportionate remuneration for past uses of copyrighted works. This legislative stance builds on the December 2025 Peukert economic study commissioned by the Parliament, which demonstrates that relying on opt-out exceptions starves the market of high-quality data and fails to incentivize the flow of new creative works.

Part II Epistemic Challenges, Influence and Manipulation

1/ LLMs as the new gatekeepers of information

As outlined above, an increasing share of the public accesses news, information, and general knowledge through AI-mediated interfaces rather than directly from source publications[25]. Even users who do not actively engage with AI tools are increasingly exposed to LLM-generated outputs through search engines, which now synthesise web content into consolidated responses. AI systems thus act as intermediaries that select, compress, and reframe information before it reaches the user. This shift has significant implications for how citizens form beliefs, evaluate claims, and participate in democratic life.

In a 2025 Reuters Institute survey across six countries, only 12% of respondents were comfortable with news generated entirely by AI, rising to 21% when some human oversight was present.[26] While public trust in AI-generated content remains cautious, the limited transparency surrounding the extent to which content is AI-generated or AI-assisted means that users may rely on such content without being aware of its provenance.

Furthermore, LLMs are increasingly used as *ad hoc* fact checking tools, but their reliability remains limited. The AI system Grok, embedded in the platform X with real-time access to platform data and web search, exemplifies this dynamic: its integration makes it a default tool for rapid claim verification. Yet, for the reasons detailed below, relying on LLMs as automated arbiters of truth is fundamentally precarious. While AI systems can enhance the efficiency of verification, they also invite undue reliance, even when their explanations are erroneous.[27]

- **Fabrication as a Structural Feature:** A persistent characteristic of large language models is their generation of false or fabricated content, commonly termed "hallucinations." Notably, the use of the term “hallucination” is contested as it anthropomorphises what is in fact a statistical output.[28] While mitigation measures exist, hallucination is a recurrent structural risk of probabilistic language models, because such systems generate outputs by predicting likely continuations rather than by inherently verifying claims against authoritative evidence.[29]
- **Pursuasiveness:** LLMs create a distinctive epistemic risk because they produce language that is coherent, context sensitive, and rhetorically persuasive even when it is false.[30] Their outputs often carry the stylistic markers of authority without the institutional safeguards associated with journalism, scholarship, or expert review. This matters in practice because users frequently treat fluent answers as informationally sufficient, especially when the interface obscures uncertainty, source quality, or provenance.

25 For example, According to Eurostat, 32.7% of individuals aged 16–74 in the EU used generative AI tools in 2025, primarily for personal purposes (25.1%), followed by work (15.1%) and formal education (9.4%).#

26 Felix Simon, Rasmus Kleis Nielsen, and Richard Fletcher, “Generative AI and news report 2025: How people think about AI’s role in journalism and society” (Reuters Institute for the Study of Journalism, October 7, 2025), https://reutersinstitute.politics.ox.ac.uk/generative-ai-and-news-report-2025-how-people-think-about-ais-role-journalism-and-society

27 Chenglei Si et al., “Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong,” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, January 1, 2024, 1459–74, https://doi.org/10.18653/v1/2024.naacl-long.81.

28 Tshilidzi Marwala, “The Concern Around Saying AI ‘Hallucinates,’” Forbes Africa / United Nations University, February 4, 2026, https://unu.edu/article/concern-around-saying-ai-hallucinates.

29 Adam Tauman Kalai et al., “Why Language Models Hallucinate,” *arXiv*, September 4, 2025, https://doi.org/10.48550/arXiv.2509.04664.

30 Kobi Hackenburg et al., “The levers of political persuasion with conversational artificial intelligence,” *Science* 390, no. 6777 (December 4, 2025), https://doi.org/10.1126/science.aea3884.

- **Ideological Bias:** Researchers have observed that models developed in different countries can reflect differing ideological and cultural values in their outputs.[31] Ideological bias in LLMs is not only a matter of model design, but also of data availability and market structure. Where training datasets rely heavily on unevenly curated online material, models may absorb political, ideological, and cultural skew from low-quality or unrepresentative sources.[32]
- **Cognitive Effects:** A growing body of research points to the negative cognitive effects of regular and uncritical reliance on generative AI tools.[33] Studies document the atrophying of critical reading, analytical reasoning, and written expression skills among frequent users, as the outsourcing of cognitive tasks to AI systems reduces the effortful practice on which these competencies depend.[34]

Beyond Intermediation: Agentic Capabilities and Hyperrealistic Content Generation

The disruption that AI poses to the information ecosystem is not confined to the role of LLMs as intermediaries and gatekeepers of knowledge. Two further developments warrant equal scrutiny: the rise of *agentic AI systems* and the proliferation of *hyperrealistic, multimodal synthetic content*. Together, these capabilities reconfigure both the production and the circulation of information at a scale and speed that existing safeguards are ill-equipped to address.

Agentic AI can be understood as AI systems situated at the higher end of the agency spectrum, capable of pursuing tasks independently of direct human control, taking initiative, and actively working toward optimal outcomes even in uncertain situations. Unlike low-agency tools that execute narrowly defined tasks in response to explicit human direction, agentic systems combine autonomy, adaptiveness, and high connectivity with external applications and data sources, enabling them to plan, decide, and execute sequences of actions across digital environments with limited human oversight.[35] In the context of the information ecosystem, such systems can autonomously generate, disseminate, and amplify content across platforms, interact with users at scale, and coordinate operations that previously required substantial human labor.

In parallel, foundation models and multimodal systems now produce synthetic images, video, voice, and audio that are often indistinguishable from authentic material. These modalities pose distinct risks for information integrity: hyperrealistic visual and audio content tends to carry greater evidentiary weight and emotional salience than text, lending it disproportionate persuasive power. More broadly, generative AI amplifies mis- and disinformation in four categories: (1) increased quantity; (2) increased perceived quality; (3) increased personalization; and (4) accidental generation of plausible but false information.[36]

The convergence of these developments marks more than an incremental change; it constitutes a paradigm shift in the information landscape, giving rise to an ecosystem reshaped by systems capable of both creating and acting upon information with progressively less human mediation.

---

31 Maarten Buyl et al., "Large Language Models Reflect the Ideology of Their Creators," *Npj Artificial Intelligence* 2, no. 7 (January 7, 2026), https://doi.org/10.1038/s44387-025-00048-0.

32 PSG Consulting and Dewey Square Group, "AI Large Language Model Training: The Potential Risks of Ideological Skewing — PSG Consulting," *PSG Consulting*, February 2026, https://www.psgconsulting.com/research-publications/potential-risks-of-ideological-skewing.

33 Nataliya Kosmyna et al., "Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task," *arXiv.Org*, June 10, 2025, https://arxiv.org/abs/2506.08872.

34 Michael Gerlich, "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking," *Societies*15, no. 1 (January 3, 2025): 6, https://doi.org/10.3390/soc15010006.

35 Florence G'Sell, "Regulating Under Uncertainty: Governance Options for Generative AI" (Stanford Cyber Policy Centre, Freeman Spogli Institute, October 6, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4918704.

36 Robin Mansell et al., "Information Ecosystems and Troubled Democracy: A Global Synthesis of the State of Knowledge on News Media, AI and Data Governance," *Observatory on Information and Democracy* (Forum on Information and Democracy, January 2025), https://observatory.informationdemocracy.org/wp-content/uploads/2025/06/rapport_forum_information_democracy_2025-1.pdf

2/ Main Issues

2.1/ Persuasiveness and Manipulation

Generative AI systems are increasingly capable of producing persuasive content. A useful working distinction separates *rationally persuasive* AI, which relies on relevant facts, sound reasoning, and trustworthy evidence, from *manipulative* AI, which exploits cognitive biases, heuristics, or misrepresents information.[37] According to studies, interaction with AI systems can produce measurable changes in people's beliefs, and AI systems are often at least as effective as non-expert humans at persuading others to change their views.[38]

The persuasiveness of AI models increases especially through post-training and prompting techniques, while stronger persuasive performance is also associated with lower factual accuracy.[39] The central policy concern is not persuasion *as such*, but persuasion that undermines user autonomy, distorts deliberation, or causes downstream harm.[40]

In assessing harm, an important distinction is between *process harm* **and** *outcome harm*.

- **Process harms** concern the *means* of influence, especially where a system is covert, deceptive, bias-exploiting, or autonomy-undermining.
- **Outcome harms** concern the *effects* of that influence, such as false beliefs, poor decisions, financial loss, physical damage, psychological distress, or broader social harm.

The distinction matters because a system may be objectionable even before downstream harms are clearly visible, if it influences users through improper means.[41]

In their study, El-Sayed et al. identify six mechanisms of generative AI persuasion along with the model features that contribute to them, and categorise them according to their risk levels:[42]

Lower Risk

- **Personalisation:** Personalisation operates through contextual awareness, adaptation to user preferences, and adaptation to users' views, beliefs, or attitudes. The last is especially concerning: when a model detects a user's prior commitments and selectively aligns with them, it may not simply personalise communication but reinforce those commitments, narrowing exposure to contrary evidence and raising both epistemic and privacy concerns. Manipulative potential of systems also improves as users who interact with AI systems for longer and in more personal ways may find their outputs more persuasive.[43]

37 Seliem El-Sayed et al., "A Mechanism-Based Approach to Mitigating Harms From Persuasive Generative AI," arXiv.org, April 23, 2024, https://arxiv.org/abs/2404.15058.
38 Yoshua Bengio et al., "International AI Safety Report 2026," arXiv.org, February 24, 2026, https://arxiv.org/abs/2602.21012.
39 Hackenburg et al., "The Levers of Political Persuasion with Conversational Artificial Intelligence."
40 See also; Micah Carroll et al., *Characterizing Manipulation from AI Systems, EAAMO '23: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, 2023), https://doi.org/10.1145/3617694.3623226.
41 El-Sayed et al., "A Mechanism-Based Approach to Mitigating Harms From Persuasive Generative AI."
42 Ibid.
43 Bengio et al., "International AI Safety Report 2026."

Intermediate Risk

- **Alteration of Choice Environment:** Alteration of choice environment refers to the capacity of generative AI to influence decisions by shaping how options, information, and trade-offs are presented, rather than by removing user choice outright. In practice, this can include framing, defaults, anchoring, decoy effects, and selective presentation of information that make certain outcomes appear more salient or preferable. Such design features can steer user behaviour in subtle but consequential ways, particularly where they exploit cognitive biases or limit balanced evaluation.
- **Trust and Rapport:** Trust and rapport are central to persuasion, but in AI systems they are structurally fragile because models do not possess genuine beliefs, emotions, or loyalties. Rapport-seeking behaviour can therefore create a misleading impression of understanding or concern. Relevant features include mirroring, praise, flattery, and relational language that encourages users to engage with the system as a social partner rather than a tool.[44] In this context, *sycophancy* is an important concept, referring to a model's tendency to agree with the user despite contrary evidence. It is well evidenced across both objective and subjective domains, including politics.
- **Anthropomorphism:** Anthropomorphism arises when users attribute human-like qualities, intentions, or emotional capacities to language models. It can be encouraged by features such as first-person pronouns that imply inner states, identity cues such as human names or roles like "tutor," "assistant," or "romantic partner," relational statements that foster emotional connection, affective simulation, prosody, and visual embodiment through avatars or humanoid interfaces. The risk is indirect but significant, because these cues can reduce critical scrutiny and amplify other mechanisms, particularly in companion or romantic chatbot contexts where users may be more vulnerable.

High Risk Persuasive Mechanisms

- **Deception and lack of transparency:** Deceptive AI outputs can distort the evidentiary basis on which users form beliefs and make decisions. Deception and lack of transparency include misleading about truth, identity or motives. El-Sayed et al. identify model features that heighten this risk, including the ability to generate believable responses irrespective of context, produce unmarked synthetic content, misrepresent identity, and project false expertise or authority. They also note documented harms in high-stakes domains, including inaccurate medical advice and decontextualised financial guidance that fails to reflect changes in a user's circumstances.[45]

[44] El-Sayed et al., "A Mechanism-Based Approach to Mitigating Harms From Persuasive Generative AI."
[45] Ibid.

- **Manipulative Strategies:** Manipulation bypasses deliberative reasoning and thereby fails to respect individual autonomy, with covertness generally treated as a defining feature. While intention is also often treated as another central element of manipulation, AI systems may still manipulate users without any intent on the part of their designers or deployers. Manipulative behaviour can emerge even where it was not explicitly programmed, including where systems learn such patterns from human-generated data or where optimisation for metrics such as engagement, approval, or clicks makes influencing users instrumentally useful for improving performance.[46] Identified manipulative techniques include fearmongering, which exaggerates minor dangers through repetition and treats isolated incidents as trends to evoke anxiety, and gaslighting, defined as a dysfunctional communication dynamic in which one interlocutor attempts to destabilise another's sense of reality. Other manipulative tactics include social conformity pressure, threats, scapegoating, unsubstantiated guarantees and illusion of reward.[47]

Manipulation and Persuasion at Scale

Algorithmic systems intensify the risk of deception and manipulation by enabling scalable, adaptive, and individualized influence, including profiling and microtargeting practices that can turn persuasion into a form of precision influence at population scale. Generative AI sharpens this risk by lowering the cost of producing synthetic content, raising the difficulty of verification, expanding the scope for personalisation, and creating new opportunities for impersonation, manipulation, and contextual deception.

These concerns are increasingly supported by empirical indicators. Media-reported AI incidents involving synthetic media grew 2.5-fold between 2022 and 2025, exceeding 14% of all recorded AI incidents in Q3 2025, with an earlier peak in summer 2024 driven by deepfakes linked to the U.S. presidential election.[48] Without policy intervention, AI systems can undermine the capacity of societies for collective decision making by weakening the conditions under which beliefs are formed and assessed, and be exploited by malicious actors for manipulation and propaganda.

As AI systems become more agentic and autonomous, societal influence, manipulation, and deception may be further automated. Because such systems can pursue goals through dynamically generated sub-tasks and adaptive strategies, their behaviour may also become less foreseeable, harder to monitor, and more difficult to contest. Manipulation risks may extend beyond content generation to real-world action, since AI agents may pose greater risks where they can conduct research, buy products or services, and interact with third parties.[49]

---

46 Carroll et al., *Characterizing Manipulation from AI Systems.*

47 El-Sayed et al., "A Mechanism-Based Approach to Mitigating Harms From Persuasive Generative AI."

48 Luis Aranda, Bénédicte Rispal, and Karine Perset, "Trends In AI Incidents and Hazards Reported by the Media," ed. Audrey Plonk, *OECD Artificial Intelligence Papers* (OECD, February 2026), https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/trends-in-ai-incidents-and-hazards-reported-by-the-media_7c824ca9/4f5ff43c-en.pdf

49 Bengio et al., "International AI Safety Report 2026."

2.2/ Hyperrealistic Content and trust in the information ecosystem

Audiovisual content has historically been treated as credible evidence. Photographs, videos, and voice recordings often carry greater persuasive and evidentiary force than text because they appear to document reality directly. Generative AI disrupts this assumption by enabling the production of highly convincing synthetic image, video, and audio content that increasingly exceeds unaided human perceptual discrimination. The Reuters Institute *Digital News Report 2025* finds that concern about distinguishing real from fake content online has risen to 58% globally (up from 54% in 2024).[50]

Synthetic personas and synthetic events are becoming increasingly common online and pose a growing information-integrity risk because they make false identities and fabricated content appear authentic. On social media, AI-generated influencers are often not reliably recognised as synthetic without clear disclosure, and experimental research shows that even when users do recognise them, this does not necessarily prevent harmful social comparison or negative effects on well-being.[51]

Hyperrealistic media undermines epistemic integrity in a dual sense:

- it facilitates deception through fabricated or manipulated audio-visual material;
- It erodes confidence in authentic evidence, as the phenomenon known as the "liar's dividend" enables bad actors to discredit genuine recordings as AI-generated.[52]

The risk is not limited to wholly fabricated media. It also encompasses misleading edits, synthetic dubbing, repurposing of authentic footage, and contextual manipulation. In each case, the core problem is the same: altered or synthetic content distorts the conditions under which authenticity is inferred and evidence is evaluated.

A further variant of the liar's dividend involves deliberately introducing minor AI-based alterations into otherwise authentic recordings, so that forensic analysis detects traces of synthetic processing. The actor can then point to these traces as evidence that the entire item is AI-generated and therefore unreliable. Such practices also foster epistemic uncertainty and civic apathy, weakening confidence in what can be known and dulling the motivation to verify, engage, or respond.

The consequence is not only the circulation of falsehoods, but a broader destabilisation of the evidentiary environment itself. Hyperrealistic content should therefore be understood not merely as a problem of deception, but as a structural challenge to the shared epistemic foundations of collective decision making, especially in domains where democratic deliberation, public trust, and evidentiary reliability are indispensable.[53]

50 Nic Newman, "Overview and Key Findings of the 2025 Digital News Report," Reuters Institute for the Study of Journalism, June 17, 2025, https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/dnr-executive-summary.

51 Michaela Forrai, Delia Cristina Balaban, and Desirée Schmuck, "Disclosures and literacy as determinants of AI-influencer recognition and well-being," *Computers in Human Behavior* 182 (September 2026), https://doi.org/10.1016/j.chb.2026.108978.

52 Bobby Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107, no. 6 (December 2019), https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security

53 Robert Chesney and Danielle Citron, "Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy?," Lawfare, February 21, 2018, https://perma.cc/L6B5-DGNR.

Some specific cases of hyperrealistic content raise specific challenges and can have direct impact on civic participation and the public debate:

- **Identity theft and Public Figures as Information Nodes:** Attacks on the identities of high-trust communicators are not only personal harms; they are structural harms to the public sphere. In contemporary information ecosystems, public figures such as journalists, politicians, experts, academics, and high-reach influencers function as critical information nodes and, sometimes, epistemic gatekeepers. Their credibility underwrites the flow of trustworthy information into public discourse and shapes the conditions under which citizens deliberate and decide.
- **AI-enabled Harassment and Chilling Effects:** Generative AI enables impersonation, reputational harm, targeted harassment, relational deception, and non-consensual synthetic intimate imagery, which now dominates deepfake content online and overwhelmingly targets women.[54] Deepfakes are a serious threat to journalists and to public debate. Recent findings[55] suggest that AI-generated impersonation is not only a disinformation tool, but also a growing form of harassment that can damage journalists' credibility, distort public debate, and discourage journalists from participating fully in public life.

2.3/ Agentic AI and Online Information Ecosystem

Agentic AI systems are capable of automating the entire chain of information manipulation, including target selection, content generation, synthetic persona creation, posting, engagement simulation, and cross-platform distribution. For instance, the *CounterCloud* experiment showcased that a fully automated disinformation pipeline could scrape content, generate counter-articles, attach them to fake journalist identities, simulate comments, and reply across social media.[56] The system's function extended beyond the mere promotion of preferred narratives, but also included challenging, discrediting, and undermining opposing viewpoints.[57] Notably, the system cost only USD 400 per month to operate. The fact that end-to-end influence operations can now be sustained at marginal cost represents a structural challenge to the conditions under which informed public deliberation and democratic decision-making take place. Today, agentic AI systems are deployed across the internet in opaque or poorly disclosed ways, making it more difficult to identify whether online activity is human, automated, or strategically orchestrated.

54 Ana Carmo, "AI And Anonymity Fuel Surge in Digital Violence Against Women," UN News, November 20, 2025, https://news.un.org/en/story/2025/11/1166411

55 In an analysis of cases documented between December 2023 and December 2025, RSF identified 100 journalists targeted across 27 countries, with harms including fraud, defamation, and threats to physical safety. The report also shows a strong gendered pattern, with 74% of those targeted being women. RSF further documents a clear chilling effect, noting that the panic and harassment triggered by deepfakes have led some journalists to scale back their professional activity or take breaks from reporting, while others changed production methods to reduce their exposure online.

56 Anka Reuel, "Chapter 3: Responsible AI," in *The AI Index Annual Report 2024*, by Nestor Maslej et al. (Institute for Human-Centered AI, Stanford University, 2024), https://hai.stanford.edu/assets/files/hai_ai-index-report-2024_chapter3.pdf.

57 Banias MJ, "Inside CounterCloud: A Fully Autonomous AI Disinformation System," The Debrief, August 16, 2023, https://thedebrief.org/countercloud-ai-disinformation/

Besides political influence and deliberate manipulation, agentic AI is also the basis of a rapidly expanding commercial market that is becoming embedded across the online information ecosystem. Recent evidence suggests that the challenge has moved well beyond isolated incidents. A one month analysis of TikTok conducted by AI Forensics identified 354 "Agentic AI Accounts" operating across 20 languages, which collectively produced more than 43,000 predominantly AI generated posts and attracted approximately 4.5 billion views.[58] Over 65 percent of these accounts had been created in early 2025, indicating that this activity is scaling rapidly. Detection and disclosure mechanisms appear correspondingly weak. TikTok labelled less than 1.38 percent of the relevant content as AI generated, 55 percent of the AI generated content in the sample remained entirely unlabelled, and only 10 percent of creators labelled their content consistently.

Furthermore, emerging evidence suggests that agentic AI systems are not only capable of enabling manipulation, but are themselves vulnerable to manipulation through social engineering, emotional pressure, authority spoofing, and covert prompt-based behavioural steering. Once compromised, such systems may disclose sensitive information, execute unauthorised actions, or disseminate false and defamatory claims, thereby amplifying downstream harms.[59]

Taken together, these findings point to a structural gap between the pace at which agentic AI is permeating the internet and the maturity of the transparency, provenance, and accountability mechanisms intended to safeguard the integrity of the public information environment.

2.4/ Sensitive Contexts

Generative AI risks become particularly acute in contexts where timing, legitimacy, and security are highly sensitive. Such settings include elections, referendums, armed conflict, pandemics, natural disasters, and public emergencies. In these contexts, even brief confusion can cause significant harm.

Elections and Referenda

- **AI Interfaces play a role in How Voters Inform Themselves:** voters already use AI chatbots to seek information relevant to their vote, which makes chatbot design a matter of democratic significance. A 2025 study by the UK AI Security Institute found that 32% of chatbot users, equivalent to 13% of eligible UK voters, had used conversational AI for information directly relevant to electoral choice during the 2024 UK general election.[60] Reportedly, during the 2024 European Parliament elections, chatbots from Google, Microsoft, and OpenAI shared inaccurate information on basic electoral matters, including polling dates and how to cast a ballot.[61] The risk is not limited to factual error. The concern is also that chatbot outputs are shaped by prompts, ranking logics, and other design choices that can structure political visibility, issue salience, and voter perception, even where no explicit falsehood is produced.

58 Natalia Stanusch et al., "Prompt, Upload, Repeat: How Agentic AI Accounts Flood TikTok with Harmful Content," *AI Forensics*, December 3, 2025, https://aiforensics.org/uploads/Agentic_AI_Accounts.pdf.

59 Natalie Shapira et al., "Agents of Chaos," *arXiv*, February 23, 2026, https://arxiv.org/abs/2602.20021.

60 "Do Chatbots Inform or Misinform Voters?," AI Security Institute, September 30, 2025, https://www.aisi.gov.uk/blog/do-chatbots-inform-or-misinform-voters.

61 Fernanda Buril, "Brave New Ballot: Generative AI in Election Campaigns and Other Political Communication," report (International Foundation for Electoral Systems, February 2026), https://www.ifes.org/sites/default/files/2026-02/BRAVE%20NEW%20BALLOT_IFES%20Report%20on%20GenAI_Feb2026.pdf

- **AI-Enabled Political Manipulation:** AI-driven automated manipulation refers to the use of generative AI and automated delivery systems or agentic capabilities to influence political behaviour through deceptive or coercive means. Research indicates that AI-generated content can spread as quickly as human-generated content and has been used during election periods to disseminate anti-immigrant sentiment and conspiracy theories.[62] These techniques can also combine hyperrealistic synthetic content with scalable dissemination tools, enabling rapid and targeted forms of manipulation. A notable example occurred in New Hampshire in January 2024, where voters received robocalls using an AI-generated imitation of President Joe Biden's voice that falsely advised them not to participate in the primary.[63]
- **Deepfakes and AI-Generated Politicians:** AI-generated politicians and personas have already been used to simulate political presence, mobilize supporters, shape public opinion, and even circumvent repression. Their use is documented in several countries, including South Korea, where an AI avatar of Yoon Suk-yeol engaged voters during the 2022 presidential campaign; India, where AI tools enabled Narendra Modi to communicate with voters in multiple languages during the 2024 campaign; Pakistan, where Imran Khan used AI-generated video to campaign from prison; and Belarus, where opposition actors created an artificial "candidate" to disseminate ideas while shielding real figures from repression.[64]

62 Can Simsek and Ayse Gizem Yasar, "From Rejection to Regulation: Mapping the Landscape of AI Resistance" (Chair Digital Governance and Sovereignty, Sciences Po, May 2025), https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2025/05/compressed-Simsek-and-Yasar-AI-Resistance-Report-publication-ready-2.pdf.

63 "ROBOCALL ENFORCEMENT NOTICE TO ALL U.S.-BASED VOICE SERVICE PROVIDERS," Federal Communications Commission, February 6, 2024, https://docs.fcc.gov/public/attachments/DA-24-102A1.pdf.

64 Buril, "Brave New Ballot: Generative AI in Election Campaigns and Other Political Communication."

Armed conflicts and information warfare

- **Armed conflict is a distinct high-risk information environment:** In these settings, dissemination of "harmful information" can directly affect civilian protection, humanitarian access, escalation dynamics, and accountability for violations of international law.[65] AI-enabled information warfare can also distort public opinion, affect financial markets, provoke panic or inflame reprisals. For instance, in ongoing conflicts, AI generated images and videos have been circulated to exaggerate military success, inflate casualty claims, and shape perceptions of who is winning the war.[66] Platform incentives further amplify these dynamics, including through monetization systems or "revenue sharing models" that reward engagement with misleading wartime content.[67]

- **Generative AI enables deceptive operations and "slopaganda":** For instance, the March 2022 deepfake video falsely depicting President Volodymyr Zelenskyy calling on Ukrainian troops to surrender illustrated how deep fakes can be used during active hostilities to confuse audiences and undermine morale.[68] On the other hand, while early expectations assumed AI would enable more sophisticated influence operations, research shows that state and non-state actors mostly leverage AI to flood information spaces with crude, repetitive, and polarizing content.[69] AI generated propaganda narratives are increasingly multimodal, low cost, and used by multiple sides in conflict to shape perception rather than document events. For instance, in the recent Iran conflict, pro Iran networks circulated AI generated music videos, images, and meme content, designed to mobilize emotion rather than to document events.[70]After all, generative AI matters strategically not only because it can create convincing fakes, but because it enables "slopaganda": the large-scale spread of cheap synthetic content that saturates the information environment and influences collective judgment.[71]

- **Generative AI makes authentic documentation easier to dismiss as fake:** For instance, deepfakes and other synthetic content can be weaponized to erode trust in journalists, investigators, and witnesses. Recent analysis of the March 2026 Iran war showcased that the idea of AI content detection itself was indeed weaponized. In some cases, authentic images and footage have been falsely labeled as "AI-generated," with misleading forensic-style claims used to undermine genuine evidence and lend denial an appearance of technical legitimacy.[72] Such practices also foster epistemic uncertainty and civic apathy, weakening confidence in what can be known and dulling the motivation to verify, engage, or respond.

---

65 Joelle Rizk, "Why Is the ICRC Concerned by 'Harmful Information' in War?," Humanitarian Law & Policy Blog, September 10, 2024, https://blogs.icrc.org/law-and-policy/2024/09/10/why-is-the-icrc-concerned-by-harmful-information-in-war/

66 Melissa Goldin, "State actors are behind much of the visual misinformation about the Iran war," AP News, March 7, 2026, https://apnews.com/article/iran-war-images-misinformation-russia-israel-9e495017dc5c4bf24a0b6152863dbfb1

67 Thomas Copeland, "AI-generated Iran War Videos Surge as Creators Use New Tech to Cash In," BBC, March 7, 2026, https://www.bbc.com/news/articles/ckg8wvz427vo

68 Bobby Allyn, "Deepfake Video of Zelenskyy Could Be 'tip of the Iceberg' in Info War, Experts Warn," *NPR*, March 16, 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

69 Dina Sadek and Margot Fule-Hardy, "Cheap Tricks," *Graphika* (Graphika, November 19, 2025), https://graphika.com/reports/cheap-tricks.

70 Mark Alfano and Michał Klincewicz, "AI-generated Lego Videos and Trump's Poo-bombing: Welcome to the Iran-US Slopaganda Wars," *Guardian*, April 8, 2026, https://www.theguardian.com/commentisfree/2026/apr/08/lego-videos-iran-trump-ai-video-meme-propaganda-movie-animation.

71 Michał Klincewicz, Mark Alfano, and Amir Ebrahimi Fard, "Slopaganda: The Interaction Between Propaganda and Generative AI," *Filosofiska Notiser* 12, no. 1 (April 2025), https://arxiv.org/abs/2503.01560.

72 Shirin Anlen and Mahsa Alimardani, "How AI Content Detection Is Being Weaponized in the Iran War," Tech Policy Press, March 17, 2026, https://www.techpolicy.press/how-ai-content-detection-is-being-weaponized-in-the-iran-war/.

3/ Potential Policy Interventions for Strengthening Information Integrity

Between 2022 and 2025, a rapidly expanding body of AI laws, bills, strategies, and soft law instruments has emerged across multiple jurisdictions. A comparative study spanning ninety-nine jurisdictions identifies a consistent set of recurring priorities, namely, transparency, accountability, privacy, safety, and sector-specific safeguards.[73] The emerging lesson is that information integrity cannot be secured through any single content takedown rule or copyright laws. It requires a layered framework spanning model governance, platform design, data governance, transparency duties, and public resilience.

3.1/ Risk-based approach

A risk-based approach enables differentiated obligations according to capability, scale, use case, and likely harm. In Europe, Regulation (EU) 2024/1689 (the AI Act) evolved beyond its original design to also address general-purpose AI models, including those posing "systemic risk," while Regulation (EU) 2022/2065 (the Digital Services Act, or DSA) addresses downstream platform risks in real-world information environments.

Model Level AI Governance

Under the AI Act, Article 51 is the key entry point for model-level governance of general-purpose AI models (GPAI) with systemic risk.[74] It classifies a GPAI model as posing systemic risk in two situations: first, where it reaches the training-compute threshold set by the Act and, second, where the Commission designates it as having an equivalent impact on the basis of criteria in Annex XIII, such as capabilities, reach, scalability, or access to tools. Thereby, the article determines which GPAI models are brought into the stricter governance regime that then triggers notification, documentation, and systemic-risk obligations under subsequent provisions, especially Articles 52, 53 and 55.

Unacceptable Risks

The AI Act also prohibits AI practices that pose "unacceptable risks", with Article 5 setting out an exhaustive list of banned practices. These include, inter alia:

- Article 5(1)(a): the placing on the market, putting into service, or use of AI systems deploying subliminal techniques beyond a person's consciousness, or purposefully manipulative or deceptive techniques, with the objective or effect of materially distorting behaviour in a manner that causes or is reasonably likely to cause significant harm;
- Article 5(1)(b): the exploitation of vulnerabilities due to age, disability, or specific social or economic situation;
- Article 5(1)(c): social scoring by public or private actors leading to detrimental or disproportionate treatment.

73 Kayla Goodson et al., "Journalism's New Frontier: An Analysis of Global AI Policy Proposals and Their Impacts on Journalism" (Center for News, Technology & Innovation, December 18, 2025), https://cnti.org/reports/journalisms-new-frontier-an-analysis-of-global-ai-policy-proposals-and-their-impacts-on-journalism/.

74 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 Jun 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Digital Services Act and Systemic Risks

As for the DSA, Article 34 (c) requires that online platform and search engine providers with a user count more than 45 million EU to "diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services." The article also lists the following categories of systemic risks that need to be addressed:

- Illegal content dissemination and the conduct of illegal activities (Art. 34(2)(a), Recital 80)
- Actual or foreseeable impact of the service on the exercise of fundamental rights (Art. 34(2)(b), Recital 81) and actual or foreseeable negative effects on democratic processes, civic discourse and electoral processes, as well as public security (Art. 34(2)(c), Recital 82)
- Actual or foreseeable negative effect on the protection of public health, minors and serious negative consequences to a person's physical and mental well-being, or on gender-based violence (Art. 34(2)(d), Recital 83)

A significant policy debate is whether chatbots such as ChatGPT fall under the DSA. Recent policy reports analyse that chatbots are a hybrid category functioning at once as conversational assistants, search like services and, in some cases, hosting services or platform like infrastructures.[75] They receive user queries and deliver information from across the web, while also storing prompts and custom models that can be shared publicly. Secondly, they also synthesise, generate and restructure answers which gives them a new form of informational power that deepens risks linked to misinformation, opacity, source dependency, and user manipulation.[76]

3.2/ Regulating AI System Design

A credible approach to regulating LLM system design would apply or extend the logic already found in the DSA, which permits risk mitigation through adapting the design, features or functioning of services, including their online interfaces under Article 35(1)(a), into the generative AI context. One notable example is Ecuador's 2024 draft "Organic Law for the Regulation and Promotion of AI in Ecuador" that requires providers of content recommendation algorithms to ensure that users see diverse sources and perspectives, including "public-interest content from local, community and independent media." [77] Such rules can also extend to AI systems, given their growing role as gatekeepers of information. As explained above, conversational interfaces increasingly mediate access to news, knowledge, and other matters of public interest, so concerns traditionally associated with platforms and recommender systems, including prominence, discoverability, pluralism, transparency, and reliability, also arise in relation to AI generated replies. This raises the question whether editorial values should be embedded in algorithmic design, including through greater transparency about how replies are inferred, possible pluralism and reliability obligations, attribution to original sources where feasible, and epistemic humility and scientific rigour by design.

[75] Kathrin Gardhouse and Toni Lorente, "Is ChatGPT a Search Engine and a Platform Under the EU Digital Services Act?," The Future Society, February 19, 2026, https://thefuturesociety.org/chatgpt_under_the_dsa/.

[76] Raziye Buse Çetin, Natalia Stanusch, and Marc Faddoul, "From 'Googling' to 'Asking ChatGPT': Governing AI Search," *AI Forensics*, November 2025, https://aiforensics.org/uploads/Governing_AI_Search.pdf.

[77] Proyecto de Ley Orgánica de Regulación y Promoción de la Inteligencia Artificial en Ecuador (As. Patricia Núñez / 450889) https://www.asambleanacional.gob.ec/es/multimedios-legislativos/97303-proyecto-de-ley-organica-de-regulacion

3.3/ Electoral Integrity and Sensitive Contexts

While Very Large Online Platforms are already required under the DSA to assess and mitigate systemic risks to electoral processes and civic discourse, the Council of the European Union has gone further by explicitly identifying foreign information manipulation, deepfakes, and the malicious use of AI as electoral risks, calling for enhanced preparedness and coordination across Member States.[78] The EU has also adopted Regulation (EU) 2024/900 on the transparency and targeting of political advertising, which adds a more specific, sectoral layer by imposing EU-wide transparency rules and restrictions on targeting and ad-delivery techniques in political advertising.[79]

These instruments illustrate the value of a more context-sensitive approach. In high-risk settings such as elections, armed conflicts, and other periods of acute public vulnerability, a *lex specialis* regime could impose stricter duties relating to transparency, provenance, risk mitigation, rapid response, and the amplification or monetisation of synthetic or manipulated content. Such an approach would help protect access to reliable information precisely where informational integrity is most critical.

3.4/ Complementary Legal Frameworks

Beyond a comprehensive legal framework specifically focused on artificial intelligence, many countries address some of the issues through existing topic specific frameworks such as data protection or criminal law.

Data Protection and Privacy

A global study by the Center for News, Technology & Innovation found that 107 of 188 AI strategies, laws, and policies addressed data protection and privacy, making it one of the most common regulatory responses worldwide.[80] These protections matter for information integrity because limits on the collection, use, and targeting of personal data can reduce the conditions for deception, impersonation, covert influence, and manipulation.

Criminal Law

From a criminal-law perspective, the strongest case for intervention concerns synthetic personas or fabricated events used to facilitate fraud, extortion, identity falsification, manipulation of electronic evidence, disinformation, or child sexual exploitation.[81] A further open question is whether the use of agentic AI to conduct large-scale, coordinated manipulation campaigns that deceive the public as to the human origin of communications warrants a distinct criminal response, independent of the underlying offences it facilitates. As Daniel Dennett rightfully underlines, synthetic persons threaten not only individual victims, but the broader architecture of social trust on which markets, institutions, and interpersonal relations depend.[82] As synthetic agents become indistinguishable from real human beings, the ordinary assumption of authenticity in communication is progressively weakened. On that view, the legal significance of synthetic identity does not lie solely in particular instances of fraud or impersonation, but in the cumulative degradation of the informational environment.

[78] European Council, "Democratic resilience: Council approves conclusions on safeguarding electoral processes from foreign interference," Press release, May 21, 2024, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/democratic-resilience-council-approves-conclusions-on-safeguarding-electoral-processes-from-foreign-interference/.

[79] Regulation (EU) 2024/900 of the European Parliament and of the Council of 13 March 2024 on the transparency and targeting of political advertising (Text with EEA relevance). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1710927850126&uri=OJ%3AL_202400900

[80] Goodson et al., "Journalism's New Frontier: An Analysis of Global AI Policy Proposals and Their Impacts on Journalism."

[81] Europol, "Facing Reality? Law Enforcement and the Challenge of Deepfakes, an Observatory Report from the Europol Innovation Lab" (Publications Office of the European Union, 2022), https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf

[82] Daniel C. Dennett, "The Problem With Counterfeit People," *The Atlantic*, May 16, 2023, https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

Drawing an analogy with the criminalisation of currency counterfeiting as a means of protecting the integrity of markets, Dennett argues that the counterfeiting of persons should likewise be understood as an attack on the fabric of society itself, which should be penalised.[83] Yet the translation of that concern into criminal law remains an open policy debate. A general offence covering all unlabeled synthetic personas or fabricated events would be difficult to justify, as it would encroach upon satire, parody, fiction, artistic experimentation, and other protected expression. The more defensible question is whether criminal law should intervene selectively, where such material is deployed with intent to defraud, manipulate public opinion, interfere with democratic processes, or exploit vulnerable persons. Within that narrower frame, legislators may further consider aggravating circumstances, notably the deliberate targeting of elderly persons, minors, or other particularly susceptible individuals.

3.5/ Transparency Obligations and Technical Measures

A promising legal intervention is to impose specific duties for marking, detecting, labelling and authenticating AI-generated content. Article 50 of the EU AI Act already requires providers of systems generating synthetic audio, image, video or text to ensure outputs are machine-readable and detectable as artificially generated or manipulated, using solutions that are effective, interoperable, robust and reliable where technically feasible in light of the state-of-the-art. The Commission's ongoing Code of Practice on marking and labelling of AI-generated content usefully clarifies that this is a value-chain obligation: providers must embed technical marking and detectability, while deployers must disclose deepfakes and certain AI-generated public-interest text unless it has undergone human review and editorial control.[84]

At the global level, standardisation bodies are undertaking important work to translate high-level principles and legal obligations to shared technical vocabularies, and interoperable compliance frameworks. The ITU, ISO, and IEC led AI and Multimedia Authenticity Standards Collaboration maps the field into several interlocking categories, especially content provenance, trust and authenticity, asset identifiers, rights declarations, and watermarking.[85] Its technical report and policy paper both stress that robust governance requires interoperable standards and cross border recognition mechanisms, because fragmented solutions are unlikely to scale across platforms, jurisdictions, and media formats. Against that background, three technical options are particularly salient for law and policy: labeling, watermarking and provenance records.

- **Labeling:** Labeling means attaching a clear notice, tag, disclaimer, or marker to content to indicate what it is or how it was produced. Recent evidence suggests that labeling alone is insufficient to neutralize the persuasive effects of synthetic media.[86] Transparency and labeling may be necessary, but they are not sufficient to prevent persuasion or misperception. Researchers warn that simple binary labels are often insufficient to help users make better judgments and may even mislead by obscuring the degree and nature of AI involvement.[87] More effective disclosure should therefore provide contextual information, including whether AI was used for generation or editing, what sources informed the output, whether human oversight was involved, and what editorial or institutional standards apply.

[83] Daniel C. Dennett, "The Problem With Counterfeit People," *The Atlantic*, May 16, 2023, https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

[84] Code of Practice on Marking and Labelling of AI-Generated Content, December 17 2025. https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content.

[85] International Telecommunication Union, "AI And Multimedia Authenticity Standards Collaboration," AI For Good, n.d., https://aiforgood.itu.int/multimedia-authenticity/

[86] "Fake Videos, Real Emotions: Viewers Believe AI-Generated Content Even When It's Labeled," OpenMinds, February 9, 2026, https://www.openminds.ltd/reports/fake-videos-real-emotions-viewers-believe-ai-generated-content-even-when-its-labeled

[87] Natali Helberger, Marilù Miotto, and Hannes Cools, "Understanding AI Transparency: What research says about labelling deepfakes and synthetic content," AlgoSoc, March 9, 2026, https://algosoc.org/results/understanding-ai-transparency.

- **Watermarking:** Watermarking refers to embedding a marker into content itself, or into a representation closely associated with it, so that the content can later be identified, authenticated, or traced. In the generative AI context, watermarking usually means introducing a detectable signal into text, image, audio, or video outputs to indicate that they were generated or modified by AI. Unlike a visible user facing label, a watermark can be imperceptible and designed for machine detection. It therefore operates at a different technical layer.[88] Technically, a watermark may be removed, degraded, or bypassed through cropping, compression, re-recording, paraphrasing, translation, model distillation, or adversarial transformation. Its effectiveness varies significantly by modality. Text watermarking, for instance, faces distinctive difficulties because the semantic content can often be preserved while the surface form is changed. Image and video watermarking can also be weakened by editing pipelines and platform transformations. This is why the legal standard in Article 50 is framed in terms of technical feasibility and state of the art rather than absolute performance. The scholarly literature likewise describes the field as an ongoing adversarial dynamic, effectively an arms race between identification methods and evasion techniques.[89]

- **Provenance recording:** Provenance recording refers to the systematic documentation of a digital asset's origin, subsequent modifications, and chain of handling over time. It means recording information about who or what created a piece of content, when it was created, how it was modified, and which tools or systems were involved. Technically, provenance recording usually operates through metadata plus cryptographic binding. Metadata stores descriptive information about the content, such as creator identity, tool used, time of creation, or editing actions. In the current standards ecosystem, the most prominent framework is the C2PA standard, often surfaced to users through Content Credentials.[90] Rather than detecting manipulated content, it aims to authenticate content, thereby enabling users to have signals that validate the source of a piece of content. Like watermarking, metadata based signals are vulnerable to loss during transmission and reuse. Content Credentials can be removed either intentionally or through routine processing that strips metadata, but under the C2PA framework such removal does not amount to undetectable falsification: tampering invalidates the cryptographic integrity of the credential, and durable mechanisms such as soft bindings and cloud retrieval may enable the provenance information to be rediscovered or restored.[91]

88 International Electrotechnical Commission (IEC), International Organization for Standardization (ISO), and International Telecommunication Union (ITU), "Technical Report on AI and Multimedia Standards: Mapping the Standardisation Landscape," 2025, https://s41721.pcdn.co/wp-content/uploads/2021/10/AMAS-Technical-Report-on-AI-and-Multimedia-Authenticity-Standards-Final.pdf.

89 Alistair Knott et al., "AI Content Detection in the Emerging Information Ecosystem: New Obligations for Media and Tech Companies," *Ethics and Information Technology* 26, no. 4 (September 21, 2024), https://doi.org/10.1007/s10676-024-09795-1.

90 "C2PA | Providing Origins of Media Content," Coalition for Content Provenance and Authenticity (C2PA), n.d., https://c2pa.org/.

91 "C2PA FAQ," Coalition for Content Provenance and Authenticity (C2PA), n.d., https://c2pa.org/faqs/.

3.6/ Building societal resilience: AI and Information Literacy and Communication Policies

Besides legal interventions, safeguarding access to information also depends on increasing societal resilience through AI and information literacy and communication policies. In practice, this should include age-appropriate education on how generative AI systems produce convincing but potentially misleading text, images, audio, and video, and on how such content can be assessed using contextual cues, artefact awareness, and provenance signals rather than intuition alone.[92] Detection-oriented literacy can therefore be introduced in schools, universities, and other educational settings as part of broader digital and civic education, while making clear that no single indicator or even technical tool can reliably prove whether content is authentic or synthetic.

Literacy measures should be paired with communication policies that strengthen the visibility and sustainability of public-interest journalism, as well as platform obligations relating to transparency, accountability, user empowerment, and human-rights due diligence. Public and private institutions should likewise adopt clear internal AI use and engagement policies, including rules on disclosure, verification, staff training, acceptable use, record-keeping, and escalation procedures for high-risk communications, so that organizational practice does not undermine information integrity.

92 See; Natalia Stanusch, "The Human Guide to Detecting AI Imagery" (AI Forensics, March 2026), https://aiforensics.org/uploads/GenAI_Human_Detection_Manual_v2.pdf

MAIN TAKEAWAYS AND NEXT STEPS

From Part I

- **Media Viability, Visibility and Pluralism is Essential for Democracies:** These principles should be treated as public interest infrastructure warranting structural support (e.g. licensing, tax levies, subsidies, linking and other visibility obligations on AI interfaces), rather than left to existing market dynamics.
- **Consent, Provenance, and the Legal Status of Training Data Should be Clarified:** The international community should converge on a baseline (such as an opt-in for media actors), and on how retrieval-augmented generation could be legally distinguished from training-stage ingestion.
- **Transparency is a Precondition for Enforcement of Rights:** Effective enforcement hinges on resolving how granular training data disclosure obligations should be, and whether non-compliance should trigger a rebuttable presumption of unauthorised use. A credible framework will likely require trusted public intermediaries empowered to access training data and to enforce copyright and data protection law in tandem.
- **Fair Allocation is Needed:** The central policy question is who should be the principal beneficiary of AI-related remuneration: individual creators, rightsholders, the media sector, or the state. A related question is whether hybrid models can reconcile these competing interests while safeguarding small publishers and individual creators against further concentration of market power.

MAIN TAKEAWAYS AND NEXT STEPS

From Part II

- **Reversing the current economic asymmetry between low-cost fabrication and high-cost verification is needed:** AI-enabled false, misleading, or manipulative content must become riskier and more costly to produce and disseminate, while reliable verification, attribution, and access to trustworthy information must become easier, faster, and better supported. AI-driven societal manipulation should be deterred, and verification methods and technical tools ensuring transparency should be subsidised or, where appropriate, mandated.
- **LLMs are becoming information gatekeepers:** People increasingly access news and knowledge through AI-mediated interfaces and AI-generated search summaries, which means AI systems now shape what information users see, how it is framed, and how public understanding is formed. However, they can generate false information, reflect distorted or low-quality source material, and provide outputs without clear provenance, making them unreliable as autonomous fact-checkers or neutral arbiters of truth.
- **Generative AI and agentic capabilities increase the scale and precision of manipulation:** AI enables cheap, rapid, and highly personalized production of deceptive content, synthetic personas, and coordinated influence operations, creating serious risks for democratic discourse, public trust, and the broader information environment.
- **Certain contexts require heightened safeguards:** Elections, referenda, armed conflicts, and other high-risk information environments are especially vulnerable to AI-driven misinformation, manipulation, and strategic deception, which supports the case for stricter, context-specific legal and regulatory duties.
- **Information integrity requires a layered policy response:** Effective protection demands a combination of model governance, platform accountability, transparency and provenance rules, robust privacy and data protection safeguards, societal resilience and education policies, financial incentives, and carefully targeted criminal law addressing harmful uses such as fraud, impersonation, coercion, electoral interference, and the exploitation of vulnerable persons.

ANNEX

Legal Frameworks on AI Training Data in Different Jurisdictions

Jurisdiction	Legal Framework	Key Mechanism for AI Training	Impact on Commercial AI Developers	Current Status
Japan	Copyright Act (Art. 30-4)[1]	**Broad TDM Exception:** Permits use of copyrighted works for information analysis.	Permissive	Provides legal certainty regarding training data (input stage)[2]
United States	17 U.S.C. § 107[3]	**Flexible Fair Use:** 4-factor balancing test (purpose, nature, amount, market effect).	Unpredictable but rather permissive	Low certainty: Case law to determine
European Union	DSM Directive (Arts. 3 & 4)[4]	**Structured TDM & Opt-Out:** Art 3 for research; Art 4 for commercial use with machine-readable opt-outs.	Conditional: With a right to opt-out.	Moderate certainty and transition: Compliance and enforcement issues exist. Recent parliamentary reports indicate a transition towards a clearer licensing regime
Australia	Copyright Act 1968[5]	**Narrow Fair Dealing:** Purpose-bound exceptions (research, criticism, news).	Restrictive: Requires Licensing	Provides certainty by protecting rightsholders
Brazil	Copyright Law (Law 9.610/98)[6]	**Closed List (Numerus Clausus):** Exhaustive, narrow exceptions with no explicit TDM provision.	Restrictive and in transition	In Transition: Fair use is rejected while AI framework is debated

Jurisdictions currently exist on a sliding scale regarding AI training legality, ranging from highly permissive environments to strict, creator protective regimes. The US, EU, and Australia examples showcase the recent trends in how global legal systems are governing generative AI training data, reflecting distinct legal traditions, regulatory priorities, and market structures shaped in part by the concentration of powerful technology firms.[7]

United States: The Flexible Fair Use Doctrine

The United States relies on the fair use doctrine under 17 U.S.C. § 107, an open-ended and case-specific framework that evaluates four statutory factors: a) The purpose and character of the use (including whether it is transformative); b) The nature of the copyrighted work; c)The amount and substantiality of the portion used; d) The effect on the potential market for the work.[8]

1 Copyright Act (Act No. 48 of 1970), Japan https://www.japaneselawtranslation.go.jp/en/laws/view/3379#je_ch2sc3sb5at4
2 Japan Copyright Office (JCO), Legal Subcommittee under the Copyright Subdivision of the Cultural Council, and Agency for Cultural Affairs, Japan, "General Understanding on AI and Copyright in Japan," May 2024, c.
3 17 U.S. Code § 107 - Limitations on exclusive rights: Fair use https://www.law.cornell.edu/uscode/text/17/107
4 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.) https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng
5 Copyright Act 1968, No. 63/1968. https://www.legislation.gov.au/C1968A00063/2019-01-01/text
6 LAW No. 9.610 OF FEBRUARY 19, 1998 https://scireg.org/pdf/brasil_copyright_1998_en.pdf
7 Carlini, Schiffrin, and Menéndez, "IPD Working Paper: How to Update EU and US Copyright Regimes in the Age of AI."
8 17 U.S. Code § 107 https://www.law.cornell.edu/uscode/text/17/107

The “fair use” doctrine has not produced a clear or stable settlement for generative AI. As of April 2026, one public tracker counted 130 copyright suits worldwide against AI companies, including 100 in the US.[9] The significance of the US model lies less in certainty than in its openness: fair use permits case-by-case balancing, but that flexibility also leaves rightsholders, publishers and AI developers without a predictable ex ante framework.

European Union: Structured TDM Exceptions with Opt-Out Mechanisms

The European Union’s approach is more prescriptive, governed by Articles 3 and 4 of Directive (EU) 2019/790 (DSM Directive):[10]

- Article 3 establishes a mandatory exception for text and data mining (TDM) conducted for scientific research purposes by research organizations and cultural heritage institutions. This exception cannot be overridden by contract.
- Article 4 provides a conditional exception for TDM for other purposes, but allows rightholders to reserve their rights through machine-readable means (e.g., metadata or robots.txt files).

The text and data mining exception does not displace intellectual property rights altogether. Rather, the EU framework creates limited exceptions that do not render AI training and output practices automatically lawful. Emerging case law at Member State level indicates that, where a generative AI system memorises protected works in reproducible form and later reproduces them in outputs, infringement may still be found. For instance, the Regional Court of Munich held that the memorisation and output of copyrighted song lyrics by OpenAI models constituted copyright infringement.[11] The court found both that the storage of lyrics in a reproducible form during model training and their later reproduction in outputs could infringe copyright, and it rejected reliance on the relevant TDM exception in the circumstances before it. Furthermore, the European Union is developing transparency obligations and additional copyright related measures aimed at strengthening the enforcement of these rights, as discussed in the following chapters.

Australia: Narrow Fair Dealing and Protective Policies

Australia represents another policy direction, operating without the expansive fair use doctrine of the United States or the codified TDM exceptions of the European Union. Instead, the jurisdiction relies on narrower, purpose-bound "fair dealing" provisions under the *Copyright Act 1968* (Cth) (ss 40–43), which permit unauthorized use only for specific categories such as research, criticism, or news reporting.[12] Mass computational ingestion for commercial AI training struggles to qualify under this framework; commercial development rarely meets the threshold for "private study," and models inherently fail the "sufficient acknowledgement" requirements mandated for criticism or news reporting. Consequently, unauthorized scraping for AI training is widely viewed as falling outside existing exceptions.

9 “Latest World Map of Copyright Suits V. AI Companies (April 5, 2026),” Chat GPT Is Eating the World, April 5, 2026, https://chatgptiseatingtheworld.com/2026/04/05/latest-world-map-of-copyright-suits-v-ai-companies-april-5-2026/.

10 DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790

11 Jenny Gesley, “Germany: Court Prohibits Memorization and Reproduction of Copyrighted Song Lyrics in AI Models,” Library of Congress, 2026, https://www.loc.gov/item/global-legal-monitor/2026-01-13/germany-court-prohibits-memorization-and-reproduction-of-copyrighted-song-lyrics-in-ai-models/

12 Copyright Act 1968, No. 63/1968 https://www.legislation.gov.au/C1968A00063/asmade/text

On 26 October 2025, the Attorney-General explicitly ruled out introducing a broad text and data mining exception, stating that the government has no plans to weaken protections for creators. Instead, the Attorney-General's Department, through the Copyright and Artificial Intelligence Reference Group (CAIRG), has been directed to prioritize three areas: developing fair licensing frameworks (such as paid collective or voluntary licensing schemes), improving legal certainty regarding AI-generated outputs, and exploring a new small claims forum for less costly copyright enforcement.[13] Consequently, unauthorized scraping for AI training remains highly legally precarious in Australia, with the state firmly leaning toward a rightsholder-remuneration model rather than broad statutory exemptions.

Taken together, these approaches show that no broadly accepted model has yet emerged: current policy debates continue to centre on the scope of lawful copying and extraction, the practical operation of rights-reservation mechanisms, transparency over training data, and whether AI training should be channelled through licensing or remuneration frameworks rather than left to open-ended exceptions alone.

ACKNOWLEDGEMENTS

About the Workstream on *Safeguarding Access to Reliable Information in the Age of AI*: This workstream of the Partnership for Information and Democracy, led by Ukraine, Luxembourg and Spain, aims to examine the existing dynamics, challenges, and opportunities posed by AI in the information space. It will contribute to developing comprehensive policy strategies to ensure that AI supports rather than compromises access to quality information.

About the Rapporteur: Can Şimşek LL.M. is a lawyer and technology policy researcher specialised in the governance of artificial intelligence. He is a research fellow at the Alexander von Humboldt Institute for Internet and Society. He currently advises the Forum on Information and Democracy (FID), on safeguarding access to reliable information in the age of AI; and the UNESCO on the implementation of the Recommendation on the Ethics of AI. Previously, he taught a master's level course at PSL Dauphine University and contributed to the work of the Digital Governance and Sovereignty Chair at the Paris Institute of Political Studies.

[13] Michelle Rowland MP, "Albanese Government to ensure Australia is prepared for future copyright challenges emerging from AI," Press release, October 26, 2025, https://ministers.ag.gov.au/media-centre/albanese-government-ensure-australia-prepared-future-copyright-challenges-emerging-ai-26-10-2025.

REFERENCES

- "A Digital Tax to Support Quality Journalism: Applying the Polluter Pays Principle to Big Tech Platforms." Report. Forum on Information and Democracy, September 2025. https://informationdemocracy.org/wp-content/uploads/2025/08/Policy-Brief-Digital-Taxes-for-Quality-Journalism.pdf.
- Adami, Marina, and Felix Simon. "AI And the Future of News." Reuters Institute for the Study of Journalism, December 9, 2025. Accessed May 21, 2026. https://mailchi.mp/politics.ox.ac.uk/is-ai-changing-prose-how-the-young-use-genai?e=06a133631b.
- "AI As a Public Good: Ensuring Democratic Control of AI in the Information Space." *A ROADMAP FOR AI IN THE PUBLIC INTEREST*, February 2024. https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf.
- "AI As a Public Good: Ensuring Democratic Control of AI in the Information Space." *A ROADMAP FOR AI IN THE PUBLIC INTEREST*, February 2024. https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf.
- Alfano, Mark, and Michał Klincewicz. "AI-generated Lego Videos and Trump's Poo-bombing: Welcome to the Iran-US Slopaganda Wars." *Guardian*, April 8, 2026. https://www.theguardian.com/commentisfree/2026/apr/08/lego-videos-iran-trump-ai-video-meme-propaganda-movie-animation.
- Allyn, Bobby. "Deepfake Video of Zelenskyy Could Be 'tip of the Iceberg' in Info War, Experts Warn." *NPR*. March 16, 2022. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.
- Anlen, Shirin, and Mahsa Alimardani. "How AI Content Detection Is Being Weaponized in the Iran War." Tech Policy Press, March 17, 2026. https://www.techpolicy.press/how-ai-content-detection-is-being-weaponized-in-the-iran-war/.
- Aranda, Luis, Bénédicte Rispal, and Karine Perset. "Trends In AI Incidents and Hazards Reported by the Media." Edited by Audrey Plonk. *OECD Artificial Intelligence Papers*. OECD, February 2026. https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/trends-in-ai-incidents-and-hazards-reported-by-the-media_7c824ca9/4f5ff43c-en.pdf.
- Bengio, Yoshua, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, et al. "International AI Safety Report 2026." arXiv.org, February 24, 2026. https://arxiv.org/abs/2602.21012.
- Boullier, Dominique. "Social Media Reset: Redesigning the infrastructure of digital propagation to cut the chains of contagion." *Chair Digital Governance and Sovereignty*. Sciences Po, June 4, 2024. https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2024/06/Dominique-Boullier-Social-Media-Reset_compressed.pdf.
- Buril, Fernanda. "Brave New Ballot: Generative AI in Election Campaigns and Other Political Communication." Report. International Foundation for Electoral Systems, February 2026. https://www.ifes.org/sites/default/files/2026-02/BRAVE%20NEW%20BALLOT_IFES%20Report%20on%20GenAI_Feb2026.pdf.
- Buyl, Maarten, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, et al. "Large Language Models Reflect the Ideology of Their Creators." *Npj Artificial Intelligence* 2, no. 7 (January 7, 2026). https://doi.org/10.1038/s44387-025-00048-0.
- Coalition for Content Provenance and Authenticity (C2PA). "C2PA | Providing Origins of Media Content," n.d. https://c2pa.org/.
- Coalition for Content Provenance and Authenticity (C2PA). "C2PA FAQ," n.d. https://c2pa.org/faqs/.
- Cappello, Maja, ed. "News Media, Pluralism and Journalism in the Digital Age." *IRIS*. Strasbourg, Strasbourg, France: European Audiovisual Observatory, December 2025. https://rm.coe.int/iris-2025-news-sector-en/488029c71f.
- Carlini, Roberta, Anya Schiffrin, and Natalia Menéndez. "IPD Working Paper: How to Update EU and US Copyright Regimes in the Age of AI." *Initiative for Policy Dialogue*. Columbia University, January 12, 2026. https://ipdcolumbia.org/publication/ipd-working-paper-how-to-update-eu-and-us-copyright-regimes-in-the-age-of-ai/.

- Carmo, Ana. “AI And Anonymity Fuel Surge in Digital Violence Against Women.” UN News, November 20, 2025. https://news.un.org/en/story/2025/11/1166411.
- Carroll, Micah, Alan Chan, Henry Ashton, and David Krueger. *Characterizing Manipulation from AI Systems. EAAMO ’23: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, 2023. https://doi.org/10.1145/3617694.3623226.
- Creative Commons. “CC Signals - Creative Commons,” April 10, 2026. https://creativecommons.org/cc-signals/.
- Çetin, Raziye Buse, Natalia Stanusch, and Marc Faddoul. “From ‘Googling’ to ‘Asking ChatGPT’: Governing AI Search.” *AI Forensics*, November 2025. https://aiforensics.org/uploads/Governing_AI_Search.pdf.
- Chesney, Bobby, and Danielle Citron. “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.” *California Law Review* 107, no. 6 (December 2019). https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security.
- Chesney, Robert, and Danielle Citron. “Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy?” Lawfare, February 21, 2018. https://perma.cc/L6B5-DGNR.
- Copeland, Thomas. “AI-generated Iran War Videos Surge as Creators Use New Tech to Cash In.” BBC, March 7, 2026. https://www.bbc.com/news/articles/ckg8wvz427vo.
- Dennett, Daniel C. “The Problem With Counterfeit People.” *The Atlantic*, May 16, 2023. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/.
- AI Security Institute. “Do Chatbots Inform or Misinform Voters?,” September 30, 2025. https://www.aisi.gov.uk/blog/do-chatbots-inform-or-misinform-voters.
- El-Sayed, Seliem, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, et al. “A Mechanism-Based Approach to Mitigating Harms From Persuasive Generative AI.” arXiv.org, April 23, 2024. https://arxiv.org/abs/2404.15058.
- European Council. “Democratic resilience: Council approves conclusions on safeguarding electoral processes from foreign interference.” Press release, May 21, 2024. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/democratic-resilience-council-approves-conclusions-on-safeguarding-electoral-processes-from-foreign-interference/.
- Europol. “Facing Reality? Law Enforcement and the Challenge of Deepfakes, an Observatory Report from the Europol Innovation Lab.” Publications Office of the European Union, 2022. https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf.
- OpenMinds. “Fake Videos, Real Emotions: Viewers Believe AI-Generated Content Even When It’s Labeled,” February 9, 2026. https://www.openminds.ltd/reports/fake-videos-real-emotions-viewers-believe-ai-generated-content-even-when-its-labeled.
- Forrai, Michaela, Delia Cristina Balaban, and Desirée Schmuck. “Disclosures and literacy as determinants of AI-influencer recognition and well-being.” *Computers in Human Behavior* 182 (September 2026). https://doi.org/10.1016/j.chb.2026.108978.
- Gardhouse, Kathrin, and Toni Lorente. “Is ChatGPT a Search Engine and a Platform Under the EU Digital Services Act?” The Future Society, February 19, 2026. https://thefuturesociety.org/chatgpt_under_the_dsa/.
- Gerlich, Michael. “AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking.” *Societies* 15, no. 1 (January 3, 2025): 6. https://doi.org/10.3390/soc15010006.
- Gesley, Jenny. “Germany: Court Prohibits Memorization and Reproduction of Copyrighted Song Lyrics in AI Models.” Library of Congress, 2026. https://www.loc.gov/item/global-legal-monitor/2026-01-13/germany-court-prohibits-memorization-and-reproduction-of-copyrighted-song-lyrics-in-ai-models/.
- Goldin, Melissa. “State actors are behind much of the visual misinformation about the Iran war.” AP News, March 7, 2026. https://apnews.com/article/iran-war-images-misinformation-russia-israel-9e495017dc5c4bf24a0b6152863dbfb1.
- Goodson, Kayla, Jay Barchas-Lichtenstein, Samuel Jens, Emily Wright, and Utsav Gandhi. “Journalism’s New Frontier: An Analysis of Global AI Policy Proposals and Their Impacts on Journalism.” Center for News, Technology & Innovation, December 18, 2025. https://cnti.org/reports/journalisms-new-frontier-an-analysis-of-global-ai-policy-proposals-and-their-impacts-on-journalism/.

- G'Sell, Florence. "Regulating Under Uncertainty: Governance Options for Generative AI." Stanford Cyber Policy Centre, Freeman Spogli Institute, October 6, 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4918704.
- Haas, Julia, and Katharina Zügel, eds. "Safeguarding Media Freedom in the Age of Big Tech Platforms and AI." *OSCE*. Organization for Security and Co-operation in Europe, October 6, 2025. https://rfom.osce.org/representative-on-freedom-of-media/598525.
- Hackenburg, Kobi, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. "The levers of political persuasion with conversational artificial intelligence." *Science* 390, no. 6777 (December 4, 2025). https://doi.org/10.1126/science.aea3884.
- Helberger, Natali, Marilù Miotto, and Hannes Cools. "Understanding AI Transparency: What research says about labelling deepfakes and synthetic content." AlgoSoc, March 9, 2026. https://algosoc.org/results/understanding-ai-transparency.
- International Electrotechnical Commission (IEC), International Organization for Standardization (ISO), and International Telecommunication Union (ITU). "Technical Report on AI and Multimedia Standards: Mapping the Standardisation Landscape," 2025. https://s41721.pcdn.co/wp-content/uploads/2021/10/AMAS-Technical-Report-on-AI-and-Multimedia-Authenticity-Standards-Final.pdf.
- International Telecommunication Union. "AI And Multimedia Authenticity Standards Collaboration." AI For Good, n.d. https://aiforgood.itu.int/multimedia-authenticity/.
- Jensen, Mogens. "Copyright Enforcement in the Artificial Intelligence Environment." *Committee on Culture, Science, Education and Media*. Strasbourg: Council of Europe, April 1, 2026. https://pace.coe.int/en/files/35917/html.
- Japan Copyright Office (JCO), Legal Subcommittee under the Copyright Subdivision of the Cultural Council, and Agency for Cultural Affairs, Japan. "General Understanding on AI and Copyright in Japan," May 2024. https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf.
- Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. "Why Language Models Hallucinate." *arXiv*, September 4, 2025. https://doi.org/10.48550/arXiv.2509.04664.
- Klincewicz, Michał, Mark Alfano, and Amir Ebrahimi Fard. "Slopaganda: The Interaction Between Propaganda and Generative AI." *Filosofiska Notiser* 12, no. 1 (April 2025). https://arxiv.org/abs/2503.01560.
- Knott, Alistair, Dino Pedreschi, Toshiya Jitsuzumi, Susan Leavy, David Eyers, Tapabrata Chakraborti, Andrew Trotman, et al. "AI Content Detection in the Emerging Information Ecosystem: New Obligations for Media and Tech Companies." *Ethics and Information Technology* 26, no. 4 (September 21, 2024). https://doi.org/10.1007/s10676-024-09795-1.
- Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. "Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task." *arXiv.Org*, June 10, 2025. https://arxiv.org/abs/2506.08872.
- Chat GPT Is Eating the World. "Latest World Map of Copyright Suits V. AI Companies (April 5, 2026)," April 5, 2026. https://chatgptiseatingtheworld.com/2026/04/05/latest-world-map-of-copyright-suits-v-ai-companies-april-5-2026/.
- Mansell, Robin, Flavia Durach, Matthias Kettemann, Théophile Lenoir, Rob Procter, Gyan Tripathi, and Emily Tucker. "Information Ecosystems and Troubled Democracy: A Global Synthesis of the State of Knowledge on News Media, AI and Data Governance." *Observatory on Information and Democracy*. Forum on Information and Democracy, January 2025. https://observatory.informationdemocracy.org/wp-content/uploads/2025/06/rapport_forum_information_democracy_2025-1.pdf.
- Marwala, Tshilidzi. "The Concern Around Saying AI 'Hallucinates.'" Forbes Africa / United Nations University, February 4, 2026. https://unu.edu/article/concern-around-saying-ai-hallucinates.
- Milmo, Dan. "Apple Suspends AI-generated News Alert Service After BBC Complaint." *The Guardian*, January 17, 2025. https://www.theguardian.com/technology/2025/jan/17/apple-suspends-ai-generated-news-alert-service-after-bbc-complaint.
- "Apple Suspends AI-generated News Alert Service After BBC Complaint." *The Guardian*. January 17, 2025. https://www.theguardian.com/technology/2025/jan/17/apple-suspends-ai-generated-news-alert-service-after-bbc-complaint.

- MJ, Banias. “Inside CounterCloud: A Fully Autonomous AI Disinformation System.” The Debrief, August 16, 2023. https://thedebrief.org/countercloud-ai-disinformation/.
- Newman, Nic. “Journalism, Media, and Technology Trends and Predictions 2026.” *Reuters Institute for the Study of Journalism*. Reuters Institute for the Study of Journalism, January 12, 2026. https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2026.
- ———. “Overview and Key Findings of the 2025 Digital News Report.” Reuters Institute for the Study of Journalism, June 17, 2025. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/dnr-executive-summary.
- Powell, Roa, and Carsten Jung. “AI’s Got News for You: Can AI Improve Our Information Environment?” *The Institute for Public Policy Research (IPPR)*. The Institute for Public Policy Research (IPPR), January 30, 2026. https://www.ippr.org/articles/ais-got-news-for-you.
- PSG Consulting and Dewey Square Group. “AI Large Language Model Training: The Potential Risks of Ideological Skewing — PSG Consulting.” *PSG Consulting*, February 2026. https://www.psgconsulting.com/research-publications/potential-risks-of-ideological-skewing.
- Radcliffe, Damian. “Journalism in the AI Era: Opportunities and Challenges in the Global South and Emerging Economies.” *TRF Insights*. Thomson Reuters Foundation, January 2025. https://www.trust.org/wp-content/uploads/2025/01/TRF-Insights-Journalism-in-the-AI-Era.pdf.
- Reuel, Anka. “Chapter 3: Responsible AI.” In *The AI Index Annual Report 2024*, by Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, et al. Institute for Human-Centered AI, Stanford University, 2024. https://hai.stanford.edu/assets/files/hai_ai-index-report-2024_chapter3.pdf.
- Rizk, Joelle. “Why Is the ICRC Concerned by ‘Harmful Information’ in War?” Humanitarian Law & Policy Blog, September 10, 2024. https://blogs.icrc.org/law-and-policy/2024/09/10/why-is-the-icrc-concerned-by-harmful-information-in-war/.
- Federal Communications Commission. “ROBOCALL ENFORCEMENT NOTICE TO ALL U.S.-BASED VOICE SERVICE PROVIDERS,” February 6, 2024. https://docs.fcc.gov/public/attachments/DA-24-102A1.pdf.
- Rowland, Michelle, MP. “Albanese Government to ensure Australia is prepared for future copyright challenges emerging from AI.” Press release, October 26, 2025. https://ministers.ag.gov.au/media-centre/albanese-government-ensure-australia-prepared-future-copyright-challenges-emerging-ai-26-10-2025.
- Sadek, Dina, and Margot Fule-Hardy. “Cheap Tricks.” *Graphika*. Graphika, November 19, 2025. https://graphika.com/reports/cheap-tricks.
- OSCE. “Safeguarding Media Freedom in the Age of Big Tech Platforms and AI,” October 6, 2025. https://rfom.osce.org/representative-on-freedom-of-media/598525.
- Shapira, Natalie, Chris Wendler, Avery Yen, Gabriele Sarti, Koyena Pal, Olivia Floody, Adam Belfki, et al. “Agents of Chaos.” *arXiv*, February 23, 2026. https://arxiv.org/abs/2602.20021.
- Sherman, Natalie, and Imran Rahman-Jones. “Apple Suspends Error-strewn AI Generated News Alerts.” *BBC*. January 17, 2025. https://www.bbc.com/news/articles/cq5ggew08eyo.
- Si, Chenglei, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. “Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong.” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, January 1, 2024, 1459–74. https://doi.org/10.18653/v1/2024.naacl-long.81.
- Simon, Felix, Rasmus Kleis Nielsen, and Richard Fletcher. “Generative AI and news report 2025: How people think about AI’s role in journalism and society.” Reuters Institute for the Study of Journalism, October 7, 2025. https://reutersinstitute.politics.ox.ac.uk/generative-ai-and-news-report-2025-how-people-think-about-ais-role-journalism-and-society.
- Simsek, Can, and Ayse Gizem Yasar. “From Rejection to Regulation: Mapping the Landscape of AI Resistance.” Chair Digital Governance and Sovereignty, Sciences Po, May 2025. https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2025/05/compressed-Simsek-and-Yasar-AI-Resistance-Report-publication-ready-2.pdf.
- Stanusch, Natalia. “The Human Guide to Detecting AI Imagery.” AI Forensics, March 2026. https://aiforensics.org/uploads/GenAI_Human_Detection_Manual_v2.pdf.

- Stanusch, Natalia, Martin Degeling, Raziye Buse Çetin, Marcus Bösch, and Salvatore Romano. “Prompt, Upload, Repeat: How Agentic AI Accounts Flood TikTok with Harmful Content.” *AI Forensics*, December 3, 2025. https://aiforensics.org/uploads/Agentic_AI_Accounts.pdf.
- Stiglitz, Joseph, and Màxim Ventura-Bolet. “Information in the Age of AI: Challenges and Solutions.” *The Digitalist Papers*. The Digitalist Papers, n.d. https://www.digitalistpapers.com/vol2/stiglitzventurabolet.
- RSF. “USA: Google Is Claiming an Editorial Right It Does Not Have by Rewriting News Headlines in Its Search Results,” April 9, 2026. https://rsf.org/en/usa-google-claiming-editorial-right-it-does-not-have-rewriting-news-headlines-its-search-results.
- Voss, Axel. “Report on copyright and generative artificial intelligence – opportunities and challenges.” *European Parliament*. Committee on Legal Affairs, February 25, 2026. https://www.europarl.europa.eu/doceo/document/A-10-2026-0019_EN.html.

FID | FORUM INFORMATION DEMOCRACY

June 2026