

Forum: General Assembly 6 (GA6)

Issue: Legal Measures to Tackle Rumors, Misinformation, and Conspiracy Theories in Order to Protect Democracy

Student Officer: Dhruv Iswaran

Position: Deputy Chair

Introduction

The rapid propagation of rumours, disinformation, and conspiracy theories poses an existential threat to democratic politics, institutional trust, and the enforcement of international law. Deceptive narratives are used with effective digital communication technologies and modern AI systems to undermine democratic institutions, public discourse, and societal trust (WEF, 2024), which is a state in which propaganda and rumor-based strategies have existed historically, but has now evolved and been enhanced by new digital communication platforms and GenAI technologies. This modern state of affairs has created new avenues to disseminate content faster, to much wider and more specific public bodies than ever possible before (UNESCO 2023). Addressing information manipulation has posed a human rights dilemma regarding the spread of good, bad, true, and false information from a legal perspective.

On one hand, there is a risk of restricting freedom of opinion and expression that is protected under Article 19 of the ICCPR (UN General Assembly, 1966). On the other hand, failure to regulate online information will allow states and non-state actors to engage in coordinated inauthentic behavior, manipulate election processes, and undermine national security. On an international level, weaponised disinformation can now be used as part of asymmetric hybrid warfare that will question traditional principles of sovereignty and non-intervention that were entrenched in customary international law(United Nations, 2021).

Definition of Key Terms

Disinformation

False or misleading information that is created, disseminated, and circulated to deceive, distort, or manipulate the information environment for the benefit of the attacker (UNESCO, 2021).

Misinformation

False information shared by individuals who believe it to be true but was not created with malicious intent (WEF, 2024). Is used in some nations by ill actors to sway public opinion on the current government regime/stance.

Rumour

An unverified piece of information distributed among a large group of people whose accuracy cannot be verified independently (DiFonzo & Prashant, 2007). Can be unintentionally or intentionally spread.

Conspiracy Theory

An informational narrative device about a major political and/or social event(s) taking place as a result of a secret, clandestine, and highly influential coalition of powerful people and organisations (Council of Europe, 2021).

Synthetic Media (Deepfakes)

Digitally manipulated or synthetic visual, audio, or multimodal content created with the aim of deliberately generating false representations of real people or events by using AI algorithms (EU, 2024). It is used by some to make government officials look bad in the public eye.

Coordinated inauthentic Behaviour (CIB)

Organised efforts by networks of automated systems, sockpuppets, and fake accounts to artificially amplify content and/or to distort the online discourse to influence the political or social debate.

Background Information

Technological Amplifiers and Media Transformation

As changes are taking place to convert editorial media to digital media, increasingly, companies are following an engagement business model with recommendation algorithms actively designed to optimise outrage and polarization to keep users online, where falsehoods reach audiences 6 times faster than true news (Meyers, 2022). Compounding this issue, generative AI makes the production of convincing audio, visual, and text content inexpensive and accessible, leading to an explosion of synthetic disinformation.

Hyper-realistic deepfake audio and visual content are used for candidate impersonation and deception and undermine the general public's ability to fact-check (EU, 2024). Furthermore, coordinated inauthentic behaviour (CIB) exploits organised bots, trolls, and front groups to automate networks with online adversaries to generate fake consensus, manipulate trending topics, and undermine democratic processes (Reuters Institute, 2024).

PROPORTION THAT SAY THEY TRUST NEWS FROM EACH SOURCE – ALL MARKETS

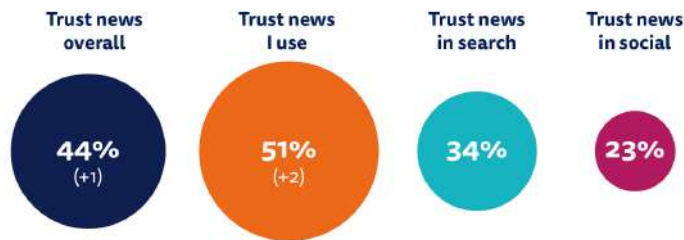


Figure 1: Reuters Institute data on global trust in news via varying outlets.

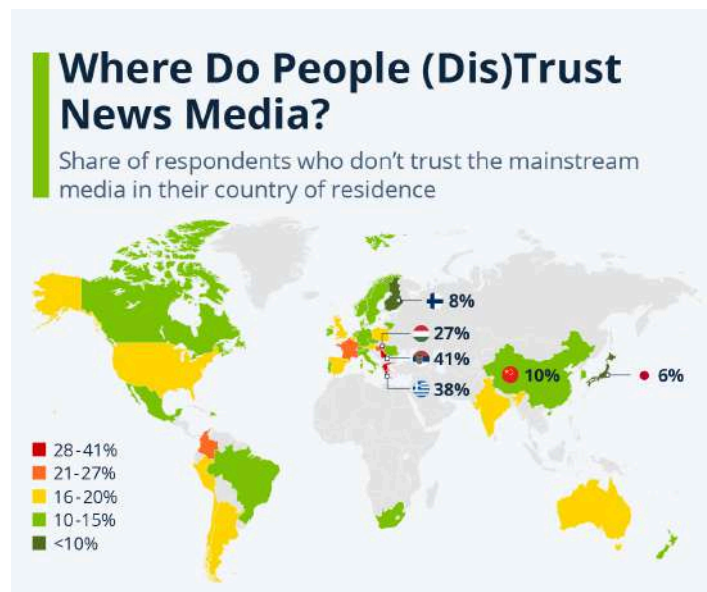


Figure 2: Statista map on national populations' trust in their mainstream news, compared across the globe.

Major Recent Developments

In recent global election seasons, many false operations generated by state-sponsored entities used generative AI models to impersonate candidates, create scandalous audio, and disenfranchise voters. At the same time, campaigns of foreign information manipulation and interference (FIMI) against transnational partnerships expanded, leading international judicial and legislative institutions to search for obligatory means of regulation to uphold electoral integrity.

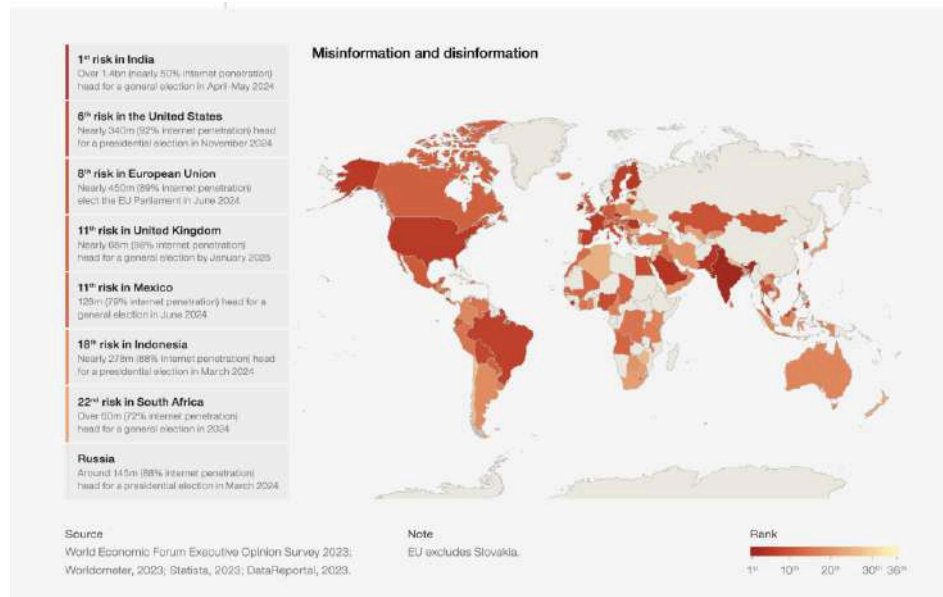


Figure 3: AI's Role in General Elections Globally, as well as its prominence on the internet of the respective country, from GNET.

Prominent Claims

Foreign and domestic disinformation ranks as the most significant short-term threat to the world, even above extreme weather events and armed conflict in the World Economic Forum Risks Report 2024 (WEF, 2024). A survey of public opinion in 15 democracies by Ipsos and Statista revealed that a staggering 85% of citizens in these countries are very or extremely worried about the consequences of online disinformation on upcoming elections (Statista, 2024). Over 65% of respondents reported they regularly encountered online hate speech and false information on major platforms (Statista, 2024).



Figure 4: Statista infographic detailing the national impact of misinformation and disinformation posing the biggest threat in the sociopolitical scheme, compared across the globe.

Global Regulatory Models and Legal Frameworks

By adopting a co-regulatory platform accountability model, the EU DSA can impose mandatory systemic risk assessments, independent algorithmic audits and a transparency

report on content moderation and ad targeting to VLOPs (EU, 2022). In contrast, through constitutional speech protection, US anti-disinformation rules only allow moderation by voluntary platforms and data-sharing imposed from above; they are limited by the First Amendment's prohibition on suppressing speech by targeting "content neutrality" on the platform unless the speech incites imminent lawless action (Meyers, 2022). The government can impose immediate correction or take-down commands for online disinformation under Singapore's POFMA (Republic of Singapore, 2019), and messaging traceability and mitigation of content go together in India's IT Rules (Republic of India, 2023). Lastly, platform liability and transparency require a liability rule for amplified payment and a transparency mandate for the automation of individual messaging accounts, from Brazil's Facebook Law 2630/2020 ("Fake News" bill) (Federative Republic of Brazil, 2020).



Figure 5: Map depicting prominent regional alliances across the globe, and their member countries, from *Politics for India*.

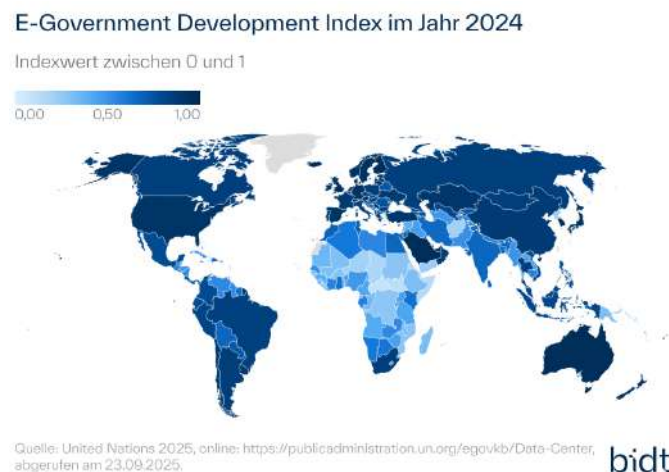


Figure 6: Map ranking each country on the global E-Government Development Index(EGDI), from the *Bavarian Research Institute for Digital Transformation (BIDT)*.

Past Resolution Attempts

UN (Resolutions and Treaties)

- ICCPR (1996): Articles 19 & 20 define international standards of free expression such that restrictions should only be made in cases of absolute necessity for the

protection of the rights of others, the security of the State, or the prevention of hate speech (UN General Assembly, 1966).

- UNGA Resolution 76/227 (2021): Countering disinformation for the promotion and protection of human rights and fundamental freedoms: Calls on Member States to counter disinformation while ensuring the respect and promotion of the obligations of States under international human rights law (UN General Assembly, 2021).

UN Agencies (Programmes)

- UNESCO Policy Guidelines for the Governance of Digital Platforms (2023): Sets a comprehensive set of international standards for the protection of freedom of expression and the regulation of digital platforms (UNESCO, 2023).
- UNDP iVerify System (2021): An automated electoral verification system developed by the United Nations Development Programme for use in EDCs to respond to electoral disinformation (UNDP, 2021).
- ITU (International Telecommunication Union): They are key to technical standardisation and digital infrastructure security. ITU is working with standards organisations across the world to develop the technical standards for synthetic media, cryptographic watermarking, and verification of AI content.

Organisations and Major Countries Involved

EU (European Union)

The EU leads the way on binding digital regulation with its Digital Services Act and AI Act, prioritising responsibility while heavily fining tech giants (EU, 2022).

EU Code of Practice on Disinformation (2022): Co-regulatory code of conduct through which participating tech companies abide by self-governing principles over ad monetisation, political ad transparency, and data access (EU, 2022).

Council of Europe Budapest Convention (2001/2023): Provides for international legal cooperation to research and prosecute cyber-enabled deceptive tactics and foreign information (Council of Europe, 2023).

USA (United States of America)

The US, where many of the world's largest tech companies reside, has a less interventionist approach to regulating legal content due to First Amendment safeguards, relying instead on moderation from platforms (Meyers, 2022).

Singapore

In Singapore, executive intervention levels legislation with the introduction of POFMA, the Ability to act quickly and correct false statements online, although international observers fear a rise in executive power (Republic of Singapore, 2019).

India

As the world's largest democracy, India ranks #1 globally for executive content removal, message traceability on encrypted platforms, and state-led fact-checking of viral rumours (Republic of India, 2023).

Brazil

Brazil fights disinformation ahead of elections through active law enforcement by its Supreme Federal Court and legislation currently before the legislature for platform accountability (Federative Republic of Brazil, 2020).

UNESCO (United Nations Educational, Scientific and Cultural Organisation)

UNESCO has been leading the UN's fight for freedom of speech through a set of global platform governance principles by supporting Member States to increase digital media literacy (UNESCO, 2023).

Possible Solutions

Mitigating the spread of damaging information

Platform Algorithms and Accountability: Member States should legislatively mandate that platforms proactively implement independent systemic risk analysis (e.g., annually) and recalibrate optimisation recommendation devices that systematically promote outrage, conspiracy theories, and/or incite unrest.

Establishing rules for content creation through AI:

Mandatory Synthetic Media Watermarking and Authentication Standards: Governments should enshrine a legislative code of practice for cryptographic data watermarks (e.g., C2PA) to be added to generative media by developers before dissemination.

Securing Election Processes:

Independent Electoral Verification Coalitions: National election bodies in agreement with UNDP and civil society, should roll out automated verification systems (e.g., iVerify) and prevent states from monopolising fact-checking to avoid politicisation.

Strengthening Public Knowledge:

All-Hazards Digital Media and Information Literacy Curriculum: Governments should include digital Media and Information Literacy (MIL) that is aligned with UNESCO's agenda in the national education system to enable their citizens to participate in participatory discourse (e.g., active participation) and counter-narrative manipulation (e.g. D.A.R.V.O.).

Bibliography

Council of Europe. "Combating Disinformation and Conspiracy Theories." Council of

Europe Portal, 2021, <https://www.coe.int/en/web/freedom-expression/disinformation>.

DiFonzo, Nicholas, and Prashant Bordia. Rumor Psychology: Social and Organizational Approaches. American Psychological Association, 2007.

European Union. "Regulation (EU) 2022/2065 on a Single Market For Digital Services (Digital Services Act)." Official Journal of the European Union, 2022, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2065>.

Federative Republic of Brazil. "Draft Law No. 2630/2020 (Brazilian Internet Freedom Law)." National Congress of Brazil, 2020, <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2256735>.

Meyers, Zach. "Will the Digital Services Act Save Europe from Disinformation?" Centre for European Reform, 21 Apr. 2022, <https://www.cer.eu/insights/will-digital-services-act-save-europe-disinformation>.

Republic of India. "Information Technology Amendment Rules." Ministry of Electronics and IT, 2023, <https://www.meity.gov.in/content/information-technology-amendment-rules-2023>.

Republic of Singapore. "Prevention from Online Falsehoods and Manipulation Act 2019 (POFMA)." Singapore Statutes Online, 2019, <https://sso.agc.gov.sg/Act/POFMA2019>.

Reuters Institute for the Study of Journalism. "Digital News Report 2024." Oxford University, 2024, <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024>.

Statista. "Where False Information Is Posing the Biggest Threat." Statista Research Department, 19 Jan. 2024, <https://www.statista.com/chart/31605/rank-of-misinformation-disinformation-among-selected-countries/>.

Statista. "2 in 3 People Often Encounter Hate Speech Online." Statista Research Department, 21 Oct. 2024, <https://www.statista.com/chart/33299/online-hate-speech-encounters/>.

United Nations Development Programme. "iVerify: Automated Fact-Checking Platform." UNDP Digital Centre, 2021, <https://www.undp.org/digital/verify>.

United Nations Educational, Scientific and Cultural Organization. "Guidelines for the Governance of Digital Platforms." UNESCO Publishing, 2023,

<https://www.unesco.org/en/internet-trust/guidelines>.

United Nations General Assembly. "Countering Disinformation (A/RES/76/227)." UN Official Document System, 24 Dec. 2021, <https://undocs.org/A/RES/76/227>.

United Nations General Assembly. "International Covenant on Civil and Political Rights." UN Treaty Series, vol. 999, 16 Dec. 1966, p. 171, <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>.

World Economic Forum. "The Global Risks Report 2024." WEF Geneva, 10 Jan. 2024, <https://www.wef.org/reports/global-risks-report-2024>.

Zandt, Florian. "Infographic: Where Do People (Dis)Trust News Media?" Statista Daily Data, 28 June 2024, <https://www.statista.com/chart/32510/share-of-respondents-who-dont-trust-the-mainstream-media-in-their-country-of-residence>.

10] Regionalization of World Politics - Politics for India. 9 June 2025, <https://politicsforindia.com/10-regionalization-of-world-politics-psir/>.

Bayerisches Forschungsinstitut für Digitale Transformation. "E-Government Development Index | body." bidt DE, <https://www.bidt.digital/zahlenmonitor/e-government-development-index/>. Accessed 15 Sept. 2026.

Trust, Misinformation, and the Declining Use of Social Media for News: Digital News Report 2018 | Reuters Institute for the Study of Journalism. <https://reutersinstitute.politics.ox.ac.uk/news/trust-misinformation-and-declining-use-social-media-news-digital-news-report-2018>.

Anadi. "Deep Fakes, Deeper Impacts: AI's Role in the 2024 Indian General Election and Beyond." GNET, 11 Sept. 2024, <https://gnet-research.org/2024/09/11/deep-fakes-deeper-impacts-ais-role-in-the-2024-indian-general-election-and-beyond/>.

Appendix

I. OHCHR Rabat Plan of Action (6-Part Threshold Test for Incitement)

A. This introductory section of the plan elaborates on the six-part threshold test

that was developed by the Office of the United Nations High Commissioner for Human Rights (OHCHR) to help distinguish speech that is protected expression (permissible under ICCPR Article 19) from unlawful incitement or hate speech (permissible under ICCPR Article 20)

1. 6 Threshold Factors:

- a) Context: Economic, social, and political context of the speech/act
- b) Speaker: Role, authority, and functionary status of the speaker disseminating the content.
- c) Intention: Clear intention by the speaker to promote hatred, incite violence, or cause harm.
- d) Content & Delivery: Style, tone, argumentation, and degree of provocation.
- e) Scope & Audience: Reach, dissemination channels, public accessibility, and frequency of the speech act.
- f) Likelihood & Imminence: Objectively assessed probability of actual harm resulting from the speech/act

B. Resource Link: [OHCHR Rabat Plan of Action Guidelines](#)

II. Content Provenance and Authenticity (C2PA) Technical Framework

Technical Mechanism	Function & Implementation	Vulnerability & Limitations
Signed Metadata	Hidden cryptographic manifest embedded into files; records software used for creation, edit history of such files, and author signatures	Is easily stripped away when re-encoded or uploaded to social media platforms
Invisible Watermarking	Pixel/audio level noise alterations applied to AI-generated single/multimodal content(e.g., Google SynthID, Anthropic Text	Requires specialised algorithms to detect it; decays after severe/heavy file compression; can be disabled with tailored

	Watermarking)	algorithms(e.g., Claude AI skills to remove text watermarks)
Perceptual Fingerprinting	Content hashing reverse-searched against central media repositories	Matching is probabilistic; it can't be fully trusted without human review/oversight/input

III. The Santa Clara Principles 2.0 on Content Moderation

A. A set of human rights standards for platform accountability and algorithmic transparency.

1. Core Principles for Draft Resolutions:

- a) Human Rights & Due Process: Requiring human review of automation-based account suspensions or content demonetisation.
- b) Transparency of State Involvement: requiring companies to publish all formal state requests for the takedown and/or censorship of user data.
- c) Cultural & Linguistic Competence: Requiring systems of content moderation to reflect regional linguistic diversity and nuanced local content.
- d) Algorithmic Explainability: Requiring independent third-party audits of recommendation engine algorithms.

B. Resource Link: [Santa Clara Principles 2.0 Official Framework](#)

IV. International Fact-Checking Network (IFCN) 5-Point Code of Principles

A. An accreditation checklist for non-partisan fact-checking upheld by the Poynter Institute

1. The 5 Fundamentals;

- a) Nonpartisanship & Fairness: Applying a uniform standard to all politicians and claims.
- b) Transparency of Sources: Naming all research sources so that others can independently verify the research.
- c) Transparency of Funding & Organization: Disclosing all corporate, NGO, and government-backed financial support.

- d) Transparency of Methodology: Developing clear standards for claim selection, research methodology, and publication.
- e) Open & Honest Corrections Policy: Following up with a correction notice when errors are found.

B. Resource Link: [IFCN Code of Principles](#)

V. Systematic Risk Assessment Standard (EU DSA Articles 34 & 35)

A. A regulatory document that calls for annual risk evaluations of online platforms with more than 45 million active monthly users.

1. 4 Required Risk Assessment Areas:

- a) Illegal Content: Facilitation of illegal material, hate speech, and illegal market trading.
- b) Fundamental Rights: Adverse effects on free expression, privacy, and equality.
- c) Civic & electoral Integrity: Strategic platform interference with public discourse, political elections, and sovereignty.
- d) Public Health & Gender-Based Violence: Coordinated disinformation campaigns related to health and targeted cyber harassment campaigns.

B. Resource Link: [EU Digital Services Act \(DSA\) Regulation Text](#)

VI. UNESCO Guidelines for the Governance of Digital Platforms

A. A global operational model to realise human rights protection and regulate the digital communications ecosystems.

1. 5 Principles for Effective Regulation:

- a) Independent Co-regulatory Bodies: Establishment of global multi-stakeholder non-governmental oversight bodies
- b) Human Rights Impact Assessments: Conduct testing before implementation of algorithmic curation changes.
- c) Media & Information Literacy (MIL): Making digital literacy classes compulsory for secondary school students nationwide.
- d) Data Sharing for Researchers: Data sharing initiatives to give independent researchers access to anonymized platform data for studying disinformation campaigns.
- e) Effective Remedies: Time-bound appeal procedures for wrongful moderation of accounts and targeted victims of online violence

B. Resource Link: [UNESCO Guidelines for Digital Platform Governance](#)