G20 BRASIL 2024
BUILDING A JUST WORLD AND A SUSTAINABLE PLANET

Mapping the Information Integrity Debate and Informing the Agenda of the G20

Integrity of Information and Trust in the Digital Economy

Digital Economy Working Group

Report prepared under the Brazilian Presidency of the G20

Knowledge partner

unesco

Mapping the Information Integrity Debate and Informing the Agenda of the G20

Integrity of Information and Trust in the Digital Economy

Digital Economy Working Group

Report prepared under the Brazilian Presidency of the G20

SECRETARIA DE COMUNICAÇÃO SOCIAL
GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO

unesco

September 2024

UNESCO is a Knowledge Partner to the Brazilian G20 Presidency to the Digital Economy Working Group, including on artificial intelligence and the integrity of information. This paper focuses on mapping the debate on information integrity and informing the agenda of the G20.

UNESCO is a leader in promoting the safeguarding of information as a public good, as recognized in the Windhoek +30 Declaration. The Government of Brazil has recognized UNESCO's added value in responding to misinformation and disinformation, as well as the relevance of UNESCO's Internet for Trust Initiative.

Acknowledgments

About this brief:

This policy brief was commissioned and coordinated by UNESCO and written by Research ICT Africa (RIA), an African digital policy, regulation and governance think tank based in Cape Town, South Africa. The bibliographic references for this work may be found via the hyperlinks in the electronic version of this document.

Authors:

Principal investigator: Dr Alison Gillwald, Executive Director, RIA (Adjunct Professor University of Cape Town)

Lead researcher: Prof Guy Berger, Distinguished Fellow, RIA (Emeritus Professor Rhodes University)

Researcher: Elizabeth Orembo, Researcher, RIA

UNESCO's team:

Tawfik Jelassi: Assistant Director-General for Communication and Information

Sylvie Coudray: Director for Freedom of Expression, Media development and Media and Information Literacy (CI/FMD)) and Secretary of the International Programme for the Development of Communication (IPDC)

Guilherme Canela: Chief of Section, Freedom of Expression and Safety of Journalists

Ana Cristina Ruelas: Senior Programme Specialist

Lucas Ferreira Novaes: Consultant

Olga Bednarek: Intern

Brazil – G20 Presidency:

The report was prepared in coordination with the following members of the Secretary of Digital Policies, at the Social Communication Secretariat at the Brazilian Presidency: João Brant, Samara Castro, Ricardo Horta, Sandro Eli, Marina Pita, Giovana Tiziani and José Renato Pereira (UNESCO consultant), under the administration of the Ministers Paulo Pimenta (until May 2024) and Laércio Portela (between May and September 2024).

Key findings from the mapping

- A vibrant and inclusive digital economy and society is highly dependent on the quality and flow of information from a diverse range of sources. The integrity of this information is needed for social and economic development at the national level, and is essential to the Sustainable Development Goals. As such, equitable and universal access to information, and ensuring its integrity, are central principles of public interest governance.
- The increasingly globalised and dynamic nature of the digital economy is characterized by a high concentration of social network platforms, AI companies and search services and the advanced data-driven technologies used by them. The integrity of the volumes of information generated and distributed online calls out for international cooperation.
- Information integrity in the digital realm faces a proliferation of challenges to its accuracy, reliability, and trustworthiness. The resulting harms and human rights abuses run directly counter to sustainable social, political and economic development.
- There are rapid increases in the speed and scale of threats to information integrity. This reflects the confluence between the falling cost of producing digital content, which is accelerated by Generative Artificial Intelligence (AI), and the distribution of rising volumes of content as afforded by social media and search services. Challenges in effective content moderation on the one hand, and the sustainability of news media on the other, are intimately bound up with the rise of risks to information integrity online.
- The benefits of digitally extending "voice," diversity, and greater citizen involvement in the wider communications landscape may be fundamentally derailed unless the growing harms associated with being, or coming, online are mitigated.
- There is recognition internationally that gaps in human rights-based governance and shortfalls in self-regulation need to be addressed. This provides an opportunity to better align platforms and AI content-related services with the imperative of information integrity.

- Human rights and fundamental freedoms, including freedom of expression, provide the foundation for governing online content in the public interest and developing safeguards for children, women and others whose rights are disrespected online.
- Revisiting governance for information integrity implicates different policies. These range from transparency of digital services and the fostering of media and information literacy, through to assuring news media's viability. Regulatory frameworks are increasingly taking account that the information ecosystem cuts across the spheres of governing media, social media, search and AI. There is also growing acknowledgement that, within the plurality of online content which meets integrity criteria, there is special value to be attached to the kind of information which counts as a public good. This includes information availed by public authorities and that produced through the professional reporting of news.
- The state of research shows there is a need to build governance capacity, increase access to data and strengthen knowledge production in order to gauge and address systemic risks. Without an evidence-base, it is complex to design digital policies and also to assess the impact of governance reforms affecting the information space.
- G20 members may wish to consider locating national strategies for information integrity within the wider landscape of digital governance debates and actors. This also entails considering the global character of the challenges of the "digitalisation", "datafication" and "platformisation" of communications. A range of possible strategies for promoting information integrity are elaborated in a companion document.

Table of contents

Figure

Table

Introduction

Freedom of expression is the bedrock of the marketplace of ideas and the circulation of information, which are in turn vital for transparency and societal progress. These are also regarded as essential for optimum outcomes in economic markets, and in science and socio-political life broadly. When the information marketplace malfunctions, buyers and sellers lose choice and become vulnerable to scams and other crimes. The online space becomes untrustworthy and unsafe, while citizens' health, environmental and democratic rights become casualties. There is a grave imbalance between the volume and virality of corrupted content on the one hand, and information that counts as a public good on the other.

So important is information to the functioning of economic and public life that historically there has been extensive governance in this realm, such as to promote universal access and to address information asymmetries which impede firms and citizens' effective participation in economy and society. Governance to this end has entailed mixes of regulation, co-regulation and self-regulation.

In contemporary society, policy-makers are increasingly seized with the access to, and the quality of, information accompanying the digitisation and platformisation of the content marketplace. Disinformation and misinformation are not new to humanity, but are reaching levels that both overshadow and undermine the flow of reliable and accurate information. This is due to the velocity and volumes at which digital content is now being generated and disseminated. Hence the increasing embrace of the concept of "information integrity" to signal a value of key policy importance for the digital economy (See "Definitions and Frameworks" below for more on this concept).

Research shows that information, if governed in the public interest and aligned to international human rights law, is a vital ingredient for protecting and promoting fundamental freedoms, scientific innovation and sustainable development, reducing poverty and hunger, and combatting corruption. Digitalisation increasingly cuts across all sectors of the economy and society, driving efficiencies, reducing costs and improving productivity through improved information flows. In this dynamism, however, there are two key conditions for progress: firstly, combating threats to information integrity, and secondly, advancing those types of data, information and knowledge which can count as a public good. In this way, information can play its role in nurturing effective and inclusive markets, fostering openness and spurring innovation.

The COVID-19 pandemic highlighted the significance of information integrity in managing an existential public health crisis. This is also increasingly evident in regard to the pressing need to mitigate climate change. Monitoring and countering climate disinformation is essential to be able to strengthen action against climate change such as at the COP30 in 2025 and the media's role therein, and in follow up on the focus of the 3 May 2024 World Press Freedom Day and the Santiago+30 Declaration. Digital inclusion and gender equality targets are put at risk by gendered disinformation and online gender-based violence such as trolling, stalking and doxing. In what is called the 2024 Year of Elections, information integrity is more central than ever to electoral integrity. The erosion of integrity of information is thus a major challenge for making (and assessing) progress towards the Sustainable Development Goals.

Governance for information integrity cannot be considered outside the political, cultural and social context in which it takes shape – therefore a whole-of--society approach is needed. Particularly significant are the globalising but highly uneven trends of digitalisation, datafication and platformisation and the imbalances in opportunities for people in most parts of the world in regard to these developments. Some of the main considerations in contextualising information integrity online include:

- The fact that more than half the world's people do not have meaningful Internet access, and for those online there are low levels of media and information literacy, including AI and data literacy. Many people appear to lack adequate critical skills to understand the online content environment, and there are operational and conceptual challenges in empirically measuring the deficit in knowledge and capacities.
- Digital inclusion depends on meaningful connectivity as well as the range of literacies needed to navigate digital transformations. The problems pertaining to online information integrity can severely impact societies at large, having at least indirect impacts on the 2.7 billion people who do not have meaningful and affordable access to digital communications. The same applies to large swathes of people disadvantaged by internet shutdowns and end-user social networking taxes, as well as high connectivity costs.
- The advantages of being connected online are inhibited by a range of deep divides in capacities, innovation and data, and by inadequate safety guardrails in available services.

- Online mis- and disinformation and hate speech are being exacerbated by the advanced data-driven technologies associated with business models reliant on data extraction and attention economics. Combined with market concentration, cuts in staff working on self-regulatory "trust and safety" compliance, an expansion of bots in the digital ecosystem, and the rise of AI-generated content, the result is an increased vulnerability for information integrity. There is increased velocity, volume and virality of content on social media and search, but a reduction in its verisimilitude. All this makes it harder to adapt digital opportunities to foster fundamental freedoms and sustainable development.

- Early assessment of the impact of massively increased output of content resulting from Generative AI, portends an intensified risk to information integrity. This goes beyond malicious deployments, to further cover the problems of generated content that is often inaccurate and which content also frequently rests upon narrow values and biased data that reflect underlying discriminatory practices.

- With the potential to undermine human rights and fundamental freedoms, disaster relief, public health management and even human agency, the loss of information integrity often harms the people – both online and offline – who are least able to mitigate the risks.

These challenges reflect what the UN Secretary General calls "a massive governance gap," that perpetuates digital inequality and data injustice, and further damages trust in the digital economy and society more broadly. Such governance is relevant to the *production*, *dissemination*, and *consumption* of content, and to users' rights in the process. At each of these three stages in the chain, measures can be taken serve to foster information integrity parameters aligned to fundamental freedoms and sustainable development. The overall outcome of renewed governance initiatives would be to mitigate threats to the information ecosystem, while at the same time strengthening those parts of it which do contribute information with integrity to the whole. This goal further entails efforts that empower users and strengthen their protection through ensuring effective reporting mechanisms and redress when suffering online harm. It calls out for the implementation of human-rights-based standards that incorporate safety by design throughout digital technology and digital services. These measures are vital for the digital economy, but also have wider societal benefit.

The questions arising from contemporary challenges to information integrity are:

- What are the main digital trends that are currently compromising information integrity and how are they doing so?
- What actors and governance mechanisms exist to counter the harms associated with these trends, and to advance information as a public good in their place?
- What are the implications for information integrity when considerable numbers of people are unable to participate in the increasingly digital public sphere, but are nevertheless significantly affected by what happens online?
- How can international co-operation help individual jurisdictions to align their interventions for information integrity to human rights law and enhance their effectiveness?

This report maps the landscape of problems in the current period, and the underlying causes. It examines how governance stakeholders are responding, and it charts the range of actors and initiatives. Finally, it signals lines of action which can inform possible solutions to address these challenges, as are elaborated in a companion report to this document.

Definitions and frameworks

Information Integrity

The UN Secretary General's policy brief 8 elaborates that information integrity "is threatened by disinformation, misinformation and hate speech". This is helpful in understanding what information integrity is not, even although it does not venture into conceptualising what its essence actually is beyond being the opposite of the threatening content.

For the purpose of this report, understanding the elements in opposition to "information integrity" draws upon a UNESCO-ITU Broadband Commission report. That study highlights the communications that are false or misleading, the degree to which people are aware of this and the corresponding motives behind such content (intentionally deceptive content as disinformation, unintentional as misinformation). The same study also points to deployment of messages for advocacy to incite violence, hostility or discrimination, as constituting hate speech. When information exists in these forms, it can be considered to be in contradiction with "information integrity".

Going further to designate the positive meaning of "information integrity", a starting point is UNDP's observation that the concept is borrowed from corporate systems, where it refers to information security and data protection within enterprises. Applied more broadly, information integrity is determined by "the accuracy, consistency, and reliability of the information content, processes and systems to maintain a healthy information ecosystem." Building upon this base are the UN Global Principles for Information Integrity. These propose that the integrity of the information ecosystem is "where freedom of expression is fully enjoyed and where accurate, reliable information, free from discrimination and hate, is available to all in an open, inclusive, safe and secure information environment". The Principles add: "Information integrity entails a pluralistic information space that champions human rights, peaceful societies and a sustainable future. It holds within it the promise of a digital age that fosters trust, knowledge and individual choice for all".

The OECD's Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity highlights the importance of taking a comprehensive approach, tailored to country contexts, that emphasises the need to create an environment for reliable information to thrive. So too does the Global Declaration on Information Integrity Online, adding that the concept offers a positive vision of a broader information ecosystem that respects human rights and supports open, safe, secure, prosperous and democratic societies.

Relevant to elaborating on information integrity as a concept are UNESCO's principles for an information ecosystem that can address the challenges of mis- and disinformation and hate speech. Its 41st General Conference endorsed the principles of the Windhoek+30 Declaration in November 2021, following a multistakeholder process that began at the global celebration of World Press Freedom Day in May of that year. The UNESCO Declaration advocates for information as a public good in the context of freedom of expression. It sets out three steps to guarantee information as a shared resource for humanity: the transparency of digital platforms, citizens being empowered through media and information literacy, and media viability being secured. Later, also through a multistakeholder process initiated by UNESCO, Guidelines on the Governance of Digital Platforms were elaborated for a coherent and comprehensive rights-based approach that advances information integrity as regards the major distribution channels of content.

It has been observed that the meaning of integrity in regard to accuracy and reliability varies from context to context and that realities on the ground are particularly affected by who controls the production, distribution and consumption of information. Further, the case is made that attention is also needed to fidelity, safety, and transparency. Academics have flagged some limits related to the origins of the concept in information security and argued that in a wider information environment, the consistency criterion has to be understood as an issue of access to information.

All this suggests that "Information Integrity" merits further elaboration in terms of definition, especially in situations where it may be given legal weight. At the same time, it is evident that the concept, understood broadly, highlights key information parameters to be valued and, conversely, those which run counter to them, and which can be treated as obstacles to human rights and sustainable development.

While states will develop varying approaches to information integrity, all are expected to align with obligations under the International Covenant on Civil and Political Rights (ICCPR) which describes freedom of expression as entailing the freedom of individuals to seek, receive and impart information, and provides for limited restrictions only under the conditions of legality, proportionality and legitimate purpose.

Considering information in economic terms

Information may be primarily public or private in its availability, but it can be of economic value in either case. Amongst publically-available information, is that content which is offered through systems like free-to-air broadcasting, the public-domain publishing of research outputs, and that availed by public authorities. Private information includes that available only through paywalls and subscriptions, as well as that content which is justifiably confidential to individuals and/or institutions. Longstanding mechanisms that complement the private provision and circulation of information, are public service broadcast services with elaborated mandates (such as multi-lingualism), "must carry" provisions for private broadcast channels, and media diversity funds. In the digital era, there are also strategies to ensure universal access to the Internet as a central element of the contemporary public sphere.

Information counts as a public good when it meets two conditions. First, it is "non-excludable," meaning that in its provision no one can be excluded (as in the case of public broadcasting or parliamentary records). Second, it is "non-rivalrous," meaning enjoyed repeatedly by anyone without diminishing the benefit to others. Information, and data, can be a pure public good, or be classified as "merit goods". This latter means the consumption of information is not a zero-sum game, even if (for example), commercial barriers exclude some actors. Both public and merit information goods have clear benefit to societies and economies, but such gains will only accrue if information has integrity.

Contemporary challenges to information integrity and their causes

Research commissioned by UNESCO in 2022 identified interrelated challenges linked to social media platforms and search services. It noted how threats, such as health dis- and misinformation, can incur major costs and cancel out the various benefits of digital communications. These problems not only persist in 2024, but appear to be worsening. Current trends include harms to: women, children, people with disabilities and vulnerable individuals, election integrity, the fight against climate change and its disruptive impact on society and the economy, the financial viability of news media, and the accountability of powerful entities. These trends, including the phenomenon of racism online, are detailed in **Appendix A**. It is evident that the situation is being exacerbated by the rise of Generative AI which super-charges the production and targeting of content that is potentially harmful to human rights and sustainable development. Current configurations of this recent technology are also threatening the prospects for information integrity by means of social, linguistic and knowledge biases, along with false content or made-up content ("hallucinations") and opaque systems that exploitatively scrape copyrighted online content sources.

Driving the proliferation of challenges to information integrity are six major factors outlined below. It follows from this outline that a corresponding set of governance measures – spanning regulation, co-regulation and self-regulation systems – would be needed to comprehensively address the full package of determinants. As elaborated (with references) in **Appendix B**, the six factors are:

- The **business model** of platforms, with associated AI recommender systems, continue to be based on what the United Nations Economist Network describes as the attention economy. Based on data-mining, this particular logic feeds the circulation of potentially harmful content, including micro-targeted and AI-generated content and advertising.
- **Automated advertising**: Opaque and centralised advertising exchanges, dominated by platform companies, continue to drive online advertising spending and placement, channelling revenues to fraudulent content as well as falsehoods about health and climate.

- **Manipulation**: The abuse of platforms for disinformation and hate speech shows no sign of diminishing, and there is evidence of AI-generated tools being used for these purposes. This is despite high-profile calls for stakeholders to refrain from doing so.
- **Spending priorities**: Platform failures in moderating content and dealing with user complaints and appeals continue at scale, reflecting insufficient investment in these areas. In the absence of legally binding safety standards in most jurisdictions, major platforms have reduced their human capacity in "trust and safety" even while automation remains inadequate. Generative AI chat services have been rushed to market with much hype, but insufficient safeguards and warnings, despite their tendency to fabrication and affordances for "prompt-hacking" working for purposes at odds with information integrity.
- **Stakeholder knowledge deficits**: With the advent of Generative AI, the provision of media and information literacy and the evaluation of its impact are facing even more complicated challenges than previously. AI literacy and data literacy, and the political economy of tech business models, are still far from being integrated into the remit of media and information literacy.
- **Problems in policies and implementation:** Major social media services have rolled back policies against hate speech and misinformation. Reports also show backsliding on election integrity and openness to electoral partnerships. There is continuation of unequal transparency standards for different jurisdictions, and, linked to spending priorities, unequal treatment of different users and jurisdictions. In terms of application of policy, generic rather than tailored formulae are deployed, such as provisions about postal votes which do not apply in the country concerned.
- **More research is needed:** The studies referenced in this mapping can assist policymakers in further defining the parameters of human rights-based debate for strengthening information integrity. Still, much more research is needed in a very fast-changing field. Further, more investment is needed in comparative studies and with a focus on the global south. The rapid developments related to Large Language Models in content production and moderation are only just becoming the object of research by different institutions. However, a key constraint, which calls out for governance action, are the inadequate (and in many cases, shrinking) transparency affordances of digital companies (platforms and AI companies), and the costs and unequal standards for data access for actors outside of Europe and the USA. The result is that independent researchers and journalists are inhibited from playing a comprehensive monitoring role that could benefit the public interest in governance that favours information integrity.

A shift towards changes in digital governance

In its review of tech companies' 2022 transparency, Ranking Digital Rights found that none of the 14 digital platforms evaluated earned a passing grade. It reported that they failed to disclose adequate information about how they conduct human rights due diligence, moderate online content, test and deploy algorithmic systems, and use personal data. The observatory concluded that the "companies are content to conduct business as usual".

Against this background, it is evident that the tide is turning in many countries away from a previous absence of external regulation around content (with a powerful exception being the US's Digital Millenium Copyright Act). Outside of copyright, the general situation has been to allow a patchwork of self/solo-regulation by individual digital services, including the newer Generative AI ones. External governance, which has relied mainly on developing norms and issuing ethical exhortations, is now considering stronger frameworks. Prominent cases in the recent past are the European Union's package of digital regulations, as well as recent experiences in many other G20 countries, including the electoral regulation in Brazil.

The trend towards stronger societal governance, incorporating both laws and "softer" measures relevant to digital services, matches public opinion in many places. For example, a survey by UNESCO/IPSOS found that in 16 countries scheduled to have elections in 2024, nine out of ten respondents were worried about disinformation and hate in their upcoming polls, and wanted actions by the platforms, government and regulatory bodies. In the US, nine of ten members of the Institute of Electrical and Electronics Engineers (IEEE) want to see stronger AI governance, and especially on data privacy and risk assessments, both of which can be relevant to information integrity.

The governance debate today has therefore become about the mix of regulatory systems and various regulatory bodies, including self-regulation systems by digital service providers, and also encompassing discussion of multi-stakeholder involvement in hybrid arrangements. The discussion further entails which actors are obligated to build societal resilience for information integrity, such as by promoting media and information literacy and fact-checking, and supporting independent news media.

Responses to these questions need to be assessed as to their appropriateness to countering online content that potentially harms human rights and sustainable development, advancing access to data and information, and fostering that type of information which counts as a public good.

The debates come at a time of technological change. What had already become a complex governance issue concerning social media and search services (and their uses of classificatory AI and algorithms for ranking, recommending, and filtering content flows) has been further complicated by the launch and massive uptake of Generative AI, such as ChatGPT, and growing interest in governing AI across a range of elements and fields of application.

Evidence shows that the spread of Generative AI is massively increasing the volume of digital content, including potentially harmful content, as well as enabling its micro-targeted customisation and distribution via the social media platforms. The main social media platforms are also now investing in Generative AI models, that will more and more underpin content-creation which in turn will disseminate via their distribution services. Questions over the reliability and provenance of content output by Generative AI and circulated at scale are compounded by the negative impacts of this on traditional content producers such as news media and cultural creators whose intellectual products are arguably often being exploited, and who also both risk being undercut and overshadowed by the big AI operators. At the same time, AI can be used to more rapidly detect disinformation and allow for earlier intervention and deployment of counter narratives.

In the confluence of enhanced capacities for low-cost content production and content distribution, there are intersecting technology and business circles in play. These make for new challenges, but also opportunities, for governance that seeks to advance information integrity in general and information as a public good in particular. The picture now is that governance of actors (and behaviours) in the interests of information integrity covers not just social media and search services and their users, but also the more wide-ranging and cross-cutting significance of AI.

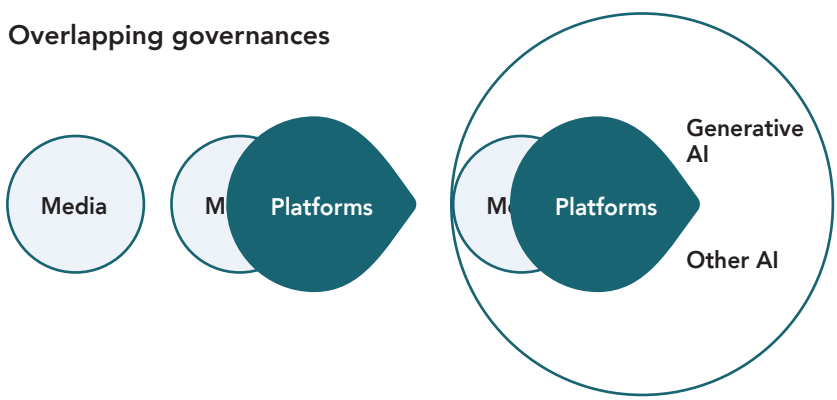
An illustration is the debate about whether open or closed models of AI merit different regulatory considerations for "safety", and at what stage of the AI lifecycle. This debate further covers what is to be meant by "safety" and what risks there are to people's rights to security, public health and equality. Rules for these matters can have an outsize impact on the information ecology and

information integrity in particular. A UNESCO publication shows many gaps which may also be relevant to the issue of content governance.

Because AI governance runs across (and beyond) aspects of content production, social media and search distribution, a holistic understanding is needed of its impact on the governance of these as regards information integrity. Thematic issues that span the combined landscape include (amongst others): company size and power; business models (and digital advertising); transparency and research access; data governance (including portability); privacy; intellectual property; human rights and fundamental freedoms.

In this light, considered narrowly, measures such as the Digital Services Act, the UK Online Safety Act, and the Canada bill are focused especially on social media and search companies. Yet, it is also necessary to assess these within the content significance of other regulatory initiatives. Examples are the EU's Digital Markets Act rules on inter-operability in messaging services, and intersections of content and AI measures in places ranging from Canada, Brazil, and China to India, the UK, the US, and other initiatives.

Figure 1. Below shows the expanding realms of governance that impact on informational space, including Generative AI and other forms of AI



Key debates

Countries are showing multiple approaches for governing the digital ecosystem – including AI and its potential intersections with content.

A key issue is the responsibility and liability along the digital communications value chain. This covers not just users, but also data and the tools. It extends to actual uses of digital content production possibilities; to the platforms for distribution and amplification; to the economic costs and benefits; and to the users/consumers (who are also often producers and disseminators of the content concerned). There are different views about whether interventions should be informed by the technological configurations and their affordances under present corporate decision-makers, or the secondary uses of these services. This is evident, for example, in debates about regulating AI as a technology, where concerns are regularly expressed about the stifling of innovation, although innovation and regulation are also not incompatible as is also evident in the regulated medical and transportation industries.

Debates exist about the mix of actors and institutions. Governance at different levels may take the form of rules that are mandatory, and/or voluntary (solo-decided rules or industry self-regulatory accords). Combinations of the two are possible, with mandatory rules setting a frame for voluntary rules and operating mechanisms. There is self-evident benefit in a field as highly complex as this, of including multistakeholder participation (with substantive influence) within both official and solo/self-regulatory mechanisms of regulation. This helps to enhance accountability and public interest considerations, as well as taps insights and interests from a range of actors.

There is value in including voluntary measures at individual company level or operating within an industry group, as part of the bigger picture of arranging governance power. Such systems encourage going beyond "box ticking" compliance with official regulations, and they can further work effectively when regulatory steps lag new technical developments. Examples of voluntary steps (some of which have multi-stakeholder involvement) include the Global Internet Forum to Counter Terrorism. the Global Network Initiative and the WeProtect Global Alliance. There are also the AI Alliance, the Accord to Combat Deceptive Use of AI in 2024 Elections, the Coalition for Content Provenance and Authenticity, and various company mechanisms such as Meta's Oversight Board and other companies' safety advisory councils.

Different existing regulators have legitimate interests related to their "turf" and varying remits. For example, there is mounting concern by election management bodies and stakeholders to hold digital actors accountable for related information environments. One example is the adoption of related principles and guidelines by the African Association of Electoral Authorities. At the same time, platform regulations during an election may well also implicate privacy and access to information regulators. As shown in various databases for 50 countries, as well as for Africa and Latin America, regulatory steps are piecemeal, and, with some exceptions, there is a dearth of mechanisms for co-ordinating the relevant regulatory and self-regulatory bodies in a jurisdiction (e.g. Audio-Visual, Consumer, Copyright, Print and Advertising bodies, etc.). On the other hand, it has been described as "low-hanging" fruit to bring together consumer-facing regulators to deal with AI-enhancements of behaviours that are already covered under existing rules.

Increasingly, digital governance entails variations and combinations of these elements, with great relevance to the governance that impacts on information integrity:

1. norms and regulations for risk-management;
2. regulations based on use cases;
3. the harnessing of existing laws (including regarding AI deployment and use); and
4. a "guidelines" approach (as distinct from law such as in the EU), as evident in the US, UK and Japan.

Output-base vs. use-case vs. rules-based vs. risk-based governance

One approach distinguishes between different AI governance approaches as to their focus on restricting particular outputs (e.g. facial recognition systems), versus that of risk-based and pre-emptive actions. Another involves regulations based on use cases. These are often post hoc and penalising in character. Use case regulations, in most cases, tend to respond to perceived immediate risks. Rules-based regulation often relates to outputs and uses, but it can also specify rules for risk-based regulation. The latter puts the onus on major actors to prepare for and mitigate potential threats. The EU's AI Act is seen by some as combining several logics.

In the UK and EU, regulations ensure that service providers fulfil a duty of care. In the UK, platforms are given a duty to moderate content so that their services strive to be free of unlawful content. Platforms and Generative AI service providers in the case of China are expected to devise moderation standards and mechanisms and share them with regulators. Turkiye had a contested law on the Regulation of Internet Publications and Combating Crimes Committed through Such Publications (Law No. 5651). India has the 2021 Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules.

Within this composite patchwork, there are measures specific to digital communications such as restrictions on non-consensual sexual imagery and flashing images, and cybersecurity and procurement issues (such as in the US's Executive Order on AI). Other digital controls that can impact content typically draw on existing or wider legal provisions such as on hate speech or intellectual property. In addition, specific groups are targeted. The protection of children is covered as a concern (including child sexual abuse materials) such as in the UK. China's regulations on AI recommender systems include the protection of seniors in its AI regulations; the US Executive Order on AI includes people with disabilities.

Distinct from content regulation, behaviours and uses and requirements for consumer redress systems, are other foci for governance relevant to information integrity. These include penalties for producing and disseminating disinformation specifically for fraud or electoral interference.

Economic and environmental regulation

The issue of market structure and dominance, which affects information integrity matters, continues to attract focus. The exercise of establishing "thresholds" of risk, or company size, is related to challenges of market dominance (in data, advertising, attention, and content production) and corresponding regulatory burdens. Balances are struck in order to not disadvantage small players or deter innovation. Relevant to information access, competition regulation aimed at achieving lower connectivity costs from service providers is in evidence. Environmental issues around technology are increasingly recognised as part of the regulatory challenges, and steps here may raise the costs of Generative AI services and server farms, thereby impacting the information ecosystem. There are debates about ex ante regulation concerning the market power of large companies, with fears that as they become even bigger, the scale of current challenges to information integrity may increase.

The EU's Digital Markets Act, effective from 2023, is a forerunner, having developed criteria to identify "gatekeepers" with significant market impact. Those designated must allow third-party inter-operability and data portability, and prevent unfair practices like favouring their own services in rankings. Non-compliance may result in fines of up to 10% of global annual turnover and even structural remedies like divesting parts of the business. The Act aims to benefit consumers, innovators and startups by promoting fair competition and consumer choice.

State and legal actions have also been evident across countries like Australia, Canada, Indonesia and South Africa to redress competitive imbalances between news media and digital companies, raising issues control of advertising exchanges of advertising exchanges and the gatekeeping terms of app stores, as well as news bargaining (as per below).

Intellectual property and the media

Contestations around compensation to news media for content contributing to digital services have raised debates about whether platform and AI regulation should include "must carry" and "must pay" provisions. Such controversy has a strong bearing on information as a public good in the wider ecology and the sustainability of news media as a "mission-critical" source of attributable and quality content. The intellectual property issue is also applicable where content is scraped for training AI systems and for developing Generative AI outputs, which draws from the content holdings of news media, platforms and websites.

Transparency

Transparency is a common focus of governance initiatives to audit processes or decision-making, yet it remains piecemeal in its interpretations, practices and legal requirements. The EU has the most elaborate regulations in this area. However, it is also the case that, on its own, transparency does not equate to actual change in policies or practices, although it is usually intended to stimulate better practice. There is a growing range of transparency measures, self-regulatory and mandatory. The complexity has been highlighted by the Integrity Institute, which outlines the data, including algorithmic and content data, needed to track harms, including in social media's design and processing activities. Researcher and journalistic access to data, as an extension to transparency, is also an issue relevant across social media, search and AI services.

Overview of regulatory impetus

There has been a surge in AI regulatory activity particularly in response to the release of Generative AI and apps such as Chat GPT. This has resulted in renewed calls for international coordination to ensure "regulatory interoperability" and to help avoid the phenomenon of "policy retrofitting".

Illustrating the plethora of initiatives in the AI space, the OECD by early 2024 provided a live repository of over 1000 AI measures from 69 jurisdictions. In the same period, the rapid development in AI policy was also evident in the launch of at least 18 AI legislation trackers. The observatory Digital Policy Alerts in March 2024 recorded 206 countries or territories involved in 229 laws, orders, standards, investigations and guidelines on AI under deliberation, and 195 as already having been adopted or implemented. These steps covered: bans on certain AI uses, mandated transparency reports and required risk assessments, and protections for the rights in regard to issues of hiring, targeted advertising or content recommendations. A total of 15 policy areas are identified by Digital Policy Alerts, ranging from content moderation, through to data governance, intellectual property and public procurement.

As is evident from this, such regulatory steps impact on many aspects of both AI services as well as those of social media and search. The following are policy threads detected by Digital Policy Alerts in March 2024 that are of (direct or indirect) relevance to online content governance and information integrity in digital space:

Table 1. Mapping digital policy threads relevant to information integrity

Policies	Number of jurisdictions[1]
International collaboration on AI governance	205 jurisdictions
Online content moderation provisions (including the multistakeholder process coordinated by UNESCO, which resulted in the Organization's Guidelines for the governance of digital platforms)	204 jurisdictions since 2008
Protection of user speech rights (and unjustified removal of media content in the case of the provisions of the European Media Freedom Act (vis-a-vis platforms' restrictions)	52 jurisdictions
Promotion of local online content	34 jurisdictions
Consumer protection in the digital economy	86 jurisdictions
The remuneration of online content	35 jurisdictions
Copyright issues in AI regulation	35 jurisdictions
The regulation of digital advertising	60 jurisdictions
Governing cookies	35 jurisdictions

Since governments work with non-binding guidelines as well as legal measures, it is also relevant to consider strategies, consultations, codes of practice, guidelines, and plans. According to the Digital Policy Alerts tracker, 571 measures cover data governance, 129 content moderation, and 78 competition.

Amongst the official regulatory steps in recent times, are a number that do not respect international human rights law as shown by UNESCO research. Regulatory interventions in some 50 countries covering a range of information-linked issues also show a lack of alignment with UNESCO's governance guidelines for digital platforms and/or its Recommendation on the ethics of AI – including the principle of institutionalising multistakeholder involvement, a principle which is strongly reaffirmed in the NETMundial+10 and WSIS +20 statement. Other cases show laws being used directly against

1. The designations employed and the presentation of material throughout this text do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, jurisdiction, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

individual users for their postings, including cases where seemingly legitimate speech is prosecuted for politically partisan reasons. All this continues a pattern detected by the Global Network Initiative in 2020.

While relatively rare, there are some policies to improve the information ecosystem promoting by alternative, non-commercial and/or decentralised social networks, However, such initiatives are yet to be significantly seen in most countries, with one exception being the European Public Digital Infrastructure Fund. Another area of enabling governance is advancing mass-scale media and information literacy, including support for citizens to critically evaluate their information sources. In the EU, Brazil and the UK, there are legal frameworks and or policies requiring the regulator or public entities to promote or take measures to develop media and information literacy, which in turn can incentivise platforms to elevate users' empowerment as a priority and encourage transparency about the impact of such steps.

The EU has advanced data portability and pluralism by mandating that messaging services be able to talk to each other. However, there is a gap in international steps to build a "data commons" that could pool data and digital infrastructure from the public and private sectors.

Many of the current and emerging regulations seek to obligate multinational service providers to tailor their services to comply with national and regional laws. However, interpretation of these rules and their enforcement is a major challenge for many countries. Inhibiting this further is the fragmented landscape of regulation, even within regional blocs. It is also not a foregone conclusion that standards in larger jurisdictions will impact others. For example, social media platforms already operate different privacy standards within and outside the EU, showing limits to the touted "Brussels Effect".

International actors in governance and information integrity

The G20 can position its actions and agreements within the wider landscape of governance developments. As elaborated in **Appendix C**, these include the UN system, with its range of norms and tools. Other international developments are occurring in the EU and the Council of Europe, the G7, the African Union, OECD, Brics, Mercosur and the Freedom Online Coalition. The G20 itself has AI principles and agreed statements of G20 digital economy ministers.

Also relevant are initiatives by private sector and by civil society actors, including several related to research and monitoring.

Conclusion

There are both persistent and growing challenges to information integrity from digital services across social media, search and AI. The information ecosystem, in general, and information as a public good, in particular, are under unprecedented change and stress. The current configuration of largely privately provisioned digital public infrastructure is widely assessed as damaging social progress in ways that can invalidate many of the advantages. Hence, it is not unexpected that governance arrangements are being widely revisited. With the increasing concentration of ownership, patterns of market dominance disadvantaging new entrants and widening the gaps between the connected and the disconnected, concerns are mounting that the inequities can harm and/or hollow out digital transformations that favour of human rights and sustainable development.

New companies using Generative AI have become forces of production highly relevant to the information space while existing companies already dominant in content distribution and discovery (e.g. Google, Meta and TikTok) are also investing in this technology. Continued solo-regulation by some of the big tech companies in the production and dissemination of digital content cannot be guaranteed to address the scale of threats to information integrity – or optimise opportunities to advance this as a critical factor in the digital economy. Automated content production and targeted delivery will feed into social media and search, and exploit their known affordances and business models. As adversaries increasingly use these capacities as weapons, as is being recognised in the industry, the volume and velocity of content that damages information integrity will be exponential. The casualty will be an informed public and a hamstrung digital economy.

These scenarios point to considering solutions, as developed in the companion report to this brief. Worth mentioning here is whether to mandate tech companies to rigorously assess systemic risks and adapt their policies, budgets and practices accordingly. It is especially important to adjust digital content production, curation, moderation and recommendation systems. The companies could also be required to ensure that user safety and empowerment is at the centre of the design process. To the extent that monopolistic tendencies and attention economics business models, powered by data mining and exploitation, are problematic factors, these could also be addressed by policies creating different incentives.

Accountability for the services of Generative AI, social media and search depends on the extent of transparency of the processes put in place to safeguard the integrity of information. This in turn is profoundly shaped by the legal regimes in which the companies operate.

Complementing action on these fronts, there is the issue of citizens' meaningful access to the digital world and the information therein. As part of safeguarding people from known harms, there is scope to build digital capabilities and critical thinking skills. Informed and educated citizens who can exercise and defend their rights online are one of the bulwarks against the erosion of information integrity. But complementary efforts will also be needed to foster the supply and dissemination of information as a public good, as the other side of the coin. These endeavours implicate open government, freedom of information regimes and support to independent media, including public service and community media.

Without change to the package of causes behind the erosion of information integrity (as elaborated in **Appendices A and B**), the future scenario for people already online, and especially for those still coming online, could be bleak. For countless persons, the availability of information as a public good will increasingly be diminished in proportion to content that contaminates information integrity.

On the other hand, innovatively scaling up governance systems to be commensurate to the exponential challenges, could turn the current tide. In this, a viable trajectory is aligning with international trends and actors, and adhering to international human rights law, not least in regard to freedom of expression, information pluralism and personal privacy. Sustainable development and the digital economy require nothing less.

Appendix A – Contemporary challenges to information integrity

- Women (not least in politics and journalism) are amongst those who continue in 2024 to face enormous harassment online with the perpetrators largely acting with impunity. An industry has emerged offering services for synthetic non-consensual intimate imagery, and exploiting platforms' systems in order to drive traffic to this content. Meanwhile, gender bias, as illustrated in a recent UNESCO study, characterises large language models. Certain groups suffer disproportional harms.
- Research and the public spotlight have revealed severe harm to the mental health, safety and self-image of young people, boosted by AI. Tech companies have created a voluntary framework for transparency around child sexual exploitation and abuse, but this has not been sufficient to stave off critical scrutiny in a number of countries concerned with systemic harms such as addictive features targeting children. Harm to individuals such as fostering eating disorders or unchecked online-bullying continues, while violent extremists promote their cause to vulnerable individuals through content available across jurisdictions.
- Further content problems online including racism, xenophobia and intolerance have been signalled by the UN Special Rapporteur on contemporary forms of racism. Racism in digital spaces was observed during Covid-19 in the context of South-South migration, and the problem has persisted. There are prominently acknowledged problems with online racial abuse and digital discrimination in facial recognition and Generative AI representations. Ambient digital racism has characterised Twitter (now X), while digital racism is also present on other platforms.
- Climate change action is being delayed due to rampant disinformation that benefits platforms and influencers financially, but continues to damage societies and economies that are affected by extreme weather patterns. There are also concerns that AI systems (themselves hugely energy-intensive) will further pollute digital content. These have adverse consequences for the development of the digital economy.

- Election integrity, which requires information integrity, is being harmed by social media platforms' increased tolerance of disinformation, including that reflecting AI-generated content.
- New trends show online content, not least that produced with Generative AI, being used extensively in online scams and fraud, and for sites featuring pirated news. There is also evidence of increased deployment of bots with additional questions about fake traffic from X (formerly Twitter) to advertisers' sites. There have also been serious increases in data breaches that violate privacy. It is evident that faked online content damages wider social trust and enables non-transparent micro-targeting of electorally-pivotal groups. Synthetic content (such as "deep fakes") has been tracked in numerous elections, armed conflict and war. A list by the Partnership on AI cites 14 potential harms from synthetic media. It is recognised that the creation and detection of manipulated media is adversarial, with continued adaptation by the various contenders.
- A growing controversy concerns the intellectual property of both the content and raw data that feed not only the giant social media platforms but also the business of Generative AI. Chat-GPT lacks source attribution in its outputs, which implies that users should take the service purely on trust even though many factual errors have come to light. Its owner OpenAI has now begun to purchase content rights from several media houses in order to transparently access some verified data sets. The wider debate has seen prominent court cases, as well as stand-offs in the face of legislation intended to compensate news producers for content benefiting the search and social media companies. Meta is publicly downgrading news on its services, and it is reported that the company's tactic against paying for news by banning such content in Canada has left the service there replete with misleading clickbait while smaller news media in that country, already struggling, have had a decline in referral traffic. However, in response to a fine by the French authorities, Google has said it will revise its method of calculating revenues generated by news on its services in that country, and offer an opt-out to news media not wishing to be scraped for AI training purposes. Meantime, news media continue to have to downsize or close, and their outputs are not recognised by the recommender systems of gatekeepers of digital distribution as meriting the visibility that would seem to befit a key element of information integrity.

- Search engine services are still dominated by Google to date despite criticism of the trends in the service such as pointing to low-quality AI content. AI-produced "junk news" content is showing up in Google searches, and there are "hallucinated" news articles being produced in response to queries put to ChatGPT. However, "search" as a digital functionality is now under competition from Generative AI "answer engines" such as Gemini, Meta-AI, Bing and Perplexity. Further, Google has also increasingly integrated answer-style features into its services ("People also ask..") , a step that reduces click-throughs that benefit the economics of content producers and also diminishes user exposure to pluralistic sources. The company in addition has its own "answer" service (AI Overview), which is shown to inherently produce falsehoods based on low-quality data in its training, and is also reported to provide dangerous advice. The ability of the public to find knowledge through direct discovery of diverse, authoritative and transparent information sources is profoundly impacted by such developments.

- As AI-generated content feeds the ecosystem, and is used as data in further AI operations, so the quality, authenticity and diversity of information results is predicted to deteriorate, akin to photocopies of photocopies. Alongside this, short-form video content consumption increasingly predominates over longer and less emotive kinds of content, displacing textual learning and adding complexities to research into informational content.

- While AI-generated content may currently form a relatively small faction of factually contested content, it is proceeding at pace, and without effective detection. "Prompt engineering" by users is able to hack "guard rails" on mainstream services to produce non-consensual pornography. This was illustrated in the infamous case of Taylor Swift, where a Microsoft tool – using provenance technology – was manipulated to generate sexual imagery (which in turn would have been included in the data sets seemingly randomly scraped for training the system). The output was then also left online on X for more than 17 hours (during which time it attracted 75 million views). Researchers have cautioned that full safety features cannot be built into AI foundation models since risks exist particularly in the context of application. They suggest that spending on "red teaming" and "stress testing" should be more focused on early warning monitoring and response, as well as on increasing AI and data transparency. Meanwhile "cheap fakes" and "misleading edits" persist as a problem even without AI involvement.

Appendix B – Factors driving challenges to information integrity

Business model

- Attention economics and data mining continue unabated by the largest platforms distributing content. The 2022 Declaration for the Future of the Internet, subscribed to by more than 60 jurisdictions including the EU, notes that "the once decentralized Internet economy has become highly concentrated and many people have legitimate concerns about their privacy and the quantity and security of personal data collected and stored online." Alternative services, such as decentralised or non-profit networks, as well as news media enterprises, find it difficult to compete in these markets against the dominance of the giants. Meanwhile, research shows YouTube's continued reliance on recommendations that drive users to more extreme content. TikTok is facing legal action for allegedly algorithmically driving depressive content that feeds suicides. The same business model that commodifies attention may also spread to Generative AI services with chat-bots becoming designed to maximise advertising and collecting data by hyping and/or delaying content results in order to optimise user "stickiness" on the platform. Google Search's personalised "discover" function is already a service that confirms the company's interest in keeping users on its site for longer, while its Gemini AI service (now prominent in its search interface) will run advertising. AI companies are beginning to pay a limited number of media houses for scraping their content, which has been described as an acceleration of the internet's extraction phase. However, as more AI-generated content enters the ecosystem, and becomes in turn data for further AI-processing ad infinitum, questions arise about the sustainability and integrity of original content producers left outside of contracts with AI services, leading to a diminishment over time of online information diversity.
- On the other hand, under EU legal pressure, Meta proposed an alternative model to advertising in offering users in that jurisdiction a subscription alternative for advert-free usage. However, the company did not disclose if

it would still collect subscriber data for micro-targeting in attention-oriented content feeds and algorithmically-driven recommendations, and the EU has deemed the proposal insufficient. The company is also developing paid service functionalities on WhatsApp which does not run adverts. There is some limited progress in companies voluntarily offering users portability through data transfer architectures, meaning that reinforcing walled gardens (and "moats" in the arena of AI companies) remains a primary feature of business operations exploiting "network effects" in ways that inhibit competition. Data collection (along with data retention and use) by Generative AI companies generally remains opaque, possibly because disclosure might raise input costs if it stimulates payment demands from the original asset owners.

Automated advertising

- Centralised and opaque advertising exchanges, dominated by the biggest search and social media platforms, continue to drive online advertising spend and placement, relying on third-party data about targeted persons which is obtained in part from tracking software ("cookies"). This problem has triggered inquiries against Google in the EU and the UK and anti-trust cases in the US. The ad-tech system also continues to channel revenues to AI-powered pirates, imposters and hate-filled sites and posts, as well as to AI-generated scam adverts. The same applies to online destinations promoting health and climate disinformation, while NGOs show that it is easy to get automated approval for adverts that blatantly contradict platform policy. The now-suspended process to phase out third-party cookies will reinforce the power of these data-driven dynamics. Meanwhile, Consumer Reports found that across a group of 709 Facebook users (with variations amongst each), a total of 186,892 companies (including data brokers, credit reports, Paypal and Amazon) were sending data to this particular social network. Privacy legislation seems stalled in many G20 countries, even though there are continued concerns around data harvesting, sharing and transfer. Information on political advertising is limited: for example, X's advertising repository states: "X is a platform that enables global conversation, and we believe that transparency is a core part of who we are", but the service only offers information on the EU. Google is slowly developing its disclosure and verification policies for electoral advertising, but by March 2024 covered under 20 countries.

Meanwhile, Ranking Digital Rights has found "virtually no transparency reporting from platforms about ad policy enforcement.

Manipulation

- The abuse of platforms for disinformation and hate speech continues, including by geo-political actors, despite the UN Secretary General's call for stakeholders to refrain from doing so. Co-ordinated information operations involving inauthentic identities and accounts are still regularly revealed by Meta, Google and TikTok, or exposed by journalistic investigations, but the actual scale is potentially much higher. Going further, covert "strategic communications" by various actors – including via paid "influencers" – continue to be enabled by the virality and targeting affordances of social media services. Disinformation-for-hire at cheap rates is increasingly an unregulated trans-border operation and boosted by AI technology. Although described as being still a minor phenomenon, the prospect is that AI can supercharge disinformation in terms of speed, scale, and personalization. The continuation of abuse, and scenarios for its increase, implicate current and future spending commitments by platforms in detecting and countering such violations of their terms of service and related policies.

Spending priorities

- Reports continue to appear about failures in moderating content that violates the particular platform's own standards. Benefiting from the absence of legally binding safety standards and metrics in most jurisdictions, major platforms have disinvested in staff working in content moderation known as "trust and safety teams". AI companies like OpenAI, reported to be opaque even to its own board, have also attracted critical attention in this regard. Despite Meta claims that harnessing Large Language Models is improving the company's ability to identify harmful content, their resort to AI is unable to provide appropriate contextual and linguistic judgement calls. Rushing AI-chat interfaces to market has resulted in services that spew out inaccurate content which can pose risks to people's health. User complaints and appeals remain poorly serviced, if at all. In 2022, Meta's "Oversight Board" received nearly 1.3 million requests from users to review the company's content moderation decisions, but in the same

year only processed 12 cases (selected for their strategic and policy implications), with the inference that huge numbers of complaints remain unaddressed. In 2024, it was announced that the Board would retrench some of its employees.

Stakeholder knowledge deficits

- The hype around Generative AI has taken focus away from the pressing policy need to improve governance of the primary vectors of content distribution – i.e. the social media platforms and search engine services. The volume of disinformation online clouds users' horizons and obscures the way that both platforms and malicious actors make money from the circulation of this type of content. In terms of ensuring informed stakeholders – including legislators and regulators, platforms' voluntary transparency policies are widely seen as insufficient, while in contrast corporate lobbying activity remains strong, not least in the EU.
- At the level of citizen competencies, media and information literacy initiatives and evaluation studies thereof – already insufficient to the challenges before the rise of dominant platforms as well as the ascent of Generative AI – are facing even more complicated challenges. Public AI literacy is relatively underdeveloped. Despite industry proposals to watermark and label content produced with Generative AI, there are insufficient signals for consumers to know when this is the case. There is little public literacy about Recommender AI systems and their role in structuring online content feeds and advertising targeting. Awareness is constrained by the lack of prominent explanations and the absence of alternative algorithmic options for users. Data literacy lags in regard to awareness and skill about the dynamic way micro-targeted synthetic media is geared to exploiting people's confirmation biases. Also lagging is public understanding of how Generative AI creates the potential for a destructive "liar's dividend" whereby people can come to distrust the entire breadth of their information sources. However, platforms like Google have taken some steps in mobilising stakeholders against disinformation, supporting fact-checkers and offering a specialised service (albeit not prominent) for searching for fact-checked content. Meta offers users a facility called "Why am I seeing this ad?" but has also eased its tolerance of advertising falsehoods that can mislead users.

Problems in policies and implementation

- One report has found that in the year prior to November 2023, Meta, YouTube and X rolled back a total of 17 policies against hate speech and misinformation. Another report shows backsliding by platforms on their commitments to electoral integrity. Geographic insensitivities and unfairness, as well as unequal treatment of users in policy application, continue to be reported. There are currently few economic incentives, outside of compliance with regulations in the EU, for platforms to invest in trust and safety. Google, TikTok and Meta have signed up to voluntary principles for election integrity. But they have been assessed as failing in actual preparedness and as shying away from suggestions about practical co-operation. Highly unmoderated platforms like Telegram are exploited by small but active communities to share misleading information and worse. As X (formerly Twitter) has shed its trust and safety capacity, the European Union (EU) has found that disinformation is proportionately most present on that service. X's owner uses the platform for partisan politics, cancelling out its potential as digital public infrastructure.

- Dominant individuals continue to have unfettered power to make corporate policy decisions, which are often arbitrary and override public interest considerations. These decisions range from specific users being allowed access or being expelled from their services, to applying/lifting "break glass measures" to protect elections. The include decisions about legal action against an NGO, changes to investment in content moderation and inadequate handling of user complaints. It is unclear which of X's policies remain meaningful given the drastic reduction in personnel under its current owner. Recently, imposter hyperlinks on the service have been found to lead users to spam websites. The integrity of information has further deteriorated with a "verified" tick now available to anyone paying for it, including outlawed groups. This is further compounded by a related policy that amplifies content produced by such "verified" users.

- It is also not at all clear how policies (where these exist) on the use of AI-generated content and advertising will be clarified, monitored, enforced and reported upon. A commitment by the biggest social media and AI companies to labelling AI won applause as a voluntary measure, but it also lacks independent auditing provisions. There are known limits to the technical ability to insert identifiers along the various stages of the digital

communications chain, and to apply them beyond images (as Meta has committed to doing with visible and invisible markers) to text, audio and video. Even with photos these technical features can sometimes be stripped out (as Meta also acknowledges), or may lead to mislabelling. Several AI companies have been found to have discrepancies between public promises and execution in regard to electoral integrity, while experiments show that safeguards for electoral image generation are easily hacked.

Mapping shows a need for independent research

- Social media platforms have come under criticism for lack of meaningful data and metrics, as to have Generative AI companies which are generally more opaque. Backward steps in transparency and researcher access have been recorded for TikTok, Meta and X. Data access possibilities for public interest researchers have particularly worsened at X, with EU authorities now investigating compliance with their compulsory access requirements. As a UNESCO report shows, YouTube and TikTok maintain limited API access for researchers in the EU and US, but not elsewhere. Meta provides access to some data sets (excluding WhatsApp meta-data), including to the Global South, via its content library and API. However, in this year of many elections, the company announced the shuttering of its Crowdtangle interface, a facility particularly used by journalists to understand Facebook behaviours. Meanwhile, the platforms sell access to their data holdings to clients, or to brokers who in turn on-sell access to third parties, at costs that exclude many academic researchers. The consequence of all this is to deprive external governance initiatives of important insight and evidence that could arise from independent study of platforms' performance and the effectiveness of their mitigation measures.

Appendix C – International landscape of digital governance actors

United Nations

Different UN entities are involved in digital governance, especially at the normative level.

The UN General Assembly passed a resolution in March 2024, highlighting the need to bridge the AI and other digital divides (including gender divides), build capacities and promote digital literacy, in the interests of safe, secure and trustworthy AI and the Sustainable Development Goals. It calls for human rights and fundamental freedoms to be respected, protected and promoted throughout AI systems' life cycles. A further call is for "domestic regulatory and governance approaches and frameworks, in line with their respective national, and where applicable subnational, policies and priorities and obligations under international law, to support responsible and inclusive artificial intelligence innovation and investment for sustainable development…". Additionally, it advocates for mechanisms of risk monitoring and management, and for securing data, as well as impact assessments as appropriate, across the life cycle of AI systems. Also encouraged are "internationally interoperable technical tools, standards or practices, including reliable content authentication and provenance mechanisms – such as watermarking or labelling, where technically feasible and appropriate, that enable users to identify information manipulation, distinguish or identify the origins of authentic digital content and artificial intelligence-generated or manipulated digital content – and increasing media and information literacy".

The resolution further champions the need for safeguards to respect intellectual property rights while promoting innovation, and the protection of personal data. It promotes "transparency, predictability, reliability and understandability throughout the life cycle of artificial intelligence systems that make or support decisions impacting end-users, including providing notice and explanation, and promoting human oversight". Linguistic and cultural diversity should be advanced within AI systems. Cross-border data flows are encouraged, as is "fair,

inclusive, responsible and effective data governance, improving data generation, accessibility and infrastructure, and the use of digital public goods". Finally, the resolution advocates for "cohesive, effective, coordinated and inclusive engagement and participation of all communities, particularly from developing countries, in the inclusive governance of safe, secure and trustworthy artificial intelligence systems".

A further resolution in July 2024 calls for international cooperation to address the AI digital divide. It focuses on the need to foster knowledge sharing and technology transfer in AI. The call is for "a fair, open, inclusive and non-discriminatory business environment" for AI design and development. It further urges international cooperation for capacity building in developing countries in order to ensure inclusive development. Capacity building is elaborated as encompassing "policy exchanges, knowledge sharing activities and the transfer of technology on mutually agreed terms, technical assistance, lifelong learning, personnel training, skilling of workforce, international research cooperation, voluntary joint international research laboratories and artificial intelligence capacity – building centres". The resolution highlights the importance of the "needs, policies and priorities of developing countries, with the aim of harnessing the benefits of artificial intelligence, minimizing its risks, and accelerating innovation and progress toward the achievement of all 17 Sustainable Development Goals". Also referenced is open-source AI and digital public infrastructure, among other methods and business models; linguistic and cultural diversity, including in training data, and the importance of proactive measures to counteract racism, discrimination and other forms of algorithmic bias. Without specifically noting AI, although this is the theme of the resolution, the UN is recognised as "playing a central and coordinating role in international development cooperation".

The Tech Envoy office of the UN Secretary General constituted an advisory panel that produced an interim report in December 2023 that highlights disparities in power around AI and proposes regulating in the public interest for data, models, benchmarks and applications. The report motivates for governance to serve as an enabler of AI for humanity and as a way to deal with risks and challenges, to be achieved through multistakeholder collaboration. Seven institutional functions for governance, to be covered by an institution or network of institutions, are highlighted. It notes further that: "At the global level, international organizations, governments, and private sector would bear primary responsibility for these functions. Civil society, including academia and

independent scientists, would play key roles in building evidence for policy, assessing impact, and holding key actors to account during implementation". The report proposes a function of international norms established through a Global AI Governance Framework endorsed in a universal setting (UN), and global harmonization of safety, and risk management standards. In addition, it calls for a specialized AI knowledge and research function, akin to the model of the International Panel on Climate Change Complementing this is a proposal for a new mechanism (or mechanisms) to facilitate access to data, compute, and talent so as to upgrade value chains and provide access to independent academic researchers, social entrepreneurs, and civil society. In the view of the report, legally binding norms (for example around lethal autonomous weapons) could be complemented by nonbinding norms. Timelines are suggested for institutionalising the functions. Former UN Special Rapporteur for Freedom of Expression and Opinion, David Kaye, has queried whether the final report will be watered down by corporate AI interests. Debate around the UN's role ranges from favouring limited options such as convening Global AI Safety Summits, to a more expanded one such as coordinating agreements on basic minimum standards for Member States.

UNESCO in 2023 finalised its "Guidelines for the governance of digital platforms: safeguarding freedom of expression and access to information through a multi-stakeholder approach". It has since been working with networks of regulators and researchers to develop and assess implementation, and has catalysed a Global Forum of Regulators promoting digital regulation consistent with the Guidelines. UNESCO developed the Guidelines consultatively with input with input from 1540 stakeholders in more than 140 countries, who provided more than 10,000 comments to the process, and they set out principles for how social media could be governed to combat misinformation, disinformation, and hate speech. The five high-level principles in the UNESCO guidelines are:

- platforms conduct due diligence on human rights;
- platforms adhere to international human rights law, including in platform design and business models, content moderation and content curation;
- platforms are transparent in their business practices, their technical design, and algorithm architecture;
- platforms make information and digital tools available for users; and
- platforms are accountable to relevant stakeholders

UNESCO's Recommendation on the ethics of AI is an agreed and adopted instrument for voluntary reporting by the organization's Member States. It spells out principles that rest on four core values (human rights and dignity, peace and justice, diversity and inclusiveness, and environmental flourishing). The instrument is accompanied by a Readiness Assessment Methodology and an Ethical Impact Assessment process. UNESCO convened its second global forum on the ethics of AI in 2024, and eight tech companies committed to apply the UNESCO Recommendation. UNESCO is also very active in Media and Information Literacy, leading the UN's Global MIL Week each year, developing resources and evaluation systems and building networks of stakeholders.

The UN Department of Global Communications has published a policy brief Information Integrity on Digital Platforms. It proposes a set of principles for Member states, digital platforms and other stakeholders covering transparency, trust and safety, data access, user empowerment, economic disincentives, and independent media. In June 2024, this was followed by Global Principles on Information Integrity – Recommendations for Multi-Stakeholder Action. The five principles are societal trust and resilience; independent, free and pluralistic media; transparency and research; public empowerment; and healthy incentives. The document encourages the formation of coalitions to advance these in practice, addressing recommendations to tech companies, AI actors, advertisers and other private sector actors, news media and fact-checkers, researchers and civil society organizations, states and political actors and the UN itself. It does not refer to governance and regulation.

UNDP has a programme to counter information contamination and in 2022 created an Action Coalition on information integrity in elections. UNDP's iVerify is a fact-checking tool that uses AI and machine learning to combat the spread of false narratives during election periods. Digital Kit 4 Dem is a suite of digital tools that support stakeholders in verifying information accuracy and this is deployed in over 15 countries. The eMonitor+ platform is a UNDP suite of digital tools designed to combat information pollution globally. The platform uses AI-driven tools to monitor and analyse online content and identify issues such as hate speech, misinformation, online violence against women, political polarization and electoral violations.

Meanwhile, the B-tech project of the Office of the High Commissioner for Human Rights (OHCHR) continues dialogues with platforms, including convening a summit on Generative AI in 2023.

UN's Member States are engaged in negotiations for the UN's Global Digital Compact (GDC) to be adopted as an Annex to the *"Pact for the Future"* by UN Member States (including those in the G20) at the Summit of the Future in September 2024. Other UN-related debates include the future of processes such as that of the ITU and many other UN agencies which continue to convene annual multistakeholder fora of the World Summit on the Information Society (WSIS). These follow action lines agreed in 2005 with attention to digital divides, capacity building and ethics. In 2025, The UN will host a high-level meeting in 2025 to review progress made over 20 years. Another process is the annual Internet Governance Forum, an output of WSIS, which continues to convene discussions and debates on contemporary topics.

Regional and multilateral processes

In 2022, the EU adopted the European Declaration on Digital Rights and Principles for the Digital Decade, to inform policies, budgets and programmes over the subsequent decade. The jurisdiction operates the Digital Markets Act, which has been assessed as having potential major impact on the gatekeeping of data, ranking systems, transparency, access and fairness.

The EU has also begun implementing the Digital Services Act (DSA), and has also adopted the AI Act (to be implemented in two years), each of which provides for stiff fines for violations. With the DSA applicable in 2024, the issue arises about voluntary compliance with the EU's 2022 code of practice on disinformation, given X-corp's withdrawal from this mechanism after receiving criticism for submitting an incomplete report. (This particular company has also not responded to requests by electoral and competition authorities for dialogue, suggesting a lack of appetite for voluntary co-operation). The DSA obliges designated platforms to publish biannual transparency reports including information about human resources dedicated to content moderation in each of the EU member states. The EU has launched proceedings to assess whether X breached the DSA in areas linked to risk management, content moderation, dark patterns, advertising transparency and data access for researchers. This arises from scrutiny of X's risk assessment report submitted in 2023, the company's 2023 Transparency report and its replies to a formal request for information. The code of practice is expected to become a compulsory code of conduct under the Digital Service Act, stimulating researchers to propose indicators and metrics for assessing compliance. The EU has also asked Facebook, Google Search, Instagram, Snapchat, TikTok, YouTube, and X about how they will mitigate risks linked to Generative AI on their services, such as viral dissemination of deepfakes and automated manipulation that misleads voters.

Many observers have recognised the value of the DSA in putting the onus on the social media platforms to demonstrate they are actually doing due diligence about problems on their services, and in the setting of standards for how they report planned and actual mitigations. Independent monitoring is needed to verify and assess the companies' claims, which requires – amongst other things – data access. Extensive transparency is needed in order to see if indeed companies are complying with their own commitments and the wider legal regime under which they operate.

The EU-U.S. Trade and Technology Council (TTC) has adopted a shared commitment to advance data access for researchers. During the 6th Ministerial Meeting of the TTC in April 2024 , Working Group 5 released the Status Report: Mechanisms for Researcher Access to Online Platform Data.

The EU's AI act designates certain AI use in elections as "high risk" and therefore as requiring companies to apply additional controls and scrutiny, plus clear and conspicuous disclosure for deepfakes. (The Act envisages tiers of risks – from unacceptable (e.g. social scoring), through to high (e.g. critical infrastructure, and for which there could be risk management systems and post-market monitoring systems), limited (requiring only transparency) and minimal (which is left to industry self-regulation). The EU has also convened a public consultation to seek views on draft DSA guidelines on the integrity of election processes.

The EU approach also covers additional transparency requirements for general purpose AI systems such as foundation models, which are expected to comply with EU copyright law, and to publish details of the content used in training.

A total of 40 collaborations between countries seeking a shared framework for AI governance have been recorded, mostly in regard to the Council of Europe. Here, 46 member states, plus 10 others, developed a treaty on AI, the Artificial Intelligence, Human Rights, Democracy and the Rule of Law Framework Convention, in March 2024. This entailed debate and compromise about application to the private sector, which resulted in provisions whereby signatory states may use voluntary measures rather than regulation, prompting criticism that this outcome makes the treaty more akin a declaration. However, the Convention does represent the first international treaty on AI that is legally binding. It establishes a standard for future work and it reflects shared values and principles to ensure AI use that respects human rights, the rule of law and democracy.

The Ibero-American Charter of Principles and Rights in Digital Environments was adopted in 2023 by 22 countries. It urges that individuals' rights such as

privacy must be protected in digital environments, and it calls for attention to youth, inclusion and connectivity. Further, it proposes that digital environments should be an inclusive, open and disinformation-free space.

The G20 AI principles of 2019, informed by the OECD Recommendation, were reaffirmed in the G20 Ministerial Statement on Trade and Digital Economy and the G20 Osaka Leaders' Declaration. Research that compares data governance systems and cross-border data flows between the G20's member countries shows heterogeneity in data processing, data flows including localisation and transfers, and sanctions regimes.

The G20 New Delhi Leaders' Declaration in September 2023 reaffirmed commitment to G20's 2019 AI Principles, and pledged a "pro-innovation regulatory/governance approach that maximizes the benefits and takes into account the risks associated with the use of AI". It encouraged "voluntary efforts to make digital public infrastructure interoperable" and recognised "the importance of data free flow with trust and cross-border data flows while respecting applicable legal frameworks". The G20's digital economy ministers in 2023 acknowledged the challenges of digital divides, including gender divides, and underlined the value of capacity building and digital public infrastructure. They also welcomed non-binding G20 High-Level Principles to Support Businesses in Building Safety, Security, Resilience, and Trust in the Digital Economy.

In May 2023, G7 leaders published a report on the Hiroshima Process on Generative AI aimed at fostering shared policy priorities and proposing a voluntary code for AI developers. It urged "appropriate data input measures and protections for personal data and intellectual property". The previous month, G7 Digital and Tech Ministers urged risk-based AI regulation. Recognising positions from earlier G20 and G7 meetings, they also called for Data Free Flow with Trust (DFFT) and to build convergence between existing regulatory approaches and instruments to that end. They further upheld practices such as stopping monetisation of disinformation content, and strengthening the accountability of digital platforms, plus encouraging platforms to allocate adequate resources which reflect language and cultural diversity for countering information manipulation and interference.

In 2023, the G7 adopted 'International Guiding Principles for Organizations Developing Advanced AI Systems'. These promote a risk-based approach, and a commitment to transparency and security throughout the AI lifecycle. The Hiroshima AI Process Friends Group comprising 52 countries and the EU expands the outreach of Hiroshima AI Process.

The G77 and China in 2024 adopted an Outcome Document of the Third South Summit urging reduction of all digital divides, and inequalities in data generation, infrastructure and accessibility. It states that "dealing with data and associated opportunities and challenges, will require a global response, with the equal participation of all countries, and stresses the need to strengthen international cooperation, and pursue greater harmonization in this regard."

The African Union's white paper on AI focusses on developing skills, infrastructure and data, creating an enabling economic climate, and a conducive climate for its deployment, and does not deal with content issues. The African Union has created a High-Level Panel on Emerging Technologies (APET), while the Association of Southeast Asian Nations (ASEAN) is active as well.

The Organisation for Economic Co-operation and Development has created OECD.AI to survey interoperability between national governance frameworks. It has also produced a report for the G7 Hiroshima process on the opportunities of Generative AI (e.g. productivity gains, innovation) and risks (e.g. disinformation/ manipulation; potential intellectual property rights infringement, threats to privacy). It recommended that the G7 could help provide tools for safety, quality control and capacity and trust building, as well as voluntary codes of conduct. Besides the already mentioned 'Facts not Fakes' report on information integrity, OECD has also published its principles for AI, its work on children online, as well as the 2023 policy considerations for Generative AI which identify policy issues of mis- and disinformation, bias and discrimination, and intellectual property rights.

The Brics countries show both convergence and divergence between their members' approaches to content governance (Brazil, Russia, India, China, and South Africa). Commonalities include stress on "digital sovereignty" and low to mixed independence of content regulatory mechanisms.

In December 2023, the Presidents of Mercosur agreed on a Declaration about information integrity and democracy in the digital environment.

The 39 Member States Freedom Online Coalition in March 2024 called on platforms to respect human rights, prioritise reliable and plural information on their services, and provide "more transparency and access to data in order to better understand how misinformation and disinformation is polluting the information ecosystem". It further urged states to promote information integrity online, and also to ensure that regulation of AI-generated disinformation should refrain from stifling freedom of expression.

In 2023, a Global Declaration on Information Integrity Online was adopted by 34 states. This calls on signatories to abstain from and condemn state-led disinformation campaigns, and avoid stifling freedom of expression under the guise of countering disinformation.

A follow up summit was held in Seoul in 2024, producing a declaration and ministerial statement on advancing AI safety, innovation and inclusivity.

A total of 38 participant countries at the 2023 UK AI Safety Summit signed the Bletchley Declaration in favour of developing a shared understanding of AI opportunities and risks.

A network is emerging of regulators dealing with online trust and safety, with a Global Forum created in June 2024. This will meet regularly along with civil society to co-ordinate efforts for platform governance.

Twelve countries have constituted the Global Partnership for Action on Gender-Based Online Harassment and Abuse. It aims for evidence-informed action to prevent, disrupt, and reduce the spread of targeted online campaigns against women political and public figures and human rights defenders.

Private sector and civil society initiatives relevant to information integrity

Complementing governance developments by authorities, are a number of private initiatives around a limited number of issues. The Partnership on AI brings together industry, civil society and academia, participates in global governance discussions and outputs recommendations on topics like transparency. Gaps include environmental considerations, and metrics for impact and implementation. A number of broad voluntary commitments concerning AI risks have been made to the US government by large players.

The Coalition for Content Provenance and Authenticity (C2PA), involving several large tech companies is developing a technical standard for verifying and securing the "origins, circulation, and trajectory of digital media". In parallel, 20 companies have signed "A Tech Accord to Combat Deceptive Use of AI in 2024 Elections" about approaching electoral risks of producing deceptive content at the level of AI platforms or foundational models, and at the level of distribution on social or publishing platforms. The response covers provenance labelling, detection using AI and creating public pathways for reporting, as well as action and engagement with civil society and academics. The accord also commits to support AI literacy. On the other hand, the initiative does not

provide for standards for "red-team" testing to model adversarial threats, nor for independent mechanisms to hold the signatories accountable and for the setting of metrics for evaluating performance. Also, not party to the accord are smaller companies documented as offering synthetic media services that are used for disinformation. A "Frontier Model Forum" of key tech actors provides for cross-company research and discussion on AI safety.

A gap in all this is commitment by platforms to positively promote information and trusted sources as a public good. They also have a way to go to demonstrate in practice their substantive and elaborated commitment to defend information integrity. A number of civil society initiatives have proposed criteria for companies to identify trusted news sources, but there is no evidence that this has stimulated companies to enhance visibility of related content on their platforms.

At the level of norms and tools from and for private actors, the World Economic Forum has launched the AI Governance Alliance which calls for "responsible AI leadership". It has produced papers covering AI safety across the data and foundation model stages of the AI lifecycle, and the governance of Generative AI. The latter deals with prioritization of harms and risks, and how governance can operate on a spectrum of open-to-closed access. It argues that "Equitable access and inclusion of the Global South in all stages of AI development, deployment and governance is critical for innovation and for realizing the technology's socioeconomic benefits and mitigating harms globally". It distinguishes between different AI governance approaches as to their focus on being risk-based (example, the EU), rules based (example, China), principles-based (e.g. Canada voluntary code of conduct), and outcomes-based (Japan).

There are aspirations to develop an international monitoring mechanism on the information environment akin to the International Panel on Climate Change. The Forum on Information and Democracy with endorsements from 50 countries has an Observatory, while the International Panel on the Information Environment is also active. Another initiative in this space is the UNESCO supported I4T Global Knowledge Network of 35 research centres from around the world.

An expert Digital Governance Discussion Group has been constituted to deliberate on scenarios, including from the global South, into the UN digital governance processes underway. Also on the civil society side, there has been a noted rise in professionalization (in tandem with increased outsourcing) of trust and safety work on various digital platforms, and this is now unfolding in AI companies.

For more information on UNESCO's work:

Guilherme Canela

Chief, Freedom of Expression and Safety of Journalists Section

Communication and Information Sector

g.godoi@unesco.org