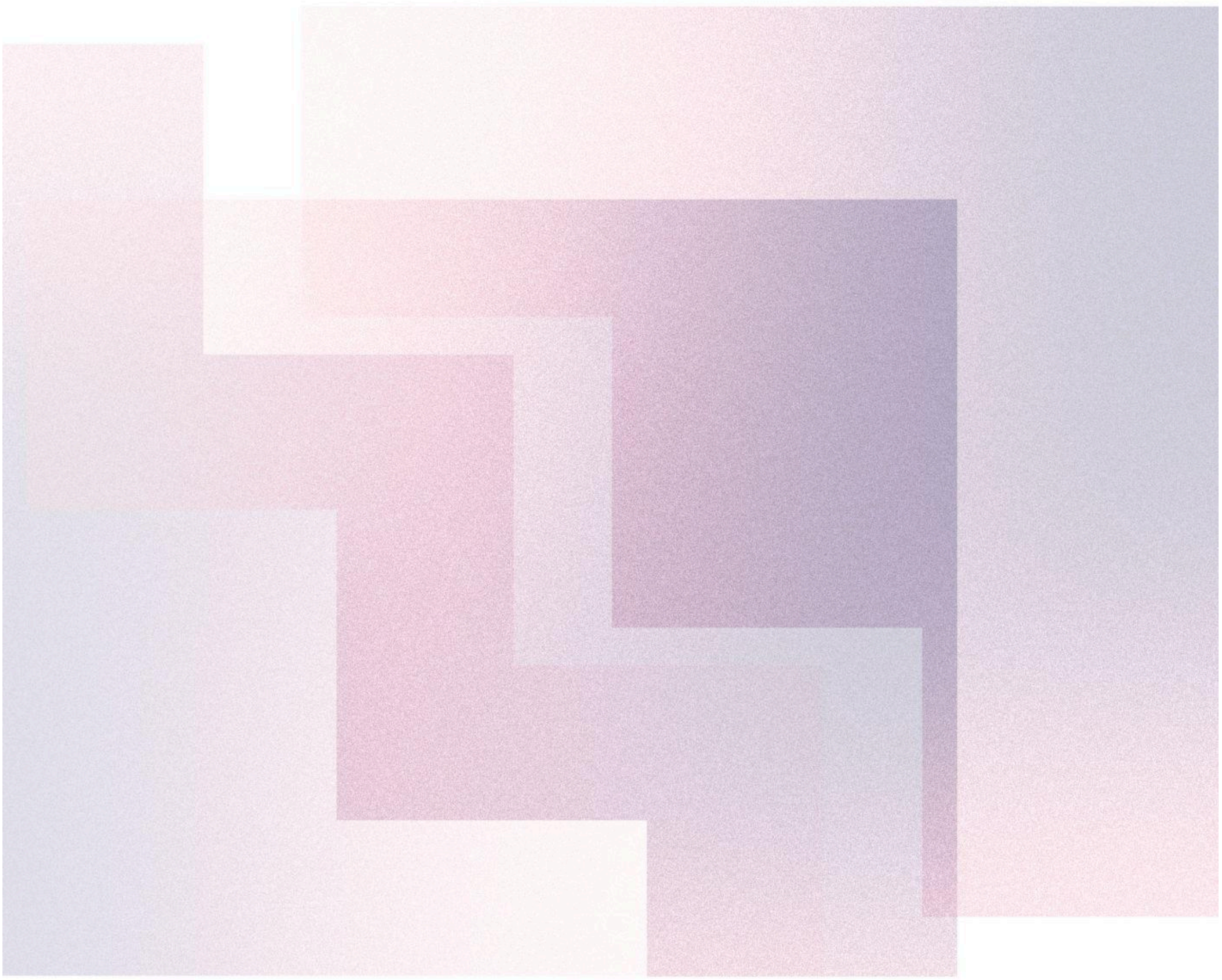
Platform Responsibilities for Information Integrity



Luca Belli, Yasmin Curzi, Rolf H. Weber, Merrin Muhammed, Anita Gurumurthy, Giullia Thomaz, Lorena Abbas, Roxana Radu, Marina Lucena, Nicolo Zingales.

Summary

How to cite this document: Belli, L., Curzi, Y., Weber, R. H., Muhammed, M., Gurumurthy, A., Thomaz, G., Abbas, L., Radu, R., Lucena, M., & Zingales, N. (December, 2024). Platform responsibilities for information integrity. Dynamic Coalition on Platform Responsibilities at the Internet Governance Forum (IGF).

Presentation

Information fabrication for political gain is not a new phenomenon. Following Posetti and Mathews[1], this practice can be traced back to 44 BC, when Octavian launched a propaganda campaign against Mark Antony. In a strategy that echoes today's social media tactics, Octavian used pithy; "Twitter-worthy slogans etched onto coins"[2] to tarnish Mark Antony's reputation, in an attempt to shape public opinion. Fast-forward to the 15th century, and the invention of the Gutenberg printing press made it possible for 'facts' to be more widely disseminated, facilitating the rapid spread of information—both factual and fabricated. As the press enabled news to circulate across Europe, it also opened the door for sensational stories and political manipulation[3].

Given the increase of disinformation and hate speech in the past two decades, international organizations and governments have been enacting policies to try to address this phenomena. The concept of information integrity has gained prominence in international discussions, especially in policies aimed at combating disinformation and promoting a reliable digital ecosystem. It has increasingly been adopted by international organizations[4] and governments, such as Brazil's in its G20 leadership[5], to describe strategies for constructing a safe and democratic informational environment. Unlike approaches that focus exclusively on mitigating negative phenomena like disinformation or

[1] For a thorough examination of democratic instabilities promoted by disinformation and hate speech campaigns, see Posetti, J., & Matthews, A. (2018). *A short guide to the history of 'fake news' and disinformation*. International Center for Journalists. https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf

[2] Idem.

[3] Idem.

[4] See Hanafin, N. (2022). *Information integrity: Forging a pathway to truth, resilience, and trust—Envisioning comprehensive and effective responses to information pollution*. Oslo Governance Centre, United Nations Development Programme (UNDP). https://www.undp.org/publications/information-integrity-forging-pathway-truth-resilience-and-trust

[5] Brazil, G20 in Brasil. (2024, May 3). *Brazil leads dialogues on information integrity and platform regulation: G20 side event discussed digital world challenges, including misinformation and hate speech, and proposed global solutions.* https://www.gov.br/planalto/en/latest-news/2024/05/brasil-leads-dialogues-on-information-integrity-and-platform-regulation

hate speech, the idea of information integrity aims at fostering a positive space where accurate and trustworthy information can circulate in a protected and accessible manner.

Nevertheless, while the idea of information integrity has been gaining traction, its adoption is not without critique. Scholars, such as Santos[6] and Ó Sióchrú and Gurumurthy (2024)[7] have been highlighting its origins on the Global North and its insufficiency for more diverse sociopolitical contexts. Besides that, the new approach to information integrity might be leaving aside the strong scholarship and expertise on platform responsibility and regulation.

The purpose of this policy brief, beyond examining the concept and its definitions, is to develop and propose a framework able to recenter information integrity within this scholarship. More specifically, it builds on previous works from the scholarship and the Dynamic Coalition on Platform Responsibility to develop possible paths to ensure platform responsibility.

[6] Santos, N. (2024, March 4). *Why do we need to discuss so-called "information integrity"?* https://www.techpolicy.press/why-do-we-need-to-discuss-socalled-information-integrity/.

[7] Ó Siochrú, S., & Gurumurthy, A. (2024, February 26). Digital platforms versus democratic political discourse: Challenges and the way forward. *Media Development, 2024*(1). https://waccglobal.org/digital-platforms-versus-democratic-political-discourse-challenges-and-the-way-forward/

Introduction: Understanding Information Integrity

The concept of information integrity finds its origins in computer science's information security studies. In this context information integrity can be traced to foundational work addressing the preservation of accuracy and consistency in secure computer systems. Among the seminal contributions to this domain is Biba's (1975) work on "Integrity Considerations for Secure Computer Systems", which introduced a dual-property system aimed at ensuring that untrusted or less reliable data does not contaminate trusted datasets, thereby upholding the integrity of sensitive information.[8] Biba's framework laid the groundwork for subsequent advancements in integrity assurance techniques and continues to inform contemporary cybersecurity practices. The origin of this concept provides a useful context to understand why and how information integrity has evolved, becoming a useful ally in the global efforts aimed at ensuring the trustworthiness of communications.

Indeed, with the increase of digital communication, it becomes crucial for democratic societies to ensure the integrity of communications, as it is a key factor for the full exercise of fundamental rights, becoming a pillar of democratic societies (OECD, 2024). As Posetti and Mathews[9] highlight, while "fake news" and "mis/disinformation" can be found in different historic periods as political strategies to affect public opinion, they emerge as topics of public relevance in 2014, during the escalation of the Russia and Ukraine conflict. Reports surfaced about the Internet Research Agency (IRA) based in St. Petersburg, with former workers revealing its operations to promote anti-Western and pro-Kremlin messages. According to leaked documents, IRA employees, often referred to as "troll armies," were required to post frequently on social media. On average, each worker managed multiple social media accounts, posting dozens of times per day across platforms like Facebook and Twitter, and

[8] Biba, K. J. (1975). Integrity Considerations for Secure Computer Systems (Technical Report MTR-3153). MITRE Corporation.

[9] Posetti, J., & Matthews, A. (2018). *A short guide to the history of 'fake news' and disinformation*. International Center for Journalists. https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf

engaging in discussions to manipulate public opinion. In Ukraine, the civil society initiative "Stop Fake", was established to counter these efforts, eventually expanding to other European countries by 2018.

The field of research on “platform responsibilities” were already established by then. In 2010, we had the Arab Spring, which marked a pivotal moment in the relationship between social media and political movements. Zeynep Tufekci[10] highlights how platforms like Twitter (currently X), Facebook, and YouTube played a crucial role in facilitating communication, organizing protests, and spreading information during the uprisings across the Middle East and North Africa. However, Tufekci — as many other authors[11] — also argued that these platforms, initially seen as tools of liberation, raised significant accountability transparency and questions, particularly in how they influence public discourse and political events.

It was also in 2014 that the Dynamic Coalition on Platform Responsibility launched its first efforts on the platform regulation agenda.[12] By bringing together various stakeholders, including tech companies, policymakers, and civil society, the Coalition’s goal was and still is to address these concerns by creating frameworks for more responsible governance of online platforms. Since then, the DCPR became a critical part of global conversations about platform accountability, content moderation, and the need for ethical standards in digital communication.This focus shifting was a direct response to the growing recognition that platforms were not neutral entities, but active participants in shaping political, social, and economic landscapes.

This was evident by other major incidents that underscored the growing influence of digital platforms, specially social networks, on political communication with the spread of mis/disinformation. In 2016, during the US presidential election, the spread of “fake news” on social

[10] Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

[11] Nardis, L. (2012). Hidden levers of Internet control: An infrastructure-based theory of Internet governance. Information, Communication & Society, 15(5), 720-738. https://www.researchgate.net/publication/241724480_Hidden_levers_of_Internet_control
Citron, D. K. (2014). Hate crimes in cyberspace. Harvard University Press. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2616790
Suzor, N. (2014, September). Promoting platform responsibility for content management. In Ninth Annual Meeting of the Internet Governance Forum 2014.

[12] The activities and outcomes of the Coalition are detailed in its webpage on the IFG website https://intgovforum.org/en/content/dynamic-coalition-on-platform-responsibility-dcpr

media reached alarming levels. One widely circulated story falsely accused Hillary Clinton of running a child abuse ring out of a pizzeria, leading a man to open fire inside the restaurant —the so-called "Pizza Gate". Additionally, Facebook disclosed that a Russian-backed operation had spent $100,000 to amplify fake news during the election period in the United States[13].

Also in 2016, significant social media activity influenced both the Brexit vote and the Philippine presidential elections. A large-scale analysis revealed that pro-Brexit supporters were not only more numerous on Instagram but also far more active than remain supporters, with similar trends observed on X (formerly known as Twitter). Foreign accounts were found to have sent hundreds of thousands of pro-Brexit tweets on polling day. In the Philippines, Rappler.com used investigative journalism, big data analysis, and fact-checking to expose state-sponsored disinformation campaigns. CEO Maria Ressa and her team faced ongoing online harassment linked to these efforts.

Meanwhile, the rise of "troll farms" and profit-driven fake news took off during the 2016 US election. A profitable troll farm run by teenagers in Veles, Macedonia, spread fabricated news, including false stories about Pope Francis endorsing Donald Trump. These operators earned significant sums through advertising revenue from sensationalist content. President Obama later referred to this phenomenon as a "digital gold rush". Similar hyperpartisan news sites that spread misinformation for profit were also emerging in the US, as revealed by a 2017 BuzzFeed investigation.

In response to these growing concerns, Facebook announced in 2016 that it would begin to flag "fake news" and take steps toward improving its content moderation practices.[14] This decision marked a key moment in the ongoing debate about the role of platforms in managing the flow of information online. However, as the following years have shown, platform responsibility remains a contentious and evolving issue, with debates intensifying over how to balance freedom of speech with the need to curb harmful mis/disinformation.

To address these concerns effectively, a proper categorization of the different types of information manipulation is crucial. Beyond platform responsibility, a field of research on

[13] Idem.

[14] See more: Jamieson, A., & Solon, O. (2016, December 15). Facebook to begin flagging fake news in response to mounting criticism. *The Guardian*. https://www.theguardian.com/technology/2016/dec/15/facebook-flagging-fake-news

dis/misinformation had emerged. Wardle and Derakhshan[15] distinguish between disinformation, misinformation, and malinformation, each representing a different dimension of information manipulation.

"Disinformation" refers to the deliberate creation and dissemination of false, inaccurate, or misleading information with the intent to harm an individual, social group, or organization. In contrast, "misinformation" pertains to false or inaccurate information shared unintentionally, without a deliberate intent to deceive. "Malinformation," on the other hand, involves the sharing of true information with the intention of causing harm, such as exposing private information to the public, like gossip. As acknowledged by several international organizations, such as OECD, these forms of false or misleading content can undermine social cohesion, erode trust in factual information, and weaken public trust in institutions.[16]

In spite of the established scholarship on both platform responsibility and dis/misinformation, the concept of information integrity has been emerging as a broader framework to enlarge or extend these concepts. It aims to address not only the accuracy and reliability of information but also the systemic and structural dynamics that shape the flow and accessibility of content in the digital space. Information integrity seeks to integrate considerations of human rights, democratic discourse, and public trust in the information environment. While it draws on prior concepts like dis/misinformation, it aims at expanding the scope to consider the role of platforms in enabling or inhibiting equitable access to reliable information, fostering social cohesion, and supporting the right to truth — that is, the platform responsibilities regarding communication human rights.

Despite growing recognition by governments, scholars, and international organizations, the concept remains far from universally defined, with significant debates about its scope, operationalization, and the challenges it faces in today's complex digital landscape. In the next section,

[15] Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/1680764c22

[16] OECD. (2024). *Facts not fakes: Tackling disinformation, strengthening information integrity*. https://www.oecd-ilibrary.org/sites/d909ff7a-en/1/3/1/index.html?itemId=/content/publication/d909ff7a-en&_csp_=037651e38039b75bbb486a006bf55a63&itemIGO=oecd&itemContentType=book

we will explore some of the key criticisms surrounding the concept of information integrity. After it, we will explore how information integrity could be resituated within the platform responsibility scholarship.

Challenges and Criticisms of the Concept of Information Integrity

The recent broader adoption of the concept of information integrity has sparked criticism, particularly in its current formulation. One key issue is the lack of consensus surrounding its definition, as evidenced by contrasting interpretations from prominent international frameworks. Comparing the recent UN Guiding Principles on Information Integrity[17] with the UNSG Common Agenda[18] reveals notable differences in conceptualization.

The Global Principles define information integrity as "entailing a pluralistic information space that champions human rights, peaceful societies, and a sustainable future," a definition that emphasizes inclusivity and long-term societal values. In contrast, the UNSG Common Agenda adopts a narrower view, describing it simply as "the accuracy, consistency, and reliability of information." This more restricted definition risks overlooking the broader social dynamics and institutional mechanisms that impact the integrity of information environments. Besides, it overlooks the nature of this environment complexity, as communicational challenges in the context of digital platforms are inherently deeply intertwined with the operational frameworks of the democratic process now (Curzi et al., 2019).

"Information integrity" arguably requires an expansion to encompass systemic and institutional factors that affect public discourse. In this sense, Ó Siochrú and Gurumurthy[19] suggest a shift towards communication integrity rather than information integrity. This change in terminology

[17] United Nations. (2024, June 24). *United Nations global principles for information integrity: Recommendations for multi-stakeholder action.* https://www.un.org/en/information-integrity/global-principles

[18] United Nations Secretary-General. (2023, June). *Our common agenda policy brief 8: Information integrity on digital platforms.* https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf

[19] Ó Siochrú, S., & Gurumurthy, A. (2024, February 26). Digital platforms versus democratic political discourse: Challenges and the way forward. *Media Development, 2024*(1). https://waccglobal.org/digital-platforms-versus-democratic-political-discourse-challenges-and-the-way-forward/

emphasizes the importance of structural context — acknowledging how platform dynamics, regulatory environments, and sociopolitical forces shape communicative interactions and information flows. A communication rights or communication justice framework could provide a more holistic understanding, as it broadens the scope to include not only the fidelity of information but also the rights and equity concerns integral to diverse, democratic discourse. Other authors such as Curzi and Belli[20], have also highlighted the need to address not only the content layer of communication but also the sociotechnical underpinnings of digital platforms that influence the spread of information.

Therefore, a more inclusive approach to information integrity might involve addressing these issues within the framework of communication rights, advocating for a digital ecosystem that safeguards access, inclusivity, and equity. Following Ó Siochrú and Gurumurthy, this perspective aligns with calls for a "communication justice" paradigm, which emphasizes the role of systemic equity and aims to empower marginalized voices within the digital information landscape. Embracing communication justice as a foundational principle for information integrity would facilitate a more comprehensive and just approach to policy formulation in the digital age, ensuring that integrity considerations are not solely focused on content accuracy but are deeply rooted in the context of communicative agency and social equity.

Another concern is that the concept of information integrity could be easily co-opted or captioned by states to legitimize censorship and control over the flow of information, depending on how it is defined and implemented. In some contexts, governments may invoke information integrity as a justification for curbing what they deem to be "misleading" or "harmful" content, which could easily lead to the suppression of dissenting voices, political opposition, or independent journalism, and leading to monopolistic information dissemination. This potential for misuse has raised alarms among critics, who argue that the term may be too easily manipulated by states to exert control over their populations' access to information.

[20] Curzi, Y., & Belli, L. (2024, April 22). *Integridade da informação no G20? Construção de um conceito e de uma agenda programática*. JOTA. https://www.jota.info/coberturas-especiais/g20-brasil/integridade-da-informacao-no-g20

Moreover, some scholars and activists[21] have critiqued the concept of information integrity as a Western-centric framework that may not align with the realities of information ecosystems in the Global South. These critics argue that the prevailing focus on content moderation, misinformation, and disinformation in discussions of information integrity overlooks the diverse ways in which information circulates and is governed across different geopolitical contexts. This challenge could be mitigated by promoting open access and interoperability, which would encourage a more diverse range of platforms and recommender systems. These measures could facilitate greater participation from a variety of stakeholders, including those from the Global South, in shaping and disseminating information. In turn, this could help to better reflect the realities of information governance in different geopolitical contexts.

Another important remark about the concept is that, currently, its debates are often framed too narrowly, focusing primarily on the negative externalities of digital platforms, such as disinformation and harmful content. While these issues are undeniably critical, discussions about information integrity should also consider positive and proactive measures to improve the quality of information on digital platforms. Rather than just addressing harmful or misleading content, there is an opportunity to foster positive information systems that encourage the production and sharing of accurate, diverse, and high-quality content.

By focusing on how digital technologies can enable individuals and communities to create, curate, and distribute trustworthy information, the conversation around information integrity could move beyond a defensive stance and promote a more proactive, solution-oriented approach to building a healthier digital public sphere. This shift in focus would help align the concept of information integrity with broader goals of enhancing democratic participation, access to knowledge, and social equity.

[21] Santos, N. (2024, March 4). *Why do we need to discuss so-called "information integrity"?* Tech Policy Press. https://www.techpolicy.press/why-do-we-need-to-discuss-socalled-information-integrity/ ; Ó Siochrú, S., & Gurumurthy, A. (2024, February 26). Digital platforms versus democratic political discourse: Challenges and the way forward. *Media Development, 2024*(1). https://waccglobal.org/digital-platforms-versus-democratic-political-discourse-challenges-and-the-way-forward/

An example relevant to information integrity concerns the electoral context. As mentioned above, the fabrication of information is not a new phenomenon. However, it can have serious consequences during elections, when it is important for voters to receive trustworthy information in order to decide how to vote and to understand and trust in the electoral process.

During elections, the impact of information related to the electoral process, candidates, and political parties might be more serious. Depending on the information quality and trustworthiness, the voter's intent and the electoral process can be affected. The concept is often highlighted as an enabler of reliable information ecosystems that could foster a positive space that includes trustworthy information. Specifically during elections, it is important that citizens receive accurate information about the voting system and the candidates. In the electoral context, it is common and expected that the candidates will express their opinions and ideas publicly, in a way to present themselves to voters. In the last few years, it has become more common for politicians, candidates, and political parties to be present and make their campaigns on social media platforms[22]. However, as we have seen, social media platforms can be used to spread false information more easily and quickly. Consequently, several laws worldwide are trying to address this problem to maintain the integrity of the electoral process. In Brazil, this has been done in a comprehensive way by the Electoral Superior Court and its Resolutions. These Resolutions are norms created by the Court which are applied during the electoral period, seeking to obtain more fairness and equality between the candidates and the process in general.

One important Resolution in this matter is n. 23.610, that regulates the electoral advertising during elections, adopted in 2019 and updated in 2024. For instance, article 7-A states that online content may only be amplified to promote and benefit candidates, not to harm (other) candidates. Article 9º-E is another relevant example and establishes that online content that is against the integrity of the electoral process must be removed by the platform immediately, or the platform will be legally responsible for them. In this case, the Brazilian Electoral Superior Courts provide a list of information that must be suppressed immediately, including hate speech, manipulated content, and threats to

[22] Recently, US elected President Donald Trump conducted his campaign with the presence of famous personalities on social media. In Brazil, another example is Pablo Marçal, former candidate for Mayor in São Paulo, who did not have allotted time for campaigning on TV and radio and yet achieved an expressive number of votes, probably due to his strength on social media.

Brazilian democracy. Another important Resolution is n. 23.714, which aims to combat disinformation and preserve the integrity of the electoral process. It grants the Tribunal the power to order digital platforms to remove not only any information that is established to be false or highly decontextualized (article 2), but also any identical content that surfaces in the future (article 3); similarly, the Tribunal may order the suspension of accounts and the prevention of creation or reactivation of accounts in the future (article 4).

This type of measure to preserve information integrity relates to the process of media and online content monitoring analysis, as defined by the UNDP Reference Manual fon Information Integrity for Electoral Institutions and Processes[23]. Another (and potentially complementary) approach seeks to build public resilience against attacks to information integrity through initiatives that promote media and Internet literacy and critical thinking skills, fact-checking and digital inclusion. Furthermore, planning and implementation of proactive, coordinated and targeted communication activities aiming to build credibility and trust in the Electoral Management Bodies and the electoral process, provide important voter information and counter disinformation narratives targeting the electoral process[24]. For instance, in Brazil, art. 9º-D, §3º of Resolution 23.610 establishes that the Electoral Justice may require the social media platform to disseminate, free of charge, informative content to clarify a seriously false or decontextualized information previously disseminated illegally.

This can be complemented by multi-stakeholder engagement to pool resources and expertise, with the aim to create a unified and effective approach to collectively analyse, prioritize and respond to threats in the information ecosystem[25]. In this regard, access to data relating to electoral campaigns and content moderation for academic researchers is a promising tool of accountability[26]. Finally,

[23] UNDP, Information Integrity for Electoral Institutions and Processes: Reference Manual for UNDP Practitioners. Available at https://www.undp.org/sites/g/files/zskgke326/files/2024-03/24119_undp_information_integrity_v07_rc_002.pdf.

[24] Id.

[25] Id.

[26] Rachelle Faust and Daniel Arnaudo, The Urgency of Social Media Data Access for Electoral Integrity . Tech Policy Press (March 2024). Available at https://www.techpolicy.press/the-urgency-of-social-media-data-access-for-electoral-integrity/.

independent and public interest media and journalism are an essential pillar of healthy information ecosystems and of inclusive and effective governance[27].

[27] Id.

Recentering information integrity within platform responsibility

The main concerns of information integrity are the risks posed by harmful content amplification, algorithmic biases, and the potential misuse of regulatory frameworks. All of those are at the heart of platform responsibility scholarship, which critically examines how digital platforms govern the flow of information, influence public opinion, and shape societal outcomes.

Thus, one of the central challenges is the role of digital platforms in amplifying harmful content. This issue is often tied to the underlying business models of these platforms, which are frequently driven by engagement metrics that prioritize attention over the quality of content. Many platforms are incentivized to maximize user engagement, typically through content that elicits strong emotional responses or generates higher levels of interaction. This business model, which heavily relies on targeted advertising, can inadvertently contribute to the spread of misinformation and sensationalized content, even if that content is not inherently false.

The algorithmic logic of content amplification on platforms like Facebook, Twitter, and YouTube tends to favor content that generates higher user engagement, which, as Zeynep Tufekci[28] and Tarleton Gillespie[29] have thoroughly pointed out, frequently includes sensationalized or polarizing content. In addition, algorithmic logic can also lead to a certain monopolization of content diffusion.

Suzor[30], Belli et al.[31], and Kaye[32] further focused on the legal and regulatory frameworks surrounding platforms, highlighting the risks posed by the current platform ecosystem, where platforms themselves largely self-regulate and prioritize profit-driven models. In this sense, scholars in

[28] Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*. https://doi.org/10.5210/fm.v19i7.4901

[29] Gillespie, T. (2014). The relevance of algorithms. *Media Technologies: Essays on Communication, Materiality, and Society*, 167-194.

[30] Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press.

[31] Belli, L. & Zingales, N. (2017). *Platform regulations: How platforms are regulated and how they regulate us*. Leeds.

[32] Kaye, D. (2019). *Speech police: The global struggle to govern the Internet*. New York University Press.

the field of platform responsibility have been solidly highlighting that this environment is ripe for exploitation, and more transparent and accountable business models are needed — ones that do not rely exclusively on programmatic advertising to fund platforms, but also consider the social and ethical implications of the content they amplify.

In response to these concerns, the UN Global Principles on Information Integrity calls for a fundamental re-evaluation of the business models driving platforms, particularly those that rely on targeted programmatic advertising. This means that platforms that depend on ad-based revenue models, where engagement metrics like clicks, likes, and shares are tied to financial success, incentivize the spread of content that may not contribute positively to public discourse or social cohesion.

Commercial incentives, which prioritize profitability over content quality, need to be reconsidered to ensure that platforms serve the public interest. Thus, platforms, stakeholders, and civil society organizations need to collaborate to develop more transparent, accountable, and socially responsible business models.

Another significant issue is the lack of diversity in platform policy development. Often, the development of content moderation policies, algorithmic design, and even strategic decision-making processes are controlled by a narrow set of interests — largely dominated by a few large tech companies headquartered in Western countries. This lack of diversity in decision-making can result in policies that are not reflective of global perspectives or local realities, especially in non-Western regions. Furthermore, the policies themselves may overlook the needs and concerns of marginalized or vulnerable groups, who may be disproportionately affected by harmful content or inadequate protections.

Furthermore, increasing trust in digital platforms is essential for the sustainability of democratic discourse and the broader information ecosystem. Trustworthiness is key to fostering a healthy digital environment, where users feel confident that the information they encounter is reliable, relevant, and responsible.[33] However, the lack of transparency in content moderation, algorithmic

[33] See Weber R.H., Transparency on Digital Platforms, Weblaw Jusletter IT, 31 August 2023, nos. 33-36.

decision-making, and the way platforms collect and use data has led to widespread skepticism regarding platform trustworthiness.

In this sense, platform responsibilities toward information integrity must include:

- An exploration of alternative monetization strategies that do not depend solely on engagement-driven advertising and may better align with the goals of promoting information integrity.
- Policies that prioritize content diversity, accuracy, and quality, and that actively reduce the incentive to spread sensationalist or harmful content.
- To enhance trust, platforms must commit to greater transparency in their operations. This includes clearly communicating how algorithms prioritize content, how content moderation decisions are made, and how user data is handled.
- Additionally, platforms should engage in ongoing dialogue with users, civil society organizations, and policymakers to ensure that their practices are aligned with public expectations and that they are held accountable for their impact on public discourse.
- Platforms should also implement mechanisms to prevent harmful content from going viral. This can include the use of "circuit breakers," such as automated flagging or human moderation, to halt the spread of harmful or false content before it reaches wide audiences. By introducing pause points, platforms can prevent dangerous content from gaining momentum while it is being reviewed.
- Building local partnerships with fact-checkers, journalists, and civic organizations can help platforms better understand the specific challenges and information needs of particular regions or communities. These partnerships should involve collaboration on content moderation strategies, media literacy efforts, and the promotion of diverse, high-quality content.
- The algorithmic systems that platforms use to curate content should be designed to prioritize diversity and inclusion, not just relevance. Platforms should explore

algorithmic models that promote content diversity—content that is both relevant to the user and offers new, diverse perspectives. The Forum on Information and Democracy's report on Pluralism outlines strategies for diversity-optimized algorithms that could help balance relevance with diversity, creating a healthier, more inclusive digital public sphere.

- Regular human rights impact assessments are necessary to evaluate the risks posed by platform algorithms and content moderation practices. Platforms should assess the potential negative effects of their services on information integrity and user rights, such as freedom of expression, privacy, and access to information. By identifying systemic risks, platforms can take steps to mitigate harm and improve their services in line with human rights principles.
- Platforms should invest in systems that predict the potential viral spread (reshare cascades) of content with reasonable accuracy. By using predictive analytics, platforms can proactively intervene to stop the viral spread of harmful content, such as misogynistic posts, hate speech, or extremist material, before it reaches large audiences.
- To better understand the impact of their services, digital platforms should provide vetted researchers with access to non-personal and pseudonymous data. This can help scholars and independent researchers analyze platform practices, understand the spread of misinformation, and assess the effectiveness of various content moderation strategies. By investing in research partnerships, platforms can contribute to a more evidence-based approach to addressing the challenges of information integrity.

Table 1: Role of platforms – Platform Responsibilities for Information Integrity

Platform Responsibility	Action/Strategy	Purpose/Goal
Alternative Monetization Models	Explore monetization strategies that do not rely solely on engagement-driven advertising.	Align platform revenue models with the goals of promoting information integrity and reduce incentives for

		harmful content amplification.
Addressing Algorithmic Biases	Engage stakeholders in addressing systemic and algorithmic biases that amplify harmful content.	Mitigate the amplification of polarizing, sensational, and misleading content through more inclusive and responsible algorithm design.
Prioritize Content Diversity, Accuracy, and Quality	Develop and enforce policies that prioritize diverse, accurate, and high-quality content, while minimizing harmful or sensationalist content.	Ensure that content circulating on platforms supports social cohesion, accurate public discourse, and inclusivity.
Regulatory Frameworks for Transparency	Policymakers should craft regulations that ensure transparency in algorithmic systems and hold platforms accountable for the societal impact of content amplification.	Strengthen platform accountability and transparency to foster trust, while ensuring platforms are responsible for the information they amplify.
Research and Advocacy by Academia & Civil Society	Encourage academia and civil society to conduct research and advocacy to highlight the importance of systemic changes and include marginalized voices in the conversation.	Amplify diverse perspectives and ensure that platform responsibility practices reflect the needs of all societal groups, particularly marginalized ones.
Transparency in Platform Operations	Platforms should clearly communicate how algorithms prioritize content, how content moderation decisions are made, and how user data is handled.	Enhance user trust by providing clarity on platform operations and decision-making processes.
Engagement with Stakeholders	Platforms must engage in ongoing dialogue with users, civil society organizations, and policymakers to align their practices with public expectations.	Ensure that platforms remain accountable to societal values, user expectations, and democratic principles.
Preventing Harmful Content Virality	Implement "circuit breakers," such as automated flagging or human moderation, to halt the spread of harmful content before it reaches large audiences.	Stop harmful or misleading content from gaining momentum and prevent further damage to public discourse.
Partnerships with Fact-Checkers	Partner with fact-checking initiatives, provide users with fact-checking labels, warnings, or context for contested information.	Combat misinformation and improve public understanding by ensuring that contested content is flagged and clarified.

Building Local Partnerships	Build partnerships with local fact-checkers, journalists, and civic organizations to better understand regional challenges and needs.	Tailor content moderation and media literacy efforts to address local needs, ensuring more effective and culturally relevant strategies.
Promoting Algorithmic Diversity	Design algorithms that prioritize content diversity, ensuring that content is both relevant and diverse, offering users new perspectives.	Foster a healthy, inclusive, and dynamic digital public sphere by introducing content diversity in algorithmic curation.
Human Rights Impact Assessments	Conduct regular human rights impact assessments to evaluate the risks posed by platform algorithms and content moderation practices.	Identify and mitigate systemic risks related to information integrity, user rights, and freedom of expression.
Predicting Viral Spread of Harmful Content	Invest in systems that predict and prevent the viral spread (reshare cascades) of harmful content, such as hate speech, misogyny, or extremist material.	Proactively intervene to stop harmful content from reaching a wide audience, protecting users from exposure to damaging material.
Access to Non-Personal Data for Research	Provide vetted researchers with access to non-personal or pseudonymous data to study platform practices and content moderation.	Support independent research to improve the evidence base for addressing misinformation, content moderation, and information integrity issues.

Policymakers and regulators

The concept of digital sovereignty has emerged as states grapple with the global nature of digital infrastructure and the data flows that shape the digital economy. Traditionally, sovereignty has been defined by territorial borders, but in the context of the digital world, sovereignty is increasingly tied to the control over digital infrastructure and data flows. While digital sovereignty can empower states to protect their national interests, an overly expansive notion of sovereignty could lead to the fragmentation of the global internet and the re-nationalization of governance structures. As philosopher Floridi (2020) warns, there is a risk that digital sovereignty could evolve into "digital sovereignism" or "digital statism," undermining the global, interconnected nature of digital communication.

Moreover, the power of private actors, such as large social media platforms and telecom companies, has created a situation where state authority is increasingly challenged by market-dominant private interests. The dominance of corporations like GAFAM (Google, Apple, Facebook, Amazon, Microsoft) in the US and BAT (Baidu, Alibaba, Tencent) in China highlights the growing influence of private entities over public discourse and information access. This tension between state sovereignty and corporate power calls for new governance models that prioritize global cooperation and uphold shared public values.

Governance interventions aimed at ensuring information integrity must align with the duty of states to protect freedom of expression, as outlined in the International Covenant on Civil and Political Rights. This requires that any regulatory action be grounded in a legal framework, be necessary and proportionate, and serve a legitimate purpose. Alongside these principles, strong human rights safeguards and transparent governance structures, including oversight of public digital infrastructure, must be implemented to ensure accountability.

While continuing to focus on aspects of current platform regulatory frameworks — such as content removal timelines, transparency, and accountability — policymakers must also consider more transformative measures to address the structural challenges of the tech-dominated digital ecosystem. This includes rethinking the role of the internet as a global public infrastructure. The internet should be seen not only as a space for verifying the accuracy and authenticity of information but also as an environment where diverse voices, including those from historically marginalized groups, are given a platform to be heard and valued.

To achieve this, regulators must confront the structural power of dominant tech companies and level the playing field for smaller, alternative platforms. Effective interventions may include stronger enforcement of competition and antitrust laws to prevent further consolidation in the tech sector. Policymakers should enforce structural separations within dominant tech firms and block mergers and acquisitions that would further entrench their market dominance. The Break Open Big

Tech Manifesto advocates for such measures, arguing that increased market competition will foster innovation and improve information diversity[34].

In addition, regulators should promote interoperability between platforms. This would allow users to easily switch between different services, thus reducing platform lock-in and fostering a competitive market environment. Furthermore, platforms' recommendation systems should be customizable, giving users more control over the content they see. Governments should also demand that platforms adopt value-sensitive recommender algorithms that are designed to promote diverse, high-quality content, especially in sensitive areas such as news distribution, political events, and elections. The Forum on Information and Democracy's Pluralism Report[35] emphasizes that content curation should highlight systemic issues, elevate marginalized voices, and spotlight journalism that reflects professional or experiential expertise.

To support the sustainability of independent and local journalism, policymakers should consider implementing revenue-sharing mechanisms between platforms and publishers, ensuring that content creators are fairly compensated. In addition, governments could explore providing public funding to support alternative, non-profit communication platforms that focus on civic missions, offering citizens a more balanced and diverse view of the world.

With this background, to ensure more effective governance of digital platforms, we highlight that:

[34] Break Open Big Tech Manifesto. (n.d.). *Beyond Big Tech: A manifesto for a new digital economy*. People vs Big Tech. Retrieved November 18, 2024, from https://peoplevsbig.tech/beyond-big-tech-a-manifesto-for-a-new-digital-economy/ ; Break Open Big Tech White Paper. (2022). *Break Open Big Tech white paper: A call for transformative change in the digital economy*. Retrieved November 18, 2024, from https://static1.squarespace.com/static/65c9daef199ea70aa66592fe/t/66f9c85ac5c3f44309088bfa/1727645794775/Break+Open+Big+Tech+White+Paper+-+FINAL.pdf

[35] Forum on Information and Democracy. (2023). *Pluralism of news and information in curation and indexing algorithms*. Retrieved November 18, 2024, from https://informationdemocracy.org/wp-content/uploads/2023/08/Report-on-Pluralism-Forum-on-ID.pdf

- Regulators should require regular audits of platform algorithms and content moderation practices to assess their impact on information integrity and public discourse. This includes human rights impact assessments to evaluate how platform operations might infringe on freedom of expression, privacy, or access to information.
- The introduction of regulatory sandboxes — pilot schemes that allow for the testing of new regulatory approaches in a controlled environment—has proven to be effective, particularly in emerging markets. Such sandbox regulations allow policymakers to experiment with innovative regulatory models while minimizing risks. However, to ensure these measures are effective, ongoing regulatory monitoring is necessary to assess their impact and adapt policies accordingly. The Alliance for Financial Inclusion's Innovative Regulatory Approaches Toolkit (2021) highlights the benefits of sandbox regulations in fostering innovation while mitigating potential harm.
- To avoid the pitfalls of digital fragmentation, regulators should ensure that any national policies related to digital infrastructure do not undermine global governance. This requires collaboration with international stakeholders, including transnational actors, to set standards that balance local interests with global cooperation. As Weber (2023) argues, standard-setting and governance must be objective-oriented and conducted in an open, transparent manner.

Table 2: Role of Policymakers and Regulators – Platform Responsibilities for Information Integrity

Platform Responsibility	Action/Strategy	Purpose/Goal
Governance & Human Rights Safeguards	Ensure regulatory actions align with legal frameworks, protect freedom of expression, and maintain transparency.	Safeguard fundamental rights while promoting information integrity, with strong accountability and oversight.
Regulatory Frameworks & Structural Challenges	Address structural challenges in tech, rethink the internet as a global public infrastructure, and consider diverse voices.	Create an open and diverse digital space where marginalized voices are heard, fostering equity and information authenticity.

Competition & Antitrust Enforcement	Enforce competition laws, block monopolistic mergers, and encourage market diversity through structural separations in dominant tech firms.	Prevent market consolidation, encourage innovation, and diversify the digital ecosystem to enhance information diversity and fairness.
Interoperability & User Control	Promote platform interoperability and customizable recommendation systems. Require platforms to adopt value-sensitive algorithms.	Enable user freedom, reduce platform lock-in, and ensure diverse, high-quality content, particularly in sensitive areas like news and elections.
Support for Independent Journalism	Implement revenue-sharing mechanisms and public funding for alternative, non-profit platforms.	Sustain independent journalism and ensure fair compensation for content creators, fostering a more balanced and diverse media landscape.
Digital Sovereignty	Balance state control over digital infrastructure with global cooperation. Avoid overreach that could fragment the global internet.	Protect national interests without undermining global interconnectedness, ensuring shared public values are upheld in digital governance.
Algorithm Audits & Content Moderation	Require regular audits of platform algorithms and content moderation to assess human rights impacts and ensure alignment with freedom of expression and information access.	Ensure platforms' operations do not infringe on fundamental rights and promote responsible content curation that supports public discourse and democracy.
Regulatory Sandboxes & Innovation	Introduce pilot schemes for testing new regulatory models in controlled environments, with ongoing monitoring.	Foster innovation in digital regulation while managing risks and adapting policies based on real-world impact.
International Collaboration & Global Standards	Collaborate with international stakeholders to create global governance standards that balance local and global interests.	Prevent digital fragmentation, promote global cooperation, and establish transparent, objective-oriented standards for digital governance

Academia and civil society organizations

The main contributions from academia and civil society are through evidence-based research, policy recommendations, teaching and advocacy. By examining the impact of algorithms with an independent and unbiased assessment process, content moderation policies and disinformation techniques for example, academia can offer meaningful contributions to policymakers, platforms and civil society. In this context, data accessibility is a crucial element. With the strengthening of R&D ecosystem academia can develop and refine technologies ensuring they are more accessible and aligned with human rights. Moreover, academics can help foster transparency by scrutinizing technology systems with unbiased assessments, provide digital literacy and engage with civil society to ensure safety online. Civil Society Organizations (CSOs) can act on the frontline of platform accountability, advocate for improved policies, and monitor government initiatives to protect citizens and uphold democratic values. CSOs work to ensure legal frameworks address emerging threats such as disinformation and algorithmic bias. They also collaborate across sectors, working alongside other stakeholders to develop comprehensive strategies that safeguard information integrity in the digital age.

In addition to advocating for greater transparency, CSOs play a critical role in enforcing stronger regulations that prioritize information integrity. They protect users and pressure platforms to disclose their practices regarding algorithmic decisions. Their efforts help ensure that digital platforms adopt a human rights-respecting approach.

Possible interventions include digital literacy programs and media campaigns focused on educating the public, especially vulnerable groups, on how to critically evaluate online information, protect their personal data, understand algorithmic biases, and be aware of their rights. These could take the form of hands-on workshops, online courses, and community outreach to stimulate critical thinking skills in digital context. Furthermore, academic institutions and CSOs could lead advocacy campaigns urging governments to regulate and monitor tech companies more closely. These campaigns would focus on holding digital platforms accountable and ensuring they operate under some standards to protect information integrity.

In addition to digital literacy and advocacy, academia and CSOs can collaborate on creating research-based policy briefs offering actionable recommendations to policymakers. By investigating the social, political, and economic impacts of misinformation and digital manipulation, these organizations can influence policy reforms that address the challenges posed by digital platforms. Finally, community based fact-checking networks can also be stimulated and established to help verify online content.

Table 3: Role of Academia and Civil Society Organizations – Platform Responsibilities for Information Integrity

Platform Responsibility	Action/Strategy	Purpose/Goal
Evidence-based Research & Policy Recommendations	Conduct independent, unbiased research on algorithms, content moderation, disinformation techniques, and digital platforms' impacts. Provide policy recommendations.	Offer data-driven insights to inform better policymaking, improve platform accountability, and address digital harms.
Data Accessibility & R&D in Technology	Strengthen the research and development ecosystem, focusing on making technology more accessible and aligned with human rights.	Ensure technologies are developed with a human rights perspective, and contribute to more equitable and accessible digital spaces.
Transparency & Independent Scrutiny	Assess technology systems through unbiased evaluations. Promote transparency in platforms' practices, including algorithms and content moderation.	Foster accountability by ensuring platforms and technologies are openly scrutinized for their impact on users' rights and information integrity.
Digital Literacy & Public Education	Launch digital literacy programs, media campaigns, workshops, online courses, and community outreach to educate the public on critical digital skills.	Empower individuals, especially vulnerable groups, with the skills to critically evaluate online content, protect their data, and navigate algorithmic biases.
Advocacy for Platform Accountability & Policy Reform	Advocate for stronger platform regulations,	Ensure platforms adopt human rights-respecting

	transparency in algorithmic decisions, and policies that protect human rights and information integrity.	approaches and are held accountable for their role in shaping public discourse.
Collaboration for Comprehensive Strategies	Partner with other stakeholders across sectors to develop strategies that safeguard information integrity and address digital threats like disinformation and algorithmic bias.	Create holistic, cross-sector solutions that protect democratic values and users' rights in the digital space.
Research-based Policy Briefs & Advocacy Campaigns	Collaborate to create actionable policy briefs and lead advocacy campaigns urging governments to regulate tech companies more closely.	Influence policymakers to enact regulations that prioritize information integrity and ensure platform accountability.
Community-based Fact-Checking Networks	Establish and stimulate community-based fact-checking initiatives to verify online content and combat misinformation.	Support the public in distinguishing reliable from misleading information, and reduce the spread of misinformation.