POLICY BRIEF

T20 SOUTH AFRICA 2025

G20 SOUTH AFRICA 2025

Information Integrity and AI Regulation: A Path Forward

2025

Carme Colomina, Senior Researcher, Barcelona Centre for International Affairs (Spain)
Javier Gonzalez, Director of Institutional Development, Ethos Innovation in Public Policy (Mexico)
José García, Postgraduate Student, Metropolitan Autonomous University (Mexico)

02

Digital Transformation

Abstract

Disinformation, particularly when amplified by artificial intelligence (AI), poses a growing threat to democracies worldwide by undermining access to reliable information – a right recognised under international law. As AI technologies rapidly evolve, they facilitate the creation and dissemination of false content at scale, raising urgent concerns about information integrity and democratic resilience. In response, a diverse array of national and regional regulatory approaches has emerged, ranging from content oversight to media literacy initiatives. However, these responses remain fragmented and uneven across countries.

This policy brief examines global trends in disinformation governance, analysing the rising number of regulatory initiatives, the geopolitical landscape, and the implications of AI-driven manipulation. It underscores the need for balanced, human rights-based policies that combine accountability with citizen empowerment, particularly in light of commitments made through frameworks like the UN Global Digital Compact and the G20 AI Principles.

The brief proposes key recommendations for the G20, including establishing a High-Level Task Force on AI and Information Integrity, encouraging ethical AI guidelines, and fostering global cooperation. It advocates for regulatory models that balance freedom of expression with protections against hate speech, based on international human rights standards such as the UN Rabat Plan of Action. It also calls on internet intermediaries to ensure algorithmic transparency and human rights accountability.

Finally, it emphasises the role of civil society, academia, and non-governmental organisations in promoting digital literacy – especially among marginalised groups – to ensure that societies can navigate the evolving digital landscape with resilience and informed agency.

Keywords: Disinformation, Artificial Intelligence, Information Integrity, AI Regulation, Content Moderation, Platform Accountability, G20 AI Principles, Global Digital Compact, Algorithmic Transparency, Ethical AI, Media Literacy, Human Rights

DISCLAIMER: This policy brief, authored independently and subjected to peer review, presents the views and opinions of its writers. These do not necessarily reflect the official policy or position of the authors' organisations or the T20 South Africa Secretariat.

Diagnosis

Globally, disinformation is an increasing threat to democracies, undermining the fundamental right to access timely and reliable information. International law recognises quality information as a public good, emphasising the need to protect its integrity – especially amid the rapid expansion of artificial intelligence (AI) technologies. AI systems enhance engagement, increase the opportunities to create realistic AI-generated fake content, and facilitate its dissemination to a targeted audience on a large scale.

Information is both a right in itself and a multiplier of other rights. Free access to information is an indispensable tool for democratic participation, as it helps promote government accountability and transparency and enables a more robust and informed public debate. However, the rise of AI systems developed in a pre-regulatory framework and their hyper-accelerated establishment pose a technological, economic, social, and geopolitical challenge that the major global powers address from diverse, sometimes contradictory, strategic visions.

The increasing number of public interventions related to disinformation and AI regulation, as depicted in Figure 1, highlights the growing need for governments to establish regulatory frameworks in this domain. The trend shows a significant rise in legislative and policy initiatives, particularly in recent years, indicating that concerns over AI's societal impact and the spread of disinformation have become more relevant.

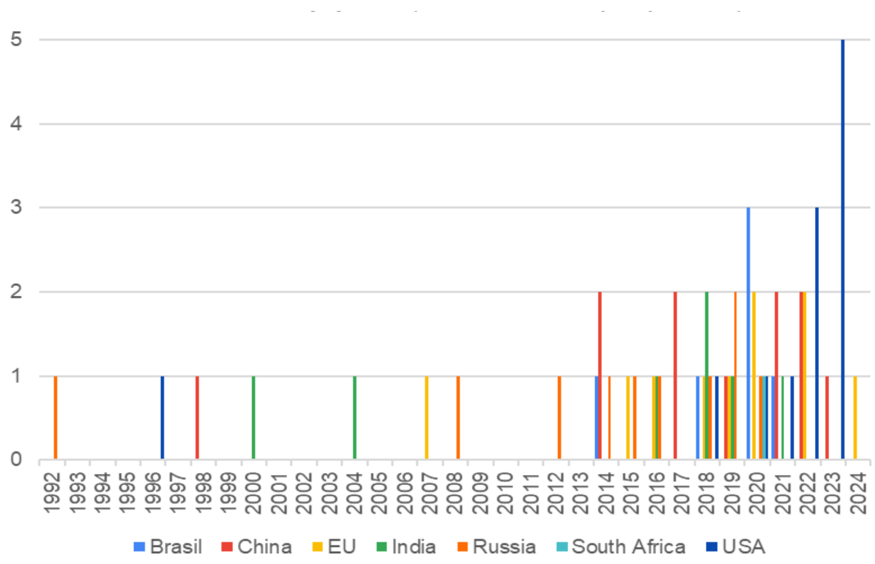

Figure 1. Number of public interventions in force and proposed in some G20 countries by year
Source: Compiled by authors

This rise in regulatory activity also reflects the extended nature of the challenge. While early interventions were limited to a few countries, recent years have seen a more widespread effort, involving diverse political and economic contexts such as Brazil, the EU, the US, China, and India. This growing regulatory momentum underscores the need for governments to not only intervene but also coordinate their efforts internationally. The complexity of AI-driven disinformation requires comprehensive policies to ensure that technology serves democratic values rather than undermining them.

As can be seen in Figure 2, the analysis highlights the divergent policy responses. While some nations have actively introduced multiple regulatory initiatives, others have remained largely absent from this trend. This disparity suggests varying levels of government intervention, influenced by political, economic, and social contexts.

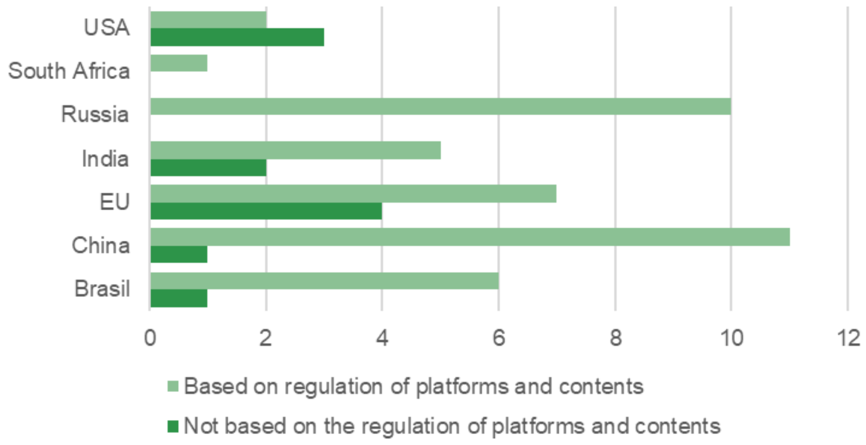

Figure 2. Number of public interventions in force in some G20 countries
Source: Compiled by authors

The nature of these policies also varies significantly. On the one hand, some governments have adopted paternalistic approaches, focusing on regulating content and platforms to limit the spread of disinformation. These interventions often involve direct government oversight, platform accountability measures, and restrictions on harmful content. With the emergence of generative AI, a recent policy approach is also focusing on content provenance mechanisms to enable the detection and tracing of AI-generated content – mainly the requirement of watermarking techniques – even if there is still a long way to go to develop interoperable international technical standards for watermarking and rules to help users distinguish content generated by AI from non-AI-generated content. On the other hand, there are also non-restrictive policies, such as media and information literacy programmes, that aim to empower citizens by improving their ability to critically assess digital content rather than imposing direct limitations.

Despite the different regulatory initiatives, there are emerging concerns on the use of AI techniques to automatically regulate content. Paradoxically, AI systems are increasingly being used to tackle disinformation, the dissemination of which is also amplified by other AI systems.

In September 2024, the UN General Assembly adopted as a principle of the Global Digital Compact[1] that international cooperation will harness digital technologies to advance all human rights and will empower all women and girls' rights. It must be ensured that[2]

> people can meaningfully and safely navigate the digital space and are protected from violations, abuses and all forms of discrimination… [cooperation] will advance a responsible, accountable, transparent and human-centric approach to the life cycle of digital and emerging technologies, with effective human oversight.

In the compact, leaders committed to initiatives to enhance digital security by ensuring greater accountability among tech companies and social media platforms, while also taking measures to address disinformation and online harms.

The UN acknowledges the role of platforms in amplifying previously unheard voices and empowering global movements, but also criticises them for "revealing the dark side of the digital ecosystem".[3] In this context, the new Global Principles for Information Integrity, published on 24 June 2024,[4] emphasised the urgent need

[1] United Nations, *Pact for the Future, Global Digital Compact and Declaration on Future Generations*, September 2024, https://www.un.org/sites/un2.un.org/files/sotf-pact_for_the_future_adopted.pdf

[2] *Idem.*

[3] UNESCO, An inclusive and safe digital ecosystem in Mexico is promoting by the UN Information Centre, UNESCO and OHCHR, 2024.

https://www.unesco.org/en/articles/inclusive-and-safe-digital-ecosystem-mexico-promoting-un-information-centre-unesco-and-ohchr#:~:text=However%2C%20these%20platforms%20have%20also,the%20right%20to%20access%20information.

[4] Available in https://www.un.org/en/information-integrity/global-principles

to address the damage caused by disinformation and hate speech while safeguarding human rights and freedom of expression. However, the voluntary and non-binding nature of these principles has had limited success in curbing these behaviours. Furthermore, legislative efforts to curb hate speech have also raised concerns that regulation may silence dissent and opposition.[5]

Multilateral debates and regulatory efforts, such as the legal framework established by the EU with the Digital Services Act (DSA), directly point to the responsibility of social media and demand transparency mechanisms to make platforms accountable for enabling or facilitating harm. In this context, in January 2025, major tech firms agreed to sign the updated Code of Conduct+, which has been integrated into the DSA. This requires major tech platforms to rigorously combat hate speech and undertake various tasks, including implementing a network of "monitoring reporters" – public or non-profit entities with expertise in illegal hate speech – to regularly monitor how their platforms review hate speech.

Link to G20 priorities

The G20 AI Principles remain a guiding paradigm for the regulatory focus of national legislation and policies. They include overarching principles such as human-centred values and fairness, transparency and accountability, digital safety and bias, privacy protection, and a systematic risk management approach at each stage of the AI system lifecycle.

[5] United Nations General Assembly, *Promotion and Protection of the Right to Freedom of Opinion and Expression*, A/74/486 (October 9, 2019), https://undocs.org/A/74/486.

The G20's discussions on disinformation have grown in importance, reflecting the increasing recognition of its impacts in the digital age. It has moved towards a more active stance, addressing the challenges posed by disinformation campaigns, particularly concerning public health crises and the integrity of online information ecosystems. Recent declarations emphasise the necessity for greater transparency from digital platforms, robust privacy protections, and the promotion of digital literacy as key strategies to mitigate the harmful effects of disinformation.

On the other hand, the G20's standpoint on AI has evolved from acknowledging its potential to a strong focus on establishing ethical and governance frameworks. The group has progressively emphasised a "human-centred" approach, promoting principles for responsible AI development and use. The G20 is actively working to balance the innovative potential of AI with the imperative to mitigate its risks, ensuring that its benefits contribute to sustainable development and digital inclusion.

Conclusion

The G20 High-Level Initiative/Task-Force on Artificial Intelligence and Innovation should take into account the commitments on information integrity set out in the Digital Global Compact.[6] However, it is fundamental that this premise consider two indispensable balances upon which the proportionality of any adopted measure and its impact on the quality of the digital public sphere and freedom of expression are based.

[6] *Pact for the Future, Global Digital Compact and Declaration on Future Generations*, *2024, p.44-45.* Particularly important is the "call on digital technology companies and developers to continue to develop solutions and publicly communicate actions to counter potential harms, including hate speech and discrimination, from artificial intelligence-enabled content. Such measures include incorporation of safeguards into artificial intelligence model training processes, identification of artificial intelligence-generated material, authenticity certification for content and origins, labelling, watermarking and other techniques".

Firstly, a balanced combination of government oversight and non-restrictive policies may be the most effective way to tackle the issue, as regulating platforms alone might raise concerns about censorship and algorithmic bias, while relying solely on literacy programmes may not be sufficient to curb the spread of disinformation. However, the disparity in regulatory strategies highlights the absence of a common global framework, which is crucial given the transnational nature of digital information. Establishing a shared approach that respects national differences while ensuring effective global coordination will be essential in addressing the challenges posed by AI and disinformation.

Secondly, taking into account the UN Rabat Plan of Action[7] – which calls for the provision of guidance to governments on the difference between freedom of expression and "incitement" (to discrimination, hostility and violence) – any regulatory effort has to present a balanced combination between fostering free expression while preventing hate speech according to the standards of international human rights law.

Actionable recommendations

- Taking into account the UN Global Principles for Information Integrity, the G20 should encourage the protection of the information space by asking governments to actively ensure and protect a pluralistic media environment. It should also encourage providing timely access to reliable and accurate information in crises; abstaining from conducting or sponsoring information operations, domestically or transnationally; and

[7] United Nations High Commissioner for Human Rights, *Report of the United Nations High Commissioner for Human Rights on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred*, A/HRC/22/17/Add.4 (January 11, 2013), https://www.ohchr.org/sites/default/files/Rabat_draft_outcome.pdf.

assessing whether and how independent regulatory mechanisms and governance may be developed to ensure enforcement and compliance by online information platforms.

- Given the growing interconnection between AI and disinformation, the G20 should encourage governments to treat AI and disinformation as complementary policy areas, advocating for clear labelling of AI-generated content to mitigate misinformation risks, and promoting frameworks that align AI governance with digital information integrity to prevent the misuse of AI in spreading false information.
- Taking into account the OECD AI Principles[8] as the first intergovernmental standard on AI, the G20 should encourage the harmonisation of AI guidelines to achieve an effective and ethical use of these tools, particularly to address mis/disinformation amplified by AI, while respecting freedom of expression and other rights protected by international law. Governments should encourage AI developers to implement mechanisms and safeguards, such as ensuring human agency and oversight, and to provide plain and easy-to-understand information on the sources of data/input, factors, processes and/or logic that led to the prediction, content, recommendation or decision. This will allow those affected by an AI system to understand the output.[9]
- Building on the commitments established in the Digital Global Compact, national regulatory frameworks should enforce that digital technology companies and social media platforms enhance the transparency and accountability of their systems. This must include terms of service, content moderation, and recommendation algorithms. This will empower users to make informed choices and provide or withdraw informed consent.

[8] Available in https://www.oecd.org/en/topics/ai-principles.html

[9] *Idem.*

- The South African G20 presidency should foster the establishment of a High-Level Task Force on AI and Information Integrity, encouraging ethical AI guidelines and fostering global cooperation. This task force should include the participation of experts, regulators, and vulnerable groups, and submit its recommendations to the authorities of the Sherpa track in the form of draft regulations, sandboxes, and best practices.
- Establish and enforce regulatory frameworks that ensure internet intermediaries uphold corporate human rights accountability, particularly in the curation, ranking of content, and automated processing of personal data. Given the pivotal role these platforms play in the digital ecosystem, governments and relevant stakeholders must implement clear standards and oversight mechanisms to safeguard fundamental rights and mitigate digital harms.
- The G20 must endorse the UN Global Compact on digital technology, calling on companies and developers to[10]

> continue to develop solutions […] to counter potential harms, including hate speech and discrimination, from artificial intelligence-enabled content. Such measures include incorporation of safeguards into AI model training processes, identification of artificial intelligence-generated material, authenticity certification for content and origins, labelling, watermarking and other techniques.

[10] *Idem*.

Appendix

Public interventions on disinformation and AI in some G20 countries

Country / region	Public intervention	Description	Regulatory level	Year	Status
Brazil	Internet Civil Framework (Marco Civil da Internet) - Law 12.965/2014]	Establishes principles, guarantees, rights, and duties for Internet use in Brazil, including the protection of freedom of expression and accountability for damages arising from content generated by third parties. This framework is relevant to the context of disinformation and how platforms manage content, though it was not designed specifically for disinformation.	mid	2014	in force
Brazil	General Personal Data Protection Law (Lei Geral de Proteção de Dados Pessoais (LGPD)) - Law 13.709/2018	While not specifically about disinformation or AI, the LGPD establishes strict rules for the collection, use, and processing of personal data. This can indirectly impact the ability to disseminate targeted disinformation and the development of AI systems that respect privacy, including limits on data used for profiling and content targeting.	mid	2018	in force
Brazil	PL 2630/2020 Brazilian Law on Freedom, Responsibility and Transparency on the Internet (aka "PL das Fake New" or "PL da Censura")	A comprehensive bill aimed at combating online disinformation in Brazil. It establishes obligations for digital platforms, such as identifying automated accounts (bots), providing transparency regarding content promotion, and removing content deemed false or harmful. It also includes provisions related to the use of AI in content moderation. This is the key piece of legislation focused on social media and disinformation.	high	2020	proposal

Brazil	PL 21/2020 (Legal Framework for the Development and Use of Artificial Intelligence) (Marco Legal Brasileño de Inteligencia Artificial)	A bill establishing principles, rights, duties, and governance instruments for the development and use of AI in Brazil. It addresses issues such as algorithmic transparency, liability for damages caused by AI, and promoting a responsible innovation environment. This includes mitigating risks related to AI-driven disinformation.	mid	2020	proposal
Brazil	Program to Combat Disinformation with a Focus on Elections (Programa de Enfrentamento à Desinformação com Foco nas Eleições)	The TSE has a permanent program to combat disinformation, with a particular focus on elections. The program includes actions to monitor social networks, awareness campaigns, partnerships with digital platforms and fact-checking agencies, and the application of sanctions in cases of dissemination of false news.	mid	2020	in force
Brazil	Brazilian Artificial Intelligence Strategy (Estratégia Brasileira de Inteligência Artificial (EBIA))	Defines the objectives and guidelines for the development of AI in Brazil, including aspects related to ethics, safety, and responsibility. The EBIA aims to promote the use of AI for the well-being of society and economic development. While it doesn't have specific provisions solely for disinformation, the ethical principles it promotes are relevant to addressing the misuse of AI.	mid	2021	in force
Brazil	National Network to Combat Disinformation (Rede Nacional de Combate à Desinformação (RNCD))	A proposed initiative that aims to bring together efforts from different sectors of society (government, academia, companies, civil society) to combat disinformation. The exact form and powers of this network are still under discussion. It represents a potential move	low	NA	in force

		towards a more coordinated national response.			
Brazil	Superior Electoral Court (TSE) resolutions for Elections	The TSE issues specific resolutions for each election, which include rules on election advertising, combating disinformation, and the use of technologies like AI. These resolutions aim to ensure the fairness of the electoral process and the integrity of public debate, including restricting the use of AI to generate deceptive content.	high	NA	in force
Brazil	Ministry of Justice and Public Security (MJSP) - Initiatives to combat cybercrimes	The Ministry of Justice and Public Security, through the Federal Police and other agencies, works to combat cybercrimes, including those related to the dissemination of disinformation that constitutes a crime (eg, libel, defamation, incitement to violence). This includes investigations into the use of AI to create and spread illegal content.	mid	NA	in force
China	"Golden Shield Project" (Great Firewall of China)	A sophisticated system of internet censorship that blocks access to websites, social media platforms, and other online content deemed politically sensitive or harmful by the Chinese government. It includes technologies such as keyword filtering, IP blocking, and DNS poisoning.	high	1998	in force
China	Ministry of Industry and Information Technology (MIIT)	Oversees the telecommunications and internet industries, including the implementation of technical measures for internet control, such as the "Great Firewall".	high	2008	in force
China	Cyberspace Administration of China (CAC)	The primary internet regulator in China. It has vast powers to enforce censorship laws, issue regulations, and oversee the operations of online platforms. The CAC is a powerful and highly influential agency.	high	2014	in force

China	Social Credit System	A nationwide system that aims to monitor and assess the trustworthiness of individuals and businesses. While still under development, it includes aspects related to online behaviour, with potential penalties for spreading "rumours" or engaging in other undesirable online activities. This can be used to pressure conformity and discourage dissent.	high	2014	proposal
China	Cybersecurity Law of the People's Republic of China	This foundational law establishes a comprehensive framework for cybersecurity and data protection in China. It grants the government broad powers to monitor and control online activity, including requiring companies to store data within China, provide technical support to security agencies, and censor content deemed harmful to national security or social order. It imposes real-name registration requirements for internet users.	high	2017	in force
China	New Generation Artificial Intelligence Development Plan	Outlines China's ambition to become a global leader in AI by 2030. It emphasises the development of AI for economic growth, national security, and social governance. While it mentions ethical considerations, the focus is primarily on technological advancement and application, including in areas like surveillance and social control. The plan acknowledges the use of AI for "social guidance" and "maintaining social stability".	mid	2017	in force
China	Provisions on the Administration of Online Information Content (aka "Content Ecology	These regulations establish a detailed system for controlling online content. They prohibit a wide range of content, including "rumours" (which can encompass anything the	high	2019	in force

	Governance Provisions")	government deems false or harmful), "negative" information, and content that "endangers national security" or "disrupts social order." They require online platforms to actively monitor and censor content, promote "positive energy," and establish "credit systems" for users.			
China	"Clear and Bright" (Qinglang) Campaign	A series of campaigns launched by the CAC to "clean up" the internet by removing "harmful" content, cracking down on "online chaos," and promoting "positive energy." These campaigns often target celebrity fan culture, online gaming, and other forms of online entertainment, but also have implications for freedom of expression and the spread of information.	high	2021	in force
China	Ethical Norms for the New Generation Artificial Intelligence	Outlines principles such as improving well-being, promoting fairness, ensuring safety and controllability and respecting privacy. However, these norms are interpreted within the framework of Chinese law and the CCP's priorities.	low	2021	in force
China	Provisions on the Administration of Algorithmic Recommendations of Internet Information Services	These regulations target the use of algorithms by online platforms to recommend content. They require companies to ensure that algorithms promote "mainstream values", prevent the spread of "illegal and harmful information", and protect user rights. They also require transparency about how algorithms work and allow users to opt out of personalised recommendations. This is a direct attempt to control how AI shapes the information environment.	high	2022	in force

China	Provisions on the Administration of Internet User Account Information (2021, updated 2022)	These regulations tighten control over online accounts, requiring real-name registration and restricting the spread of "illegal and harmful information". They also require platforms to actively manage user accounts and take action against those that violate the rules.	high	2022	in force
China	Provisions on the Administration of Deep Synthesis of Internet Information Services	Specifically targets "deep synthesis" technologies, including deepfakes. Requires providers of deep synthesis services to obtain security assessments, verify user identities, and label AI-generated content to prevent it from being mistaken for real. This aims to prevent the use of deepfakes for disinformation or other purposes deemed harmful by the government.	high	2023	in force
China	Administrative Measures for Internet Information Services	A broad set of rules governing internet content providers, including requirements for licensing, content censorship, and user data management.	high	NA	in force
EU	Audio Visual Media Services Directive (AVMSD)	Although primarily focusing on traditional broadcasting, AVMSD was revised to include some provisions on video-sharing platforms, which have implications for content moderation, including some measures related to disinformation and harmful content.	mid	2007	in force
EU	European External Action Service (EEAS) - East StratCom Task Force	This task force within the EU's diplomatic service focuses on countering disinformation campaigns originating from Russia and other external actors.	low	2015	in force
EU	General Data Protection Regulation (GDPR)	While not specifically about disinformation, the GDPR's strict rules on personal data processing have implications for targeted advertising and profiling, which are often used in disinformation campaigns.	high	2016	in force

		It limits the ability of platforms to collect and use data for creating and spreading personalized disinformation.			
EU	Code of Practice on Disinformation	A voluntary code of conduct for online platforms, advertisers, and other actors in the digital advertising ecosystem. The strengthened 2022 Code includes commitments to demonetise disinformation, enhance transparency of political advertising, empower users to report disinformation, and cooperate with fact-checkers. It's now linked to the DSA, meaning non-compliance can lead to penalties under the DSA.	high	2018	in force
EU	Directive on Copyright in the Digital Single Market (Directive (EU) 2019/790)	Includes provisions (Article 17, now Article 17 of the consolidated version) that make online platforms liable for copyrighted content uploaded by their users, unless they take certain measures to prevent unauthorised uploads. This has indirect implications for disinformation, as copyrighted material can sometimes be used in misleading ways.	high	2019	in force
EU	European Democracy Action Plan (EDAP)	A broader strategy to strengthen democracy in the EU, including measures to tackle disinformation, promote media freedom and pluralism, and protect elections. It encompasses the Code of Practice on Disinformation and other initiatives.	low	2020	in force
EU	European Digital Media Observatory (EDMO)	A network of fact-checkers, researchers, and other stakeholders that monitors disinformation, conducts research, and supports the implementation of the Code of Practice on Disinformation.	low	2020	in force
EU	Digital Services Act (DSA)	A landmark regulation that sets out comprehensive rules	high	2022	in force

		for online platforms operating in the EU, including obligations to address illegal content, disinformation, and systemic risks. It requires large online platforms to assess and mitigate risks related to the spread of disinformation, conduct independent audits, and provide data access to researchers. Crucially, it defines "very large online platforms" (VLOPs) and subjects them to stricter rules.			
EU	Digital Markets Act (DMA)	A regulation aimed at preventing large online platforms ("gatekeepers") from engaging in anti-competitive practices. While not directly focused on disinformation, it can indirectly impact the spread of disinformation by promoting competition and preventing dominant platforms from unduly favouring their own services or suppressing alternative viewpoints.	high	2022	in force
EU	Artificial Intelligence Act (AI Act)	The world's first comprehensive legal framework for AI. It classifies AI systems based on risk, with high-risk systems (including those used in social media content moderation and recommender systems) subject to strict requirements, such as transparency, data governance, and human oversight. The AI Act directly addresses the risks of AI-generated disinformation, including deepfakes.	high	2024	in force
EU	Media and Information Literacy (MIL) Initiatives	The EU promotes MIL through various programs and funding initiatives, aiming to empower citizens to critically evaluate information and resist disinformation.	low	NA	in force

India	Information Technology Act (IT Act)	The primary law governing cyberspace in India. It addresses various cybercrimes and provides a framework for electronic governance. Key sections relevant to disinformation and content regulation include: - Section 66A (Struck Down):** Formerly criminalised sending "offensive messages" online, but was struck down by the Supreme Court in 2015 for being overly broad and violating freedom of speech. - Section 69: Allows the government to issue directions for blocking public access to any information through any computer resource in the interest of sovereignty and integrity of India, defence 1 of India, security of the State, friendly relations with foreign states, or public 2 order. This is the primary legal basis for website blocking. - Section 69A: Empowers the government to issue directions to intermediaries (including social media platforms) to block access to information in the interest of sovereignty, security, etc. Includes a procedure for blocking and a review committee. - Section 79: Provides "safe harbour" immunity to intermediaries for content posted by users, provided they comply with certain due diligence requirements and takedown notices from the government. The conditions for this immunity have been significantly tightened through the IT Rules (see below).	high	2000	in force
India	Indian Computer Emergency	The national agency for responding to cybersecurity	mid	2004	in force

	Response Team (CERT-In)	incidents. While not its primary focus, CERT-In also deals with incidents related to the spread of malware and phishing, which can be related to disinformation campaigns.			
India	Ministry of Electronics and Information Technology (MeitY)	The nodal ministry responsible for formulating policy and regulations related to information technology, including those related to online content and cybersecurity.	high	2016	in force
India	National Digital Communications Policy	A broader policy framework that aims to create a robust digital communications infrastructure and promote digital inclusion. It includes some general principles related to security and trust, which are relevant to the issue of disinformation.	mid	2018	in force
India	NITI Aayog's National Strategy for Artificial Intelligence (#AIforAll)	Outlines India's vision for developing and deploying AI, with a focus on leveraging AI for social good. It includes some discussion of ethical considerations, but it doesn't have a specific focus on addressing AI-driven disinformation. The strategy emphasises "AI for All," focusing on inclusive development.	low	2018	in force
India	Press Information Bureau (PIB) Fact Check Unit	A unit within the government's Press Information Bureau that aims to counter misinformation related to government policies and initiatives. It publishes fact-checks and clarifications on its website and social media channels.	low	2019	in force
India	Information Technology Rules	These rules, issued under the IT Act, significantly expand the obligations of intermediaries, including social media platforms. Key provisions include: - Due Diligence: Intermediaries must observe	high	2021	in force

		due diligence, including publishing rules and regulations, privacy policies, and user agreements. - Grievance Redressal: They must establish a grievance redressal mechanism and appoint a Grievance Officer. - Significant Social Media Intermediaries (SSMIs): Platforms with over 5 million users are designated as SSMIs and face additional obligations, including appointing a Chief Compliance Officer, a Nodal Contact Person, and a Resident Grievance Officer (all based in India). - Traceability: SSMIs are required to enable the identification of the "first originator" of information on their platform, a provision that has raised significant privacy concerns and is being legally challenged. - Content Takedown Timelines: Intermediaries must acknowledge takedown requests within 24 hours and act on them within specific timeframes (generally 36 hours, or 15 days for certain categories of content). - Automated Tools: The rules encourage the use of automated tools for proactive monitoring and removal of illegal content.			
Russia	Security Council of the Russian Federation	A powerful advisory body that plays a significant role in shaping Russia's security policies, including those related to information security and countering perceived threats from foreign information sources.	high	1992	in force
Russia	Roskomnadzor (Federal Service for Supervision of Communications,	Russia's media and internet regulator. It has broad powers to block websites, issue takedown orders, and	high	2008	in force

	Information Technology and Mass Media)	enforce other censorship laws. It plays a central role in controlling the information landscape in Russia.			
Russia	Law on Foreign Agents	Requires individuals and organizations receiving foreign funding and engaging in "political activity" to register as "foreign agents" and face onerous labelling and reporting requirements. This has been used to target media outlets, NGOs, and journalists critical of the government, effectively stigmatizing and limiting their ability to operate.	high	2012	in force
Russia	Law on Personal Data Localization	Requires companies that collect personal data of Russian citizens to store that data on servers located within Russia. This gives the Russian government greater access to user data and makes it easier to enforce other restrictive laws.	high	2014	in force
Russia	Law on Undesirable Organizations	Allows the government to ban foreign or international organizations deemed a threat to Russia's security or constitutional order. This law has been used to shut down organizations that promote democracy, human rights, or independent media, effectively restricting the flow of information from outside sources.	high	2015	in force
Russia	"Yarovaya Law" (Anti-Terrorism Law Amendments)	Requires telecommunications and internet providers to store user data (including metadata and content of communications) for extended periods and provide access to this data to security services without a court order. This greatly enhances the government's surveillance capabilities and has implications for privacy and freedom of expression.	high	2016	in force

Russia	Ministry of Digital Development, Communications and Mass Media	Oversees the implementation of government policies related to information technology, communications, and media, including those related to internet regulation and disinformation.	high	2018	in force
Russia	Sovereign Internet Law (Federal Law No. 90-FZ)	This law grants the Russian government broad powers to control and monitor internet traffic within Russia, including the ability to isolate the Russian internet (Runet) from the global internet. It mandates the installation of "technical means for countering threats" on telecommunications networks, allowing for deep packet inspection and centralised control over internet traffic. This is ostensibly for cybersecurity but enables censorship.	high	2019	in force
Russia	Fake News Law (Amendments to the Law "On Information, Information Technologies and Information Protection")	Criminalises the dissemination of "knowingly false socially significant information" that poses a threat to public safety or order. This law has been widely used to suppress dissent and independent journalism, particularly regarding criticism of the government or coverage of sensitive topics like the war in Ukraine. Fines and jail time are potential penalties.	high	2019	in force
Russia	Amendments to the Law "On Information, Information Technologies and Information Protection" regarding social networks	Requires social networks with over 500,000 daily users in Russia to open local offices and comply with Russian laws, including those on data localization and content removal. Failure to comply can lead to blocking or fines. This pressures foreign social media companies to censor content deemed objectionable by the Russian government.	high	2020	in force
South Africa	Cybercrimes Act, 2020	This Act criminalises various cybercrimes, including data	high	2020	in force

		breaches, hacking, and unlawful interception of data. While not specifically focused on disinformation, it has provisions that could be relevant to combating malicious use of data to create and spread false information. It specifically defines and criminalises malicious communications, including "data messages" that incite violence or damage to property, or that are threatening or inherently false and intended to cause mental harm.			
USA	Section 230 of the Communications Decency Act	This foundational law provides immunity to online platforms for content posted by their users. While not designed to address disinformation specifically, it is central to any discussion of social media content regulation. There have been numerous proposals to amend or reform Section 230 to address concerns about disinformation, but none have passed.	low	1996	in force
USA	Cybersecurity and Infrastructure Security Agency (CISA)	CISA, part of the Department of Homeland Security, plays a role in combating foreign disinformation campaigns that threaten national security. They provide resources and guidance to state and local election officials, as well as the private sector.	low	2018	in force
USA	H.R.6937 - Countering Online Harms Act	It requires the Federal Trade Commission to conduct a study on artificial intelligence, and for other purposes.	low	2020	proposal
USA	National Artificial Intelligence Initiative Office (NAIIO)	Coordinates AI research and policy across the federal government. Its work on AI safety, security, and trustworthiness includes addressing the use of AI to	mid	2021	in force

		create and spread disinformation.			
USA	H.R.6796 - Digital Services Oversight and Safety Act	A broad proposal modelled after the EU's Digital Services Act. It would establish a new federal agency to regulate online platforms, with powers to address systemic risks like the spread of disinformation, require transparency reports, and impose content moderation requirements.	high	2022	proposal
USA	S.3608 - Social Media NUDGE Act	This bill requires social media platforms that have more than 20 million monthly active users to implement certain content-agnostic interventions to address social media addiction and the amplification or prioritization of certain content. Would require the National Science Foundation and National Academies of Sciences, Engineering, and Medicine to research ways to modify social media platforms to reduce the spread of harmful content, including misinformation. The FTC would then be required to implement regulations based on this research.	high	2022	proposal
USA	NIST AI Risk Management Framework	While voluntary, this framework provides guidance for organizations (including social media companies) on identifying and managing risks associated with AI, including risks related to bias, fairness, and the spread of disinformation.	low	2022	in force
USA	Executive Order 14110 - Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence	This is the most comprehensive and recent Executive Order on AI. It specifically addresses the risks of AI-generated disinformation, directing agencies to develop standards for watermarking and authenticating content.	mid	2023	in force

		It also covers broader AI safety, security, privacy, and civil rights concerns that are relevant to social media platforms.			
USA	H.R.5586 - Deepfakes Accountability Act	Focuses on malicious deepfakes (realistic but fabricated audio/video). Relevant to social networks as a major distribution channel for deepfakes. Aims to deter creation and require disclosure of deepfake content.	high	2023	proposal
USA	S.2892 - Algorithmic Accountability Act	Would require companies (including social media platforms) to assess and report on the impacts of their algorithms, including those used for content ranking, recommendation, and moderation. This is directly relevant to how social networks amplify or suppress certain types of content, including disinformation.	high	2023	proposal
USA	S.486 - Honest Ads Act	Would extend political ad transparency rules (currently applied to broadcast and print media) to online platforms. Designed to combat disinformation in political advertising.	high	2023	proposal
USA	S.483 - Internet Platform Accountability and Consumer Transparency Act or the Internet PACT Act	Seeks to reform Section 230 and increase platform accountability for illegal content and activity, while also promoting transparency and due process for users. The impact on disinformation would depend on how "illegal content" is defined and enforced.	high	2023	proposal

Methodology

The methodology for reviewing public interventions on AI regulation and disinformation across different countries involved a systematic analysis of policy initiatives and interventions, distinguishing between those already in force and those proposed. The review was conducted by identifying and categorizing legislative measures, public organisations, programmes, and policies implemented by various countries. Each intervention was classified based on its regulatory nature, distinguishing between paternalistic policies, which focus on content and platform regulation, and non-restrictive policies, such as digital literacy initiatives. Additionally, temporal trends were analysed to observe the evolution of regulatory efforts over time. This approach provided a comparative perspective, allowing for the identification of patterns, gaps, and divergences in national strategies, ultimately informing the discussion on the need for a more coordinated global response.

T20 South Africa Convenors

The Institute for Global Dialogue (IGD)

The South African Institute of International Affairs (SAIIA)

The Institute for Pan-African Thought and Conversation (IPATC)

For publication enquiries, please contact t20@t20southafrica.org

Website: www.t20southafrica.org