

ISSUE BRIEF:



Plumbing the depths of deepfakes: what's at stake, what's to be done.

April 2025

PRIORITY 4: RESOURCE PRODUCED FOR G20 DEWG ON GENERATIVE-AI AND ITS EVOLVING ABILITY TO PRODUCE HIGH-QUALITY DEEPFAKES AT LOW COST

Priority 4: Equitable, Inclusive and Just AI

Plumbing the depths of deepfakes: what's at stake, what's to be done.

An Issue Brief for G20 Digital Economy Working Group April 2025

A knowledge resource produced in support of the G20 Digital Economy Working Group (DEWG) Deliverable 2.8.4: Workshop on generative AI and its evolving ability to produce high quality deep fakes a lower cost, and the impact on information integrity, and consideration of possible recommendations.

Knowledge partners: **RIA (South Africa), CETIC (Brazil)**



Synopsis

This Issue Brief for the South African G20 presidency brings a digital economy lens to contemporary considerations of fact-based information environments in the light of generative AI.¹ The brief builds on DEWG's work on information integrity done under the Brazilian Presidency in 2024. Deepfakes are conceptualised as those synthetic media products in which events or a person's likeness (particularly faces or voices) are rendered for purposes of deception, diminishing reputation or distraction from trustworthy content. The brief addresses the problem of how to develop policy responses within a fast-moving technical context, and how G20 countries can mitigate before escalation. It points especially to the value of foresight and research in detecting the instances and extent of deepfakes. The brief finally recommends how G20 members might address these issues.

Baseline assessments

Technologically manipulated information is referenced in the [DEWG 2024 Ministerial statement](#), and since in the [Global Digital Compact](#) and many other international outputs (see Appendix 1 for background). While debates exist about concepts, terminology and the impact of false information, consensus is emerging on several points.

First, to be effective the digital economy depends on an information system where there are significant elements of [content that meet the criteria](#) of accuracy, reliability and consistency (with some debate about the [additional criteria](#) of fidelity, safety, and transparency). Within this scope, the concept of information as a public good, [endorsed](#) by UNESCO Member States in the Windhoek+30 Declaration, highlights the necessity and added value of universal access to information. Many actors also agree with the UN's [expansive conception](#) that information integrity relies on fostering societal factors such as social trust and resilience; independent, free and pluralistic media; incentives for advertisers and tech companies, as well as data access and transparency.

Second, there are significant empirical steps being taken which show that the issues around Generative-AI, and deepfakes in particular, are attracting attention for their impact on social trust, human rights and economic relations. These cover matters such as women who are victims of sextortion (including [children](#)) and [related non-consensual sexual imagery](#), and consumers who are hoaxed by [AI-personalised "phishing" attacks](#) or deepfake-enabled [identity fraud](#), [biometric system](#) compromise and [financial scams](#). [Researchers](#) argue that Generative-AI upscales existing scam tactics and enables new types of scams. One [estimate](#) is that AI-enabled fraud losses will increase significantly due to Generative-AI, reaching \$40 billion annually by 2027. An [industry analysis](#) attributes \$1.2billion in losses from deepfake scams in 2024. More broadly, a global price tag of \$78 billion has [been assessed](#) for 2020 as regards false and misleading information in general. Attention has also been drawn to deepfake content affecting [individuals' credit scores](#), while there is also awareness of [gender](#) and [other](#) stereotypes in AI-generated content that has a bearing on equal economic and career opportunities. [Cybersecurity](#) and [personal intellectual property](#) can be weakened

¹ This updates an earlier version tabled for the 2nd DEWG meeting on 8-10 April 2025 and circulated for feedback thereafter. It further draws on workshop discussions at the 3rd DEWG meeting on 10-12 June 2025.

through deepfakes. The economic impacts of deepfakes are complemented by threats in other sectors, including in [higher education](#) where risks of scientific confusion may emerge.

In addition, a number of jurisdictions have already set out provisions concerning these matters. Examples are in [the US](#) concerning electoral law as well as “digital forgeries” in cases of [non-consensual sexual imagery](#), and in the [EU](#) including [in the AI Act](#). [China](#) is requiring [AI-generated content to be labelled](#), and Spain to be [considering stiff fines](#) for failure to label. In 2024, major corporations established the [Coalition for Content Provenance and Authenticity \(C2PA\)](#) to agree on standards, including metadata and labels, for adding “Content Credentials” to AI-generated content, and an [ISO standard](#) is in process. [Governance](#) debates now encompass the deepfakes issue.

Third, the definitions of “deepfake” and the technical plus semantic challenges of identifying such are not yet settled (see Appendix 2). But what is not disputed is the need for concrete assessment of these technologies and “deepfake” outputs however they may be interpreted in different jurisdictions. This assessment is necessary in order to have an evidence-based approach to policy responses to the challenges. It has [been observed](#) that there is relatively little empirical evidence to date on the actual prevalence and impacts of deepfakes and that what exists is also limited geographically and in other ways. The challenge here is the well-known [Collingridge dilemma](#) about developing policy and regulation to pre-empt possible harms to human rights, when hard evidence is still in an emerging stage (or for some countries, where access to evidence is constrained by high paywalls or by [proprietary controls](#) on access to relevant data sets). Pausing action on a policy can leave things too late to prevent or mitigate severe consequences, while precipitate interventions may have unintended consequences that inhibit benefits or cause unforeseen harms.

Building on the conundrum outlined above, several ways forward present themselves. One is to engage in foresight and scenario planning that uses human rights and safety risk assessments as tools for preparing agile responses to how trends actually unfold in practice. A complementary response is to develop effective research methodologies to collect evidence, including in real time, about deepfakes that are circulating, as well as evidence on the effectiveness of [counter-strategies in play](#), and to extrapolate trends from such samples. Bound up with this are the development of indicators and data.

While attention has increasingly been given, [for example by regulators](#), to the wide range of measures in play to address the actors in the “deepfakes” supply chain, the matter of strategies for preparedness and informed policy-response, as discussed below, have not been foregrounded.

Foresight, Scenario Planning and Human Rights Risk Assessments (HRIAs)

By systematically exploring possible future states, [scenario planning](#) offers a [structured](#) method for developing responsive approaches to specific deepfake proliferation and possible harm. This [method](#) (with [multiple variations](#)) can enable policymakers to develop adaptive regulatory frameworks that stay relevant even as technology advances, while avoiding both ill-considered reaction or regulatory lag. Instead, they enable informed [contingency planning](#).

Policy-makers can profitably conduct such [foresight exercises](#), drawing also on partnerships with academics and private sector experts. This can be complemented by requiring designated actors (eg. Generative-AI vendors and platform distribution companies) to undertake auditable scenario planning, which in turn can be linked to compulsory HRIAs ahead of key moments (such as elections or expected natural disasters). The advantage of delegating these tasks to the core service providers is to mobilise their close knowledge of technical developments and their ability to institute safeguards at early stages.

On the basis of scenarios, a [threat analysis](#) for anticipating weaponized deepfakes can assess the likely vectors of fake identities, fake information and fake behaviours. These can inform adversarial stress testing and [red-teaming](#). A [playbook can be developed](#) to respond to different types of crisis scenarios. [Refreshed scenarios](#) are recommended to account for changes in technology, and intrinsic lags in detection tools.

Effective scenario planning for deepfakes should incorporate cross-disciplinary expertise to identify plausible trajectories of technological development and deployment. Key variables for scenarios might include the accessibility of deepfake creation tools (including those that are open-sourced and [can more easily enable users to bypass](#) technical provenance and authenticity features, as well as the simple method of photographing an image to remove technical signals about AI provenance). Also important is to assess expectations around detection capabilities, governance mechanisms for the relevant digital service providers (including platforms), and changes in societal resilience. Scenarios can test the establishment of appropriate thresholds and convergent factors which can raise the likelihood and severity of risks.

[HRIAs](#) – whether conducted by state entities, digital companies and other actors – constitute a complementary methodology that could be adjusted to evaluate the real and potential effects of deepfakes through the lens of established international human rights frameworks. HRIAs offer particular value in this domain by examining how anticipated deepfake technologies and outputs might compromise rights to privacy, security, dignity, and economic participation across different demographic groups and national settings. They can also identify deepfakes intended to intimidate people (eg. women political actors) and harm their right to free expression. These kinds of assessments can also help identify the groups most susceptible to the negative impacts of deepfakes. By [assessing risks](#) in terms of likelihood and severity, clear prioritisation can be established about attention to be given to different scenarios with different deepfakes, where and when. Also arising from a human rights orientation is the need to grade possible responses and mitigations in terms of international standards. This involves designing responses that align with the “[three-part test](#)”: legality; necessity and proportionality; and legitimate purpose.

For enduring effectiveness, HRIAs should be conducted throughout the development of Generative-AI foundation models and applications. In part, this requires establishing clear expectations for developers to incorporate human rights considerations in design and testing stages. The same consideration applies to the platform companies through which deepfakes are distributed at scale. The [UN Guiding Principles on Business and Human Rights](#) can guide these kinds of practice. G20 members may conduct or incentivise HRIAs, as well as mandate

the same for entities developing Generative-AI models and applications, as well as for those that circulate content via social networks. HRIAs can be profitably done ahead of elections such as in the [UK](#) and [South Africa](#).

Combining scenario planning and human rights assessments may also prove valuable for early warnings, such as where deepfakes may present heightened risks in contexts of high distrust, [low digital literacy](#), limited credible journalism and fact-checking infrastructure, political polarization or economic stratification. Ongoing monitoring of actual outcomes against predicted scenarios and creating feedback loops may improve assessment quality over time.

As with monitoring the prevalence of deepfakes online, conducting scenario planning research and HRIAs at scale also requires access to data from the service providers concerned. Assessment of the value of such exercises is also a practice that can be commended.

Approaches to uncovering evidence for policy

Foresight and developing playbooks are important tools that can help manage the Collingridge dilemma. Another way to respond to deepfakes is through actual empirical research guided by appropriate methodologies.

A [literature review](#) published by RAND reports that deepfake videos are more likely than fake news articles to be rated as vivid, persuasive, and credible, although they are also less persuasive than anticipated. Training audiences can help protect audiences from deepfake harm, according to some research reviewed for the report. However, *generalising from laboratories and experiments in certain countries, to other contexts and countries is a fraught exercise*. National (and contextual) differences may profoundly influence how individuals engage with deepfake content, as they relate to aspects such as trust in media outlets, individuals' media and digital literacy, prior exposure to specific facts or manipulated information, and frequency of access to each digital platform. There are also important challenges regarding confirmation and social desirability [biases](#).

Research on Generative-AI deepfakes also needs to be aware of coverage issues related to online samples, as these sometimes underrepresent Internet users with low-quality connections, fewer devices and less frequent usage. News stories may give the impression that the challenges are more widespread and influential than is actually the case. Hence, this section of this Issue Brief puts emphasis on research methods rather than findings. This focus correlates with three areas where G20 states can foster assessment of: a) the prevalence and dissemination of deepfakes; b) the impacts of this dissemination for audiences (investors, voters, consumers and vulnerable individuals who are targeted; and c) the effectiveness of mitigation interventions.

Assessing online patterns around deepfakes

Research requires effective methods to measure the prevalence of “deepfakes” online, with reference to reach, propagation velocity and algorithmic amplification, engagement and cross-platform presences as well as the differences between digital creation and distribution services. In one [study](#), the methodology of examining exponential growth, led to predicting eight million cases in 2025. Other features that can inform research methods are customisation to account for the technologies and thematics of deepfakes (including

gendered dimensions). Attention is further needed to the formats in which the content is presented (audio, video, text, etc.), and to possibly correlated variables – such as the existence or not of labels disclosing AI-generation. Detection lags – the time between publication and identification – are also significant to establish. Using watermarks for detection alone does not show the record of changes of the content, which depends then on whether there is meta-data showing provenance trajectories. However, the diverse range of watermarking and meta-data systems raise challenges for cross-platform and cross-content research.

Data about orchestrated online networks may show where commercial or other services are trafficking in deepfakes for various clients. The use of AI technologies for such analysis opens up possibilities for detecting patterns associated with deepfakes. Simulated “[Generative Adversarial Networks](#)” may be set up to develop specific decoders (also known as discriminators) for identifying where deepfake detection have gaps. [Multi-modal detection](#) systems, covering audio, text, video and still images, are needed for a comprehensive assessment of deepfakes.

Research and monitoring are needed where regulation requires actors such as political parties to disclose their use of Generative-AI. (Political parties in [Korea](#) and [Brazil](#), for example, have been targets of deepfakes and misinformation campaigns during elections). The effect of such regulations may require purposive sample methods as to the compliance levels visible online. Monitoring compliance with disclosure is a challenge, particularly in regard to detecting actors who seek to circumvent strictures. Here, the role of journalists and fact-checkers, using inter alia Open Source Intelligence Techniques, can be recognised as an important component of the ecosystem of detection. Companies providing services for AI content generation and/or distribution should be incentivised to have their systems and measures to do their own assessment of violations of their terms of service (including with regard to deepfakes in advertising content on social media platforms). Content distribution platforms using automated content moderation methods for deepfakes need to assess [under- and over-enforcement](#), as well as false negatives and positives. There is a need to independently research and [compare](#) how different platforms treat the same deepfake content in terms of detection, signalling and application of their own content policies.

Access to data from digital platforms and Generative-AI companies

For stakeholders outside the companies themselves, such as regulators and academic researchers, who seek to monitor the prevalence and character of deepfakes, access to corporate data is essential. This includes access to data sets about what content is electronically signalled as AI-generated (as per C2PA commitments). Where channels used for content distribution involve hidden dimensions, such as messaging services and emails, this requires metadata access along with creative research methods such as ethnographic sampling.

Access to privacy-protective data sets through APIs or sandboxes is a constant topic of debate. International organisations are constantly encouraging social media platforms to open up their holdings and share specific data through dedicated infrastructures. However, there appears to be a growing trend towards [more opacity](#) in this sense, with some platforms [restricting](#) researchers’ access to this data. Africa and Asia do not have open API access (for

example, from TikTok and YouTube, which provide this to the EU and North America). Meta, which previously closed its Crowdtangle tool that was also open to journalists, has replaced it with a content library and API for accredited researchers only. These facilities do [not](#) provide datasets showing where the company, its users, or AI application builders have applied Generative AI labels.

These issues further point to the need to research deepfake regulatory and self-regulatory mechanisms, as key to improving evidence for more effective policy-making and evaluation.

Assessing impacts on audiences of deepfakes

Researchers can investigate how audiences interpret and react to deepfakes, assessing cognitive, emotive, attitudinal, and behavioural dimensions. Assessing public awareness of deepfakes is significant. One [cited study](#) in 2022 found that less than a third of consumers knew what these phenomena were. Public understanding of both the technical and semantic issues around provenance and authenticity of content is an important area of research. These kinds of studies can profitably be structured in terms of types of harm to different human rights (dignity, reputation, privacy, personal property, democracy, environment, culture, etc.).

Large-scale sample surveys with audiences have been widely used to assess individuals' [time online](#), [media diets](#), [news consumption](#), skill sets, and [perceptions](#), and triangulated with studies about public trust levels. In a context where access to data on digital informational practices is restricted, these have been contributing significantly to the production and monitoring of evidence-based policies.

Such research frequently relies upon innovative methodologies for investigating how individuals interact with false and/or misleading content. [OECD](#), for example, conducted a comparative study in 21 countries and with a total of 40,765 individuals surveyed in which the respondents were asked to distinguish between false or misleading claims and verified content. Respondents are then asked to assess the veracity of the claims, which led to an individual score. The data was analysed in the light of various factors, such as content origin (human- versus AI-generated) and type, sociodemographic variables, as well as individuals' self-confidence in identifying false information.

There are also many [studies](#) that use tests for identifying the impact of fact-checking and AI-generative labels upon audiences, which measures are sometimes incorrectly presented as a [silver bullet](#). In these experiments, individuals are exposed to versions of the same content with different labels (i.e. "true" or "false"), and then asked to assess aspects such as the veracity of the claims and if (and how) they would interact with the content

Large-scale surveys as well as experimental studies can be complemented by qualitative research focused on specific audiences. Ethnographic and interview research into the producers of such deepfakes is also advised, which can help to address motivations and objectives. This triangulates with the complex methodological task of inferring the purpose of this kind of content.

Information Integrity in Brazil – producing research on informational dynamics with audiences

Measuring the informational dynamics of contemporary societies is a substantial task. In this context, the development of systematic, internationally comparable, and methodological innovative surveys with audiences plays a crucial role, aiding stakeholders in filling the evidence gaps in a particularly complex (by its subjective and multidimensional nature) and rapidly changing field.

In Brazil, the Regional Centre for Studies on the Development of the Information Society (Cetic.br|NIC.br) has been working toward a national-scale survey on the topic, leaning on the information integrity perspective encompassed by many policymakers and regulators in recent international documents. In this perspective, research and regulatory efforts assume a more holistic approach, looking, beyond countering information manipulation and hate speech, to promote healthy (diverse, reliable, consistent, evidence-based, and accurate) informational ecosystems and individuals' resilience toward informational harms. Thus, Cetic.br|NIC.br's research addresses various topics of investigation, such as audiences' information consumption/sharing practices, confidence in national media outlets and other sources of information, and skills and practices for identifying and dealing with false and misleading information (media/algorithmic/digital literacy), as well as their perspectives on the topic, with previous international surveys on the field serving as references for the questionnaire.

Cetic.br|NIC.br's [ICT Panel](#) has a specific methodology for qualifying its coverage and representativeness, in particular by estimating the coverage error of the online sample and therefore minimising the selection biases associated with non-probabilistic surveys. This is done by building a weight structure based on the [ICT Households](#) probabilistic survey to assess the propensity for each respondent to be an Internet user, according to socio-economic variables, which leads to a propensity score. By comparing both the Panel's and ICT Household's propensity scores, it is possible to estimate which part of the ICT Household's population (or if all the population) could be represented by the surveyed respondent, thus estimating the coverage error regarding the target population.

Large-scale survey inputs can be critical to the field, especially through innovative methodological approaches, and therefore can be expanded so that different aspects of deepfake circulations can be properly analysed, enhancing monitoring and evaluation of policies.

Other research experiences, often experimental in design and using A/B testing techniques, investigate the differences between engagement with and/or circulation of deepfake [videos](#) in comparison to text- or audio-based false or misleading [content](#). These seek to assess how different formats of information impact the [trust](#) in the content presented, individuals' evaluation of persuasiveness, emotional responses, and willingness to [share it](#). Further insights are gained from case studies of [high-profile incidents](#), which focus in upon a [range of dimensions](#) (ideally covering deepfake production technology, actors, motivations, output character and dissemination, reception and regulation)

Assessing the effectiveness of attempted mitigations: technical and other counter-measures

Detection tools for identifying deepfakes are constantly in need of review due to their [vulnerabilities](#). Some [research](#) identifies machine learning algorithms as flagging 92% of synthetic content via biometric anomalies, such as irregular heartbeats in videos (although deepfakes include events, not only humans). Even if detection accuracy keeps improving, [one study](#) says that 68% of deepfakes continue to be indistinguishable from authentic content. Because of continuous improvement in how Generative-AI renders voice patterns and facial expressions, and physical events, there is a need for ongoing assessment of detection technologies. Further research is needed into the datasets upon which detectors are trained, as narrow foundations (e.g. deepfakes of celebrities) may perform sub-optimally for other cases. Experimental research can explore options like blockchain for content verification.

There is also a need for studies looking at the effects of [possible interventions](#) such as labelling of media and information literacy programs, including [AI literacy](#). Eye-tracking studies can assess whether labels are noticed, and if there is impact on processing such content. Longitudinal studies may be valuable as the deepfake phenomenon evolves. According to [one study](#), there is mixed efficacy in countermeasures that help people detect fake AI-generated content, such as warning labels and watermarking, but the situation keeps evolving and requires continued investigation.

Journalists are a stakeholder group that can mitigate deepfakes by exposing deceptiveness, but their abilities [also need](#) to be assessed. A role-play simulation that assessed reporters' use of various detection tools, found many misjudgments as a result of over-reliance on flawed detectors. How journalists' [verification](#) practices keep pace with deepfakes merits attention.

Where regulation requires actors to disclose their use of Generative-AI (for example, political parties during elections), this also needs to be researched beyond the online realm and directly with engaging these stakeholder groups (e.g. concerning their maintenance of appropriate records).

Another line of research on deepfakes' impact links up with assessing educational [approaches](#) that impact on individuals' literacy about Generative-AI. A significant sub-area here is about [interventions](#) in schools where relevant knowledge [resources](#) are available, but there is little assessment of their uptake and impact. Some research relies on comparisons between broad media literacy efforts with those that are focused on deepfake video [content](#). Pilot [studies](#) have shown the effectiveness of training programs in increasing people's ability to spot deepfakes (for example, by drawing their attention to look for unnatural eye movements). But technology advances make for 'moving targets' and the obsolescence of easy detection tips.

Coalitions such as the [Partnership on AI](#) and the C2PA can also be [profitably assessed](#) with a view to strengthening their effectiveness. The use of AI tools to counter deepfakes and fraudulent behaviours further merits assessment.

Conclusion and recommendations

Holistic assessment is merited because individual mitigation strategies may have limited effectiveness. For example, counting primarily on resilience and media and information literacy is seen as placing disproportionate burden on individuals, and exempting upstream

actors in the “deepfake” supply chain from responsibility. A balanced and multilayered mix – all of which merits foresight and research - would include a combination of foresight, research, technical detection, updating regulatory institutions and co-operation, enforcement of law and of tech company ethics and policies, as well as initiatives for media and information literacy.

This Issue Brief intersects with many other points in the G20 and the DEWG, including data governance. It has relevance as well for intergovernmental organisations such as the African Union, and for private sector actors such as the C2PA. It further touches on the roles of information and privacy regulators, electoral regulators, media organisations, educators and civil society groups. Ministers responsible for the digital economy can promote [coordination efforts](#) and task forces specifically focused on deepfakes and their economic impacts. They can champion data sharing agreements, including across borders, for research. Other possibilities are to help harmonise national technical standards for content provenance, and to drive regulatory coherence. This can include cross-jurisdictional networking which can in turn inform regular high-level dialogues and knowledge exchange.

Cross-cutting recommendations, appropriate as applicable, include:

- Support for investment in research, monitoring and innovation in technical detection capabilities
- Adopting measures to ensure scenario planning and requirements for digital firms to human rights risk assessments for deepfakes ahead of sensitive contexts such as elections and crises.
- Developing guidelines and transparency/disclosure protocols for state actors about their use of Generative-AI, and also adopting conditions to selected actors, such as political parties during elections.
- Promoting comprehensive data (and metadata) access from the digital companies in the chain of deepfakes, so as to enable public-interest independent monitoring.
- Fostering societal media and information literacy on deepfakes.

Appendix 1: Background

The [DEWG Ministerial statement in September 2024](#) gave attention to “Integrity of information online and trust in the digital economy”. This statement recognised dramatic impacts on the speed, scale and reach of factually deceptive content. It noted that the absence of reliable, diverse, accurate information and knowledge may negatively affect trust in the digital economy, public institutions and governance and democratic processes.

The DEWG Ministers went on to advise that AI systems as applied to content should be ethical, transparent, auditable, accountable and compliant with legal frameworks such as data protection, human rights and intellectual property. They affirmed that “content authentication and provenance mechanisms and related technical standards may help identify AI-generated content, and enable users to identify information manipulation”. The Ministers further encouraged cooperation and information sharing on initiatives and best practices. A DEWG side event was held, at which host country Brazil announced the [Global Initiative for Information Integrity on Climate Change](#).

Since the 2024 DEWG statement, reference to “information integrity” has appeared in the [Global Digital Compact](#) (GDC), which makes the point that:

- Access to relevant, reliable and accurate information and knowledge is essential for an inclusive, open, safe and secure digital space
- AI-manipulated information can be harmful to societies, human rights and fundamental freedoms, and can negatively impact the attainment of sustainable development goals (SDGs)
- There should be strengthened governance cooperation at all levels to address the challenges of misinformation and disinformation
- Mitigating the risks of information manipulation should be consistent with international law.

The [Global Declaration on information integrity online](#), signed by 35 countries, warns of that Generative AI risks becoming “an enabling tool to spread disinformation at a far greater speed and scale than ever before”. The OECD has also produced [a recommendation](#) on information integrity, and several G20 countries remain seized with the issues, not least the challenge of deepfakes in elections. The UK’s regulator Ofcom has released a [discussion paper](#) about “deepfakes that demean, defraud and disinform”. There have been some criticisms of the terms “misinformation” and “disinformation”, but beyond the terminology, there are widely shared concerns about the [damage of deepfakes](#). The Council of Europe [Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law](#) commits parties to adopting transparency measures including as regards the identification of content generated by AI systems.

While there have been these advances within the multilateral system and some national states on safeguarding the integrity of information, self-regulation by platforms that had advanced over the past few years, has retreated. Both Meta and YouTube have walked back some of their content moderation policies and employee numbers, following the example of

X. The trend is in favour of delegating signals about the quality of content on their platforms to users via a system of “community notes” rather than professional and independent fact-checking. It is widely expected that these alterations will result in a greater volume and range of falsehoods to circulate. A [resolution](#) by the African Commission on Human and Peoples’ Rights in March 2025 stated that “community notes” are not an alternative to the companies’ corporate responsibilities nor a substitute for independent fact-checking. It added that “community notes” are “susceptible to be captured by forces that do not respect human rights”. In parallel with these developments, financial frauds and scams are becoming [ever more sophisticated](#) (even impersonating a country presidents), while Generative-AI applications and underlying foundation models continue to spread at reduced cost and owing to the growing ease of use of these technologies. Meanwhile, many users of anthropomorphised AI chat interfaces behave as if they are interacting with an actual human persona, thereby further diluting capacity to distinguish authenticity from engineered artifact. The circulation of convincing content with the particular purpose of deception can have a grave impact on trust, cohesion, electoral integrity and digital transactions. As such, the destabilising effects of deepfakes within the information environment present considerable risk to both digital and extra-digital economic life. Anticipating risks and assessing trends in this space, at the national level, as well as through international cooperation, can help enable G20 members to anticipate and develop evidence-based mitigations for deepfakes.

Appendix 2: Understanding definitions and debates about “deepfakes”

Definitions in play:

The concept of “deepfakes” is not definitively established in the literature, nor in law and regulation. The scope in some cases is used primarily to refer to visual imagery (including video) and to a lesser extent audio. However, there is also the matter of whether to include synthetic [text](#) that “can flood the zone”. In some jurisdictions, “deepfakes” refers narrowly to content about people, although depictions of events and natural phenomena could also fall within the scope. Also implicated are broader company [policies on disinformation in AI chatbots](#), where interactive engagements are structured to appear as if responses are as real as those of an actual human interlocutor.

Many documents assume an obvious meaning to “deepfakes”, or tend to use the term interchangeably with “[manipulated media](#)”, “[synthetic media](#)”, “[fabricated media](#)” and “[digital impersonation](#)”. However, distinctions can be drawn between content that is created “de novo” (synthetic), and original ‘base’ content that is digitally altered (both of which functionalities are possible in most Generative-AI applications). The significance here is for detection: where there is no singular original, techniques like reverse-image search are of little use. Further, even where there is an original, in the contested play of meanings, some people have disingenuously claimed that it is the original versions that are forgeries (a variation of the “liars’ dividend”, discussed further below).

In some instances, the concept of deepfakes is tied to Generative-AI and not to other technology. There are approaches that use the term to designate [all AI-generated content](#), with some of it being designated as harmful while other “deepfakes” are seen as having positive uses (like in gaming, virtual reality, photography, entertainment, education and support for individuals who lose their voices). In yet additional cases, the terminology references only a subset of Generative-AI output, while for some observers “deepfakes” should be entirely technology agnostic.

These variations have implications for detection, guidance, regulation and other responses to the challenges arising.

Definitions for this Issue Brief:

For current purposes, the term “deepfake” encompasses both human and non-human objects. It is used for all formats, while recognising that image and sound can often suggest greater verisimilitude than text. At the same time, it is cognisant that the persona design and plausibility cues involved in text exchanges via chat interfaces, while generally lacking specific human intentionality, are such that the process can also persuasively resemble verisimilitude to the extent of overshadowing hallucinations and biases.

Attention in this Issue Brief is given to the significance of the word “fake” in “deepfakes” and its negative connotations. In turn, this points towards deception as a primary feature for the term “deepfake”. It therefore does not use semantics of both negative and positive fakes, but limits the term to the former, and to the purpose of misleading those receiving the content.

In [differing jurisdictions](#), the criteria for deception are whether the output is believable to “a reasonable member of the public”, while others focus on malicious intention – whether the creators sought for their content to be believable. Since intentions are often hard to identify and prove, it is often necessary to infer these, and to combine such induction with the issue of whether the content does in fact appear to be credible.

In regard to the distinct nature of Generative-AI from other AI systems, and how this can be managed with regards to limiting their use in the production of deepfakes, the deceptive significance of this kind of content should not be conflated with the underlying technology. Although there are cases where X-AI and its application Grok appear to have been engineered to apply ideological perspectives, in general Generative-AI is not deceitful in the sense of intentionally manipulating people’s beliefs in what is true and what is not. Deceptive (in the sense of fraudulent) content can be produced without resort to Generative-AI. Further, very much Generative-AI content is not deceptive (i.e. is not presented as if it were a valid representation of the historically real).

Other intentions may underpin deepfakes, such as being not so much to present purported realities, but rather as weapons in order to violate people’s dignity and reputation (as in non-consensual sexual imagery), or simply as tools to create confusion about what can be believed and thereby weaken credible and authoritative content.

The technology dimension:

Keeping in mind the points made above, it is nevertheless the case that technology is also not irrelevant in terms of the power of producing deceptive content. It is also not marginal to the issue of detecting digitally processed content which is deceptive. Illustrating the historic significance of technology, “deepfakes” have often been designated in contraposition to “cheap fakes”. However, this distinction is less an absolute one than being about variations along a [spectrum](#). Furthermore, given that the distinction is also based on the quality and cost of the representations involved, it is increasingly fading as ever more powerful Generative AI technology becomes ever more available for uptake.

The concept of “deepfakes” is also sometimes counterposed to “shallowfakes” which refer to deceptive content produced by digital editing that is not AI-based, and which may nevertheless be plausible to an uninformed audience. However, more and more, digital editing technologies are embedding AI components in their functions, and Generative-AI itself is becoming more ubiquitous. This is why, in practice, there is a logic to putting attention on AI-generated content in general – which is a subset of all digitally produced content. This exponentially growing subset provides, in practice, the universe within which – going ahead – most deepfakes will increasingly be found.

Technical provenance and authenticity:

The terms “[Builders](#)” and “[Creators](#)” are sometimes used to refer to those who develop the technical tools, and those who apply them to produce content outputs. Both functions are upstream from the [distribution of deepfakes](#) which is mainly via social media platforms. The main players in Building, as well as Device manufacturers and Distributors, [committed in 2024](#) to adding signals to AI-generated content).

For the Creators of deepfake content, relying on specific devices as well as specific tools based on “Builders” foundation models, it follows that interventions at such base technical level could set parameters on possible use. Limits on “prompt-hacking” to prevent deepfake sexual imagery from being made are one example. Another is embedding – and then [detecting, interpreting and acting upon](#) - watermarked or metadata signals in devices and/or tools which will carry through in the outputs. In this regard, various actors have elaborated on the concepts of “authenticity” and “provenance”. These can be understood in two ways – technically and semantically.

Purely in terms of technical dimensions (as per the C2PA), the term “provenance” involves laying down a detailed history of content creation and direct modifications. Examples of such metadata are: source-related (device type, geolocation); modification records (time stamps, tools used); and ingredients (source materials that have fed composite outputs). “Authenticity” in technical terms covers the functionality whereby provenance metadata has not been tampered with or removed from a content item (called “cryptographic binding”) and/or where it includes digital signatures in the content (eg. “BBC”). Digital provenance thus designates a technical record, while authenticity is about the technical integrity of that record. A host of specific [technological measures](#) can be interpreted in this framework. The “how” AI-techniques have been applied can be of value in alerting users as to whether there is minor editing involved versus the fully-fledged extrusion of synthetic content.

Semantic provenance and authenticity:

These technical dimensions do not cover issues of content truthfulness or deception, which are epistemological questions about what distinguishes “fake” content from “authentic” content. Signalling the content in terms of the technology dimensions discussed above can only identify if the content is “manipulated” (as distinct from being a “raw” original representation). At the same time, such flags (even if manifested in warning labels) do not, on their own, communicate whether the output is produced or circulated with intent to deceive. Not surprisingly, false negatives and positives are endemic to any conflation of Generative AI outputs with deepfakes as such. This is where semantic interpretations of “provenance” and “authenticity” come in.

Much digitally processed content with a high degree of realism can nevertheless be a recognisable as a product of imagination or educational material, or as satire and caricature. Accordingly, adding technical signals of provenance and authenticity can only tell platforms/viewers/listeners/readers that such content *may* be deceptive even when it is not. Such disclosure about the role of generative-AI may also become of diminishing value as more and more content qualifies to have such signals attached and where user fatigue with labels sets in.

In regard specifically to content proven to be a “deepfake” (whether produced with or without Generative AI), warning labels in such cases may also serve to reduce trust in all content across the board. These matters point to the importance of research and monitoring to assess the impact of labels, and of other interventions, aimed at countering deepfakes, including unintended impacts.

In a semantic approach to “provenance”, it is possible to use less technically intrinsic methods to gauge apparent identities behind the actors creating the content and/or distributing digital content. The focus here is on whether content has an identifiable source who is authentically such. Verification methods in this regard are not always foolproof. For example, the [EU Commission](#) has found that the sale of “blue tick” marks on X without rigorous validity checks means these no longer signify authenticity and instead can enable scams and [phishing](#) attacks. Some digital signatures may also be susceptible to being excised or forged or made misleading due to other content being added to the artifact at hand.

A further complication is that it is also the case that an upfront and genuinely identified source is still capable of creating and disseminating deceptive content in a deepfake. Nevertheless, transparency about who really is involved is not without value for information integrity. State and political agencies for example, could be held accountable in most contexts if they were to use false identities to purvey deepfakes. Officials can also improve the information environment by practicing maximum transparency through optimum [public communications](#) and applying technical standards when availing authoritative content and data under proactive disclosure provisions within Right to Information laws.

Part of the challenge of “deepfakes” in regard to whether the originators and actors therein are authentic or not is that these artefacts can be constructed to serve as imposters of genuine actors, or to appear as if they are in fact real humans (and not replicas or synthetic creations). Clues can be gleaned about Generative-AI use for [coordinated influence operations](#), as well as in the context of behaviours when such operations target social media distribution platforms. In these, [metadata gives context to data](#) such as in showing behaviours around similarities in social media identity/account creation, automation of posts, nodes of leaders and peripheries of “buzzers” who amplify the fabrications. Apparently plausible convincing content linked to these behaviours may merit closer scrutiny to see if deepfakes are being deployed. The right to privacy becomes implicated where Generative-AI is used to customise and micro-target deepfakes for particular audiences and even individual targets, which in turn becomes a complex matter to detect at scale.

[Context annotations](#) are a semantic attribution measure that can help signal deepfakes, even where technical measures fail. These involve additional information linked to the content at hand, produced by fact-checkers, approved users, or ordinary users. They depend on large part on the affordances of platforms as to whether and how they appear, a matter which merits more detailed research.

Further controversies:

Where actors are encouraged or required to provide disclaimers or other labels concerning digitally-manipulated content, there are diverse rules about the wording, size, placement and frequency of these signals (including for periodic notifications in audio). There is a policy and technical challenge for distribution platforms to identify and apply labels where contributions do not include embedded “fingerprints” or creator-applied labels. Companies such as Facebook have held a “Deepfake Challenge Competition”, while at the same time some observers fear an “arms race” where publicly available detectors will be used by [adversaries](#) to build undetectable deepfakes. A diversity of governance responses is evident in the April

2024 UN General Assembly [resolution](#) saying that watermarking or labelling should be done where “technically feasible and appropriate”.² There are initiatives by platforms such as [TikTok](#) to require users to use the flag “Creator labeled as AI-generated” when using generative AI for their content on the platform, while the [Content Authenticity Initiative](#) enables users to embed watermarks for their content across all platforms.

Debates exist whether the incorporation of labels on AI-generated content may further support a “[liars’ dividend](#)” in that undetected and unlabelled deepfakes may, by implication, be taken as being non-deceptive by default. There is also the scenario of enabling lying about the authenticity of [real content](#). Some [research](#) suggests that deepfakes could reduce trust in the accuracy of professional journalism.

This scenario reflects the potential for actors to use tools or techniques that deliberately, or unintentionally, bypass C2PA-style embedded metadata and labelling practices by generative AI providers, platform distribution services and even users. Even with widespread adoption and implementation of these practices, there will be actors with interests in avoiding such identifiers. This means that relying solely on C2PA watermarking or cryptographic data will be insufficient. A combination of approaches is needed for more comprehensive detection. However, a further [consideration](#) is the extent to which detection efforts may blur into mass surveillance and thereby jeopardise personal information and privacy.

A question this raises is the extent of value to the information environment of signalling provenance and authenticity in both technical and semantic senses. The debate is whether it is better than not doing so, or whether it has unintended consequences that outweigh the benefits, and if there are priority contexts (such as of deepfake deployment in elections or in widespread financial scams) which are more important to address than others. This issue would seem to call out for credible identification and ranking of potential dangers, and the prioritization of signalling responses in a nuanced manner, ranging from high alert labels through to detailed rebuttals by credible leaders.

A further matter concerns penalties. In some cases, “deepfakes” already fall under existing legal regimes concerning consumer rights, voter rights, and financial regulations. However, these provisions may require amendment, and in other cases (e.g. non-consensual sexual imagery), new regulations may be needed. The criteria of necessity and proportionality need to inform penalties (criminal and civil) as well as redress/relief measures (e.g. extending to removal or blocking of the content concerned).

Developing coordinated responses to the challenges requires institutional initiatives. In the US, the [Deepfakes Accountability Act](#) proposed a task force to investigate national security implications, and facilitate cooperation with private sector technology enterprises and academic and research institutions. Because deepfakes can affect all realms of society, the

² “Encouraging the development and deployment of effective, accessible, adaptable, internationally interoperable technical tools, standards or practices, including reliable content authentication and provenance mechanisms – such as watermarking or labelling, where technically feasible and appropriate, that enable users to identify information manipulation, distinguish or determine the origins of authentic digital content and artificial intelligence-generated or manipulated digital content – and increasing media and information literacy”.

challenge is to develop at least a whole-of-government approach to underpin the range of governance requirements in particular priority sectors and targets (eg. [women](#), migrants) as each is affected by deceptive content.

It is worth recalling that Generative-AI is a technology that - irrespective of human intent - can output content that is [ostensibly plausible](#), but which is not correct.³ This gives added resonance to calls for disclosure by designated actors (particularly state bodies) whenever such technology is used.⁴

³ [AI Models Provide Inaccurate Information to Voters with Disabilities; How AI chatbots responded to questions about the 2024 UK election](#) Another [study](#) finds similar problems.

⁴ As an example of convincing output, which upon investigation turned out to be fabricated, one AI chatbot yielded many fake references when asked for studies about deepfakes. One example it produced: 'Chesney, R., & Citron, D. (2022). "Measurement frameworks for synthetic media prevalence: A standardized reporting protocol." *Stanford Technology Law Review*, 25(2), 189-231'. Queried on its production of hallucinated citations, the chatbot responded: "I'm creating plausible reference formats based on my training data that would match the type of information being discussed." Similar phenomena have been noted [here](#) and [here](#).