Centre for International
Governance Innovation

Digital
Policy Hub

Digital Policy Hub – Working Paper

Artificial Realities: Mitigations against Deepfakes

Ozan Ayata

Winter 2024 cohort

About the Hub

The Digital Policy Hub at CIGI is a collaborative space for emerging scholars and innovative thinkers from the social, natural and applied sciences. It provides opportunities for undergraduate and graduate students and post-doctoral and visiting fellows to share and develop research on the rapid evolution and governance of transformative technologies. The Hub is founded on transdisciplinary approaches that seek to increase understanding of the socio-economic and technological impacts of digitalization and improve the quality and relevance of related research. Core research areas include data, economy and society; artificial intelligence; outer space; digitalization, security and democracy; and the environment and natural resources.

The Digital Policy Hub working papers are the product of research related to the Hub's identified themes prepared by participants during their fellowship.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

Partners

Thank you to Mitacs for its partnership and support of Digital Policy Hub fellows through the Accelerate program. We would also like to acknowledge the many universities, governments and private sector partners for their involvement allowing CIGI to offer this holistic research environment.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Key Points

- The emergence of deepfakes, facilitated by advancements in artificial intelligence (AI), introduces new means for disseminating mis- and disinformation. These emerging methods of spreading mis- and disinformation exploit the broad accessibility of deepfake generation tools and pose a compounded challenge when considering the rapid rate at which information travels and the growing reliance on social media for news.
- This working paper delves into strategies to mitigate the negative impacts of deepfakes while simultaneously highlighting the measures that various actors, including states, social media platforms and tech companies, are taking to tackle this issue. Deepfake mitigations span across four different areas: legislative and regulatory measures, social media policies, technological tools and societal resilience.
- The paper's final section proposes policy recommendations covering these four categories. Central to these recommendations is the recognition that addressing deepfakes necessitates a collaborative approach, wherein multiple sectors work together synergistically to craft a robust response.

Background: The Rise of Deepfakes

Initially intended to serve as places to build social networks, social media platforms have evolved to swiftly disseminate information and become a primary location of news consumption (Saifuddin 2023). The increasing use of social media platforms for news, however, makes individuals more vulnerable to believing in mis- and disinformation. Emphasizing this point, a study conducted in 2021 found that individuals who heavily rely on social media platforms for news are 1.4 times more likely to believe false statements compared to those who seek information from alternative sources (Nightingale and Farid 2021). This dilemma is further exacerbated by recent developments in AI that enable the creation of various forms of content beyond conventional text. AI-generated forms of content have been termed "deepfakes," a term coined on Reddit in 2017 that blends deep learning with fakes (Lewis and Nelson 2019). Employing AI and deep-learning algorithms, deepfakes produce realistic yet fake audio, image or video content (Canadian Security Intelligence Service 2023). While historical examples of media manipulation exist (such as photography), the game-changing aspect of deepfakes is the ease with which anyone, even those with minimal technical skills, can produce AI-generated content, which can then be rapidly disseminated on the very platforms people rely on for news (Farid 2021). This ease of use highlights the ongoing urgency for researchers to analyze a range of strategies aimed at mitigating the dangers posed by deepfakes.

A Technical Overview

Before delving into mitigations, it is crucial to outline the technical underpinnings of deepfakes. Inspired by the biological structure of the human brain, artificial neural networks serve as a fundamental component of deep-learning algorithms (Carter and Manley 2020). An artificial neural network's architecture consists of input, hidden and output layers with interconnected neurons. The connectivity of neurons is determined by weights, with positive weights facilitating activation and negative weights suppressing it.[1] The positive weights trigger the activation of neurons from one layer to the next until they achieve their desired output. Artificial neural networks are trained with sample data and learn via a cost function that identifies incorrect outputs and instructs necessary adjustments, which, over successive iterations, should make the model more accurate at achieving its intended purpose (Meel, n.d.).

Deepfakes, as noted before, can be utilized to create photo, audio and video content. This paper will exemplify the process using photos, given their prevalence in the literature.

Deepfake photos are generated using general adversarial networks, which consist of two competing artificial neural networks. The first network, known as the generator, attempts to craft fabricated images, while the second network, the discriminator, distinguishes between real and fake images using a legitimate data set to guide its judgment (Byman et al. 2023). In other words, the discriminator evaluates the authenticity of each successive iteration that the generator produces. As each round passes, the generator ameliorates its previous discrepancies and produces more realistic photographs to deceive the discriminator. The process concludes when the discriminator can no longer distinguish the generated photo from its original data set (Sayler and Harris 2023). Initially, this technique was highly specialized and required lots of data; however, deepfake creation tools are now readily available on the internet, and photographs can be produced using a single image (van der Sloot and Wagensveld 2022).

The Spectrum of Deepfake Applications

Speculating on the myriad ways in which deepfakes can be deployed has garnered significant research attention (Chesney and Citron 2019a). Consequently, this paper attempts to avoid repetition in this domain and instead provides a high-level overview to help contextualize subsequent sections. For simplification, the adverse applications of deepfakes can be segmented into four units of analysis: individual, organizational, societal and international. While not mutually exclusive, this categorization is a simplified way of understanding deepfake applications.

The most prominent use of deepfakes at the individual level is non-consensual explicit material. This category of content comprises nearly 95 percent of all deepfakes produced and disproportionately targets women (Moreau and Rourke 2024). Wrongdoers can employ deepfakes for the purposes of harassment, blackmail, intimidation and many other nefarious reasons. Importantly, deepfakes can impact people regardless of their location or status; victims can be individuals a wrongdoer knows personally or strangers (such as celebrities) whom they have never met. Mitigations such as legislative measures are hence vital to protect victims and criminalize the production

1 See https://aws.amazon.com/what-is/neural-network/.

and dissemination of non-consensual explicit deepfake material. At the organizational level, various recent examples illustrate the enduring threat that fraud poses. For instance, in February 2024, fraudsters convinced an employee of a company to participate in a fabricated video conference, which included a deepfake portrayal of the chief financial officer (CFO). Convinced it was the real CFO, the employee transferred $25 million to the fraudsters at the request of the deepfakes in the conference (Chen and Magramo 2024). At the societal level, deepfakes have the potential to erode and destabilize the cohesion of a populace. A well-timed deepfake, which exploits a controversial issue, can deepen fractures and compromise agreements among different factions (Chesney and Citron 2019a). For example, if deepfaked simulations of police brutality are disseminated during periods of protests against actual instances of misconduct, existing grievances could deepen (Byman et al. 2023). The international unit of analysis encompasses two or more states and relates to claims that deepfakes could legitimize wars and uprisings (ibid.). Deepfakes could also potentially provoke both state and non-state actors into initiating or reacting with hard power, although the threshold for initiating armed conflict would in all likelihood require empirical evidence, not just alleged content (Chesney and Citron 2019b). As observed in the Russo-Ukrainian war, deepfakes have been more commonly used to sow fear, mistrust and uncertainty during a conflict rather than to directly incite violence (Kleemann 2023).

Thus far, all of the examples provided have focused on deepfakes presenting AI-generated content as real occurrences. Conversely, a counter scenario exists where people are manipulated into denying the authenticity of a legitimate event. This phenomenon, known as the "liar's dividend," could emerge as the by-product of the public's increasing awareness of deepfakes: as the awareness of deepfakes spreads, people will likely grow more skeptical about the legitimacy of videos. Consequently, the public will become more open to assertions alleging that genuine content was fabricated, which ultimately introduces a new tactic for wrongdoers to cast doubt on the credibility of evidence (Chesney and Citron 2019a).

Addressing the Deepfake Threat: Possible Mitigations and Responses

Given that deepfakes can be used for any number of nefarious reasons, the crux of this paper is to examine strategies to mitigate these impacts. Doing so also sheds light on what the actors in this space (that is, states, social media platforms and tech companies) are doing to address this need. Deepfake mitigations span four areas: legislative and regulatory measures, social media policies, technological tools and societal resilience. The following sections analyze the advantages and limitations of each of these areas, culminating in policy recommendations at the conclusion of this paper.

Legislative and Regulatory Strategies

Governments worldwide acknowledge the challenges posed by deepfakes, and some are actively working to confront this dynamic space. For example, the United States has taken a positive first step by introducing the DEEPFAKES Accountability Act at the federal level in September 2023. This act proposes three

key measures: first, it requires deepfake content to include a text box indicating it is AI-generated as well as a digital watermark for source identification; second, it prohibits the impersonation of individuals in ways they would not identify; and third, it provides victims with avenues for legal recourse.[2]

Other legislative and regulatory actions focused on protecting individuals from harassment have been implemented at the state level in the United States. Several states, including California, Hawaii, New York and Virginia, have enacted laws criminalizing AI-generated non-consensual explicit material (Farid 2022). Expanding beyond protective measures and into deepfake mis- and disinformation, Texas made it illegal to create deepfakes intended to sway election outcomes in 2019.[3] Similarly, California now allows candidates for public office to take legal action against "individuals or organizations that create or share election-related deepfakes within 60 days of an election" (Department of Homeland Security 2021, 29).

Despite these advancements, it is important to recognize the limitations surrounding what legislative and regulatory means can achieve. The primary constraint of existing legislative and regulatory measures lies in the complexity of attributing a deepfake to its creator, which can be rather difficult (Chesney and Citron 2019a). Even if this obstacle is overcome, the second limitation will be that domestic laws only pertain to a specific legal jurisdiction. In an interconnected digital world, where wrongdoers may reside in different countries, legal action against a foreign actor is thus increasingly challenging (Iyengar 2021). Additionally, given its potential ramifications on freedom of speech, regulating deepfake mis- and disinformation requires serious consideration, and the extent to which states are willing to do so will also vary. For instance, in 2023 the European Union passed the Digital Services Act, which includes measures to penalize non-compliant social media platforms with bans or fines (O'Carroll 2023). By contrast, the United States, guided by the Communications Decency Act, provides platforms with immunity from being liable for harmful content (Chesney and Citron 2019a). While one approach is not necessarily more favourable than another, each country will have to operate within the parameters of its unique regulatory appetite.

Leveraging the Role of Social Media Platforms

Social media platforms possess exceptional capabilities to adopt changes to mitigate the harmful effects of deepfakes, given that they control the environment where deepfakes are disseminated. Thus, evaluating their policies and advancements to date is crucial.

Meta has stood out as a forerunner in comparison to other social media platforms: it has not only worked to remove deepfakes but also collaborated with third-party fact-checkers to identify misleading content (Bickert 2020). When deepfakes are detected, Meta represses distribution and inserts a warning for users who encounter them on their feed. Meta also launched the Deepfake Detection Challenge, which brought together Microsoft, Amazon Web Services and other researchers to cultivate and enhance deepfake detection capabilities (Vizoso, Vaz-Álvarez and López-García 2021).

2 US, Bill HR 5586, *DEEPFAKES Accountability Act*, 118th Cong, 2023.

3 US, SB 751, *An Act Relating to the Creation of a Criminal Offense for Fabricating a Deceptive Video with Intent to Influence the Outcome of an Election*, 86, Reg Sess, Tex, 2019 (enacted).

Apart from Meta, however, other major social media platforms seem to follow a similar trajectory, wherein they task users with addressing deepfakes (for example, through labelling and reporting) while neglecting to broaden their policy coverage beyond harassment to encompass mis- and disinformation. For instance, X, formerly known as Twitter, states that it seeks to label and remove content it deems to cause serious harm, which covers threats to physical security, incitement of abusive behaviour, mass violence, stalking or obsessive attention, harassment and voter suppression.[4] Additionally, Google requires content creators on YouTube to annotate their posts if they are AI-generated (Browning 2023). Likewise, TikTok also depends on creators to label their content and users to report unlabelled AI-generated content.[5] Notably, the aforementioned social media platforms and their policies fail to provide clear guidelines regarding the acceptability of deepfake content related to mis- and disinformation. Equally significant is the absence of concerted efforts by social media platforms to collaborate with tech companies to improve automated detection capabilities, which are crucial for identifying and removing deepfakes from their platforms.

In addition to the policies governing each platform, recent collaboration on industry-wide frameworks is a positive step. One specific framework that has emerged encompasses the companies that provide AI-generated media, such as OpenAI, Microsoft and Adobe, as well as the social media platforms where such content is distributed, including Meta, Google, TikTok and X. In February 2024, parties to the framework signed a pledge to combat election-related deepfakes (Committee for Economic Development 2024) and agreed to adhere to the following seven principles: prevention, provenance, detection, responsive protection, evaluation, public awareness and resilience (AI Elections Accord 2024). Two of these principles, provenance and detection, are technological mitigations explored further in the subsequent section. The framework, however, is voluntary, and with nearly 50 percent of the world holding elections in 2024 (Ewe 2023), social media platforms need to step up their efforts.

Tech Solutions for Tech Challenges

In the battle against deepfakes, technology is both a weapon and a shield, offering solutions while simultaneously posing challenges. Deepfake detection tools are one of those potential solutions. They leverage machine-learning techniques to train models using real and fake samples and can detect AI-generated content by searching for specific indicators (Bernardo, n.d.). For example, by discerning unique human qualities, Intel's FakeCatcher can achieve a 96 percent accuracy rate in detecting deepfakes (Intel 2022). One such human quality is the changing colour of blood flow from veins as the heart pumps, a process utilized by FakeCatcher's algorithms, which create blood flow maps and determine their authenticity (ibid.). Detection tools such as FakeCatcher could then be utilized by social media platforms to label, demote and take down deepfakes.

Another technological approach to mitigating deepfakes is validating the authenticity of a piece of content by mapping its provenance. This means verifying the source of the content and tracking its passage through the internet. The Coalition for Content Provenance and Authenticity (C2PA) is a non-profit organization aimed at

4 See https://help.twitter.com/en/rules-and-policies/manipulated-media.

5 See https://support.tiktok.com/en/using-tiktok/creating-videos/ai-generated-content.

building such standards. C2PA now comprises a community of 1,500 companies, including Adobe, Arm, Intel, Microsoft and Truepic (Ryan-Mosley 2023). The logic is to add and cryptographically bundle metadata of the location, author, date and digital signature into a piece of content.[6] If the content is modified, it triggers the generation of new cryptographic hashes and a separate metadata bundle. Consequently, upon encountering this content, users can determine that it was altered by observing the separate metadata bundles displayed.

Despite the promising outlook of technological solutions, several limitations need to be addressed. C2PA's effectiveness hinges on social media platforms adopting the standard and making it a condition to post content: this is a step no platform has taken to date (ibid.). Further, devices that capture content at its source (for example, laptops and smartphones) would need to integrate C2PA standards for metadata to be properly processed. Unless these measures are affordable, in demand and present no performance barriers, this also seems unlikely (Chesney and Citron 2019b). Regarding detection, challenges concerning efficacy and scalability persist. First, wrongdoers can quickly adapt to evade detection as the methods underlying various models are publicized, mirroring a dynamic akin to an arms race between detectors and wrongdoers (Westerlund 2019). A notable instance occurred in 2018 when researchers revealed that deepfake videos "do not blink at the same rate as real humans" (Helmus 2022, 11); wrongdoers quickly counteracted this revelation and integrated more realistic blinking patterns in their content within weeks. Second, the practical challenge of identifying deepfake content at scale poses a hurdle. With approximately 500 hours of content uploaded on YouTube alone every minute (Westerlund 2019), achieving comprehensive coverage with detection tools remains a distant prospect due to the sheer volume of content posted.

Recognizing that the deepfake problem cannot be solved solely through technological fixes is essential. The social dimensions of deepfake mis- and disinformation, driven by structural trends such as growing polarization and diminishing trust in mainstream media, must also be addressed. Despite advancements in detection and provenance capabilities, these structural trends influence individuals to seek information that aligns with their biases (Allcott and Gentzkow 2017), and even dispute fact-checking methodologies when they contradict their viewpoints (Humprecht, Esser and Van Aelst 2020).

Building Long-Term Immunity: Societal Resilience

Framing deepfakes as a social challenge, as much as a technological one, constitutes the fourth and potentially most long-lasting mitigation strategy (Prusila 2022). Adopting a social lens considers the structural context, such as heightened polarization and partisanship, that characterizes the threat deepfakes pose (Schwartz 2019). Deepfakes lack inherent meaning without a social context and a person to interpret them and pass judgment on their underlying messages (Habgood-Coote 2023). In societies that prioritize digital literacy education, promote healthy skepticism and encourage critical thinking among their members, it is logical to expect that individuals would become more resilient to the effects of deepfakes.

6 See https://c2pa.org/.

Fundamentally, the resilience of a society influences the other three mitigation categories. A resilient society lessens the need for states, social media platforms and tech companies to act urgently. Conversely, irrespective of all other efforts, the aforementioned mitigations can only address short-term symptoms and not long-term root causes if the society lacks resilience. Societal resilience facilitated through digital literacy training (Shu et al. 2020), whether through early education curriculum integration or other public outreach efforts, is the most sustainable yet difficult mitigation to implement.

Recommendations

This paper's organization into four distinct mitigation categories — legislative and regulatory measures, social media policies, technological initiatives and societal resilience — naturally leads to the formulation of policy recommendations that align with these delineations. These recommendations have been selected to leverage the respective capabilities and limitations of different actors, including state entities, social media platforms and tech companies. In doing so, it seeks to capitalize on the unique capacities of each actor while addressing the impediments that hinder an effective response to deepfakes. Note that these recommendations are not aimed at any particular state, social media platform or tech company. Instead, they are designed to be high-level guiding objectives that actors in these domains should aim to fulfill.

- **Recommendation 1:** States should seek to amend or introduce legislation to protect individuals from various forms of deepfake harassment and safeguard organizations against fraud and unfair business practices.
- **Recommendation 2:** States should seek to engage in outreach efforts to inform the public about the risks associated with deepfakes and consider various avenues to deliver digital literacy programs.
- **Recommendation 3:** Social media platforms should seek to expand the scope of deepfake policies beyond harassment to cover mis- and disinformation.
- **Recommendation 4:** Social media platforms should seek to amend or introduce policies to redistribute responsibility more evenly between users and the platform, which necessitates undertaking initiatives such as exchanging data with detection companies to overcome scalability issues and incorporating these tools into the platform.
- **Recommendation 5:** Tech companies should seek to address the challenge of detection research being made publicly available to wrongdoers by exploring the creation of a research network, or any other secure information-sharing framework, that restricts the dissemination of new findings solely to verified members.

Acknowledgements

I would like to thank Alex Wilner, Daniel Araya, Matthew da Mota and the Digital Policy Hub for providing a platform to produce innovative research.

About the Author

Ozan Ayata holds a master's degree in international relations with a focus on security and defence. He currently serves as a cybersecurity analyst at the Department of National Defence. Ozan has previously led various research projects at Carleton University and collaborated with organizations such as the Canadian Defence and Security Network, Sustainable Development Technology Canada and the Forum of Federations. With the Digital Policy Hub, Ozan will conduct research on how public-private partnerships can effectively prepare for and mitigate the risks posed by emerging technologies to international security, with a specific focus on components of AI, such as deepfakes.

Works Cited

AI Elections Accord. 2024. "A Tech Accord to Combat Deceptive Use of AI in 2024 Elections." Munich Security Conference, February 16. www.aielectionsaccord.com/uploads/2024/02/A-Tech-Accord-to-Combat-Deceptive-Use-of-AI-in-2024-Elections.FINAL_.pdf.

Allcott, Hunt and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36. https://doi.org/10.1257/jep.31.2.211.

Bernardo, Vítor. n.d. "Deepfake detection." European Data Protection Supervisor. www.edps.europa.eu/data-protection/technologymonitoring/techsonar/deepfake-detection_en.

Bickert, Monika. 2020. "Enforcing Against Manipulated Media." Meta, January 6. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

Browning, Michaela. 2023. "Our approach to protecting users from the risks of AI generated media." *India Blog*, November 29. https://blog.google/intl/en-in/company-news/technology/our-approach-to-protecting-users-from-the-risks-of-ai-generated-media/.

Byman, Daniel L., Chongyang Gao, Chris Meserole and V. S. Subrahmanian. 2023. "Deepfakes and international conflict." Brookings. www.brookings.edu/articles/deepfakes-and-international-conflict/.

Canadian Security Intelligence Service. 2023. October. *The Evolution of Disinformation: A Deepfake Future*. www.canada.ca/en/security-intelligence-service/corporate/publications/the-evolution-of-disinformation-a-deepfake-future.html.

Carter, Ash and Laura Manley. 2020. "Deepfakes." Tech Factsheets for Policymakers. Belfer Center for Science and International Affairs, Harvard Kennedy School. www.belfercenter.org/sites/default/files/2020-10/tappfactsheets/Deepfakes.pdf.

Chen, Heather and Kathleen Magramo. 2024. "Finance worker pays out $25 million after video call with deepfake 'chief financial officer.'" CNN, February 4. www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

Chesney, Robert and Danielle Keats Citron. 2019a. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107 (6): 1753–820. http://dx.doi.org/10.2139/ssrn.3213954.

– – –. 2019b. "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics." *Foreign Affairs*, December 11. www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war?cid=otr-authors-january_february_2019-121118.

Committee for Economic Development. 2024. "Policy Backgrounder: Tech Companies Pledge to Combat AI Election Interference." February 21. www.conference-board.org/research/ced-policy-backgrounders/tech-companies-pledge-to-combat-ai-election-interference.

Department of Homeland Security. 2021. *Increasing Threat of Deepfake Identities*. www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

Ewe, Koh. 2023. “The Ultimate Election Year: All the Elections Around the World in 2024.” *Time*, December 28. https://time.com/6550920/world-elections-2024/.

Farid, Hany. 2021. “Detecting and Combating Deep Fakes.” *Journal of Intelligence, Conflict, and Warfare* 3 (3): 83–87. https://doi.org/10.21810/jicw.v3i3.2752.

– – –. 2022. “Creating, Using, Misusing, and Detecting Deep Fakes.” *Journal of Online Trust and Safety* 1 (4): 1–33. https://doi.org/10.54501/jots.v1i4.56.

Habgood-Coote, Joshua. 2023. “Deepfakes and the epistemic apocalypse.” *Synthese* 201 (103): 1–23. https://doi.org/10.1007/s11229-023-04097-3.

Helmus, Todd C. 2022. “Artificial Intelligence, Deepfakes, and Disinformation: A Primer.” Rand Corporation, July 6. www.rand.org/pubs/perspectives/PEA1043-1.html.

Humprecht, Edda, Frank Esser and Peter Van Aelst. 2020. “Resilience to Online Disinformation: A Framework for Cross-National Comparative Research.” *International Journal of Press/Politics* 25 (3): 493–516. https://doi.org/10.1177/1940161219900126.

Intel. 2022. “Intel Introduces Real-Time Deepfake Detector.” News post, November 14. www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.3zphbv.

Iyengar, Rishi. 2021. “Why it’s so difficult to bring ransomware attackers to justice.” CNN Business, July 8. www.cnn.com/2021/07/08/tech/ransomware-attacks-prosecution-extradition/index.html.

Kleemann, Aldo. 2023. “Deepfakes – When We Can No Longer Believe Our Eyes and Ears.” SWP Comment No. 52. German Institute for International and Security Affairs. October. www.swp-berlin.org/publications/products/comments/2023C52_Deepfakes.pdf.

Lewis, James Andrew and Arthur Nelson. 2019. “Trust Your Eyes? Deepfakes Policy Brief.” Center for Strategic and International Studies, October 23. www.csis.org/analysis/trust-your-eyes-deepfakes-policy-brief.

Meel, Vidushi. n.d. “Artificial Neural Network: Everything You Need to Know.” *Viso Blog*. https://viso.ai/deep-learning/artificial-neural-network/.

Moreau, Shona and Chloe Rourke. 2024. “Fake porn causes real harm to women.” *Policy Options*, February 8. https://policyoptions.irpp.org/magazines/february-2024/fake-porn-harm/.

Nightingale, Sophie and Hany Farid. 2021. “Examining the Global Spread of COVID-19 Misinformation.” *arXiv*, January 27. https://doi.org/10.48550/arXiv.2006.08830.

O’Carroll, Lisa. 2023. “How the EU Digital Services Act affects Facebook, Google and others.” *The Guardian*, August 25. www.theguardian.com/world/2023/aug/25/how-the-eu-digital-services-act-affects-facebook-google-and-others.

Prusila, Amanda Carlianne. 2022. “Truth, Lies and ‘Deepfakes’: The Epistemology of Photographic Depictions.” Master’s thesis, Carleton University. https://doi.org/10.22215/etd/2022-14911.

Ryan-Mosley, Tate. 2023. “Cryptography may offer a solution to the massive AI-labelling problem.” *MIT Technology Review*, July 28. www.technologyreview.com/2023/07/28/1076843/cryptography-ai-labeling-problem-c2pa-provenance/.

Saifuddin, Ahmed. 2023. “Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism.” *New Media & Society* 25 (5): 1108–29. https://doi.org/10.1177/14614448211019198.

Sayler, Kelley M. and Laurie A. Harris. 2023. “Deep Fakes and National Security.” Congressional Research Service. April 17. https://crsreports.congress.gov/product/pdf/IF/IF11333.

Schwartz, Oscar. 2019. “Deepfakes aren’t a tech problem. They’re a power problem.” *The Guardian*, June 24. www.theguardian.com/commentisfree/2019/jun/24/deepfakes-facebook-silicon-valley-responsibility.

Shu, Kai, Amrita Bhattacharjee, Faisal Alatawi, Tahora H. Nazer, Kaize Ding, Mansooreh Karami and Huan Liu. 2020. “Combating disinformation in a social media age.” *Data Mining and Knowledge Discovery* 10 (6): 1–23. https://doi.org/10.1002/widm.1385.

van der Sloot, Bard and Yvette Wagensveld. 2022. “Deepfakes: regulatory challenges for the synthetic society.” *Computer Law & Security Review* 46: 1–15. https://doi.org/10.1016/j.clsr.2022.105716.

Vizoso, Ángel, Martín Vaz-Álvarez and Xosé López-García. 2021. “Fighting Deepfakes: Media and Internet Giants’ Converging and Diverging Strategies against Hi-Tech Misinformation.” *Media and Communication* 9 (1): 291–300. https://doi.org/10.17645/mac.v9i1.3494.

Westerlund, Mika. 2019. “The Emergence of Deepfake Technology: A Review.” *Technology Innovation Management Review* 9 (11): 39–52. https://doi.org/10.22215/timreview/1282.