Computer Law & Security Review 58 (2025) 106162

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

Deepfake detection in generative AI: A legal framework proposal to protect human rights

Felipe Romero-Moreno

Schools of Law and Education, University of Hertfordshire, UK

ARTICLE INFO

Keywords:
Deepfake detection
Generative AI
XAI
C2PA
Human rights

ABSTRACT

Deepfakes, exploited for financial fraud, political misinformation, non-consensual imagery, and targeted harassment, represent a rapidly evolving threat to global information integrity, demanding immediate and coordinated intervention. This research undertakes technical and comparative legal analyses of deepfake detection methods. It examines key mitigation strategies—including AI-powered detection, provenance tracking, and watermarking—highlighting the pivotal role of the Coalition for Content Provenance and Authenticity (C2PA) in establishing media authentication standards. The study investigates deepfakes' complex intersections with the admissibility of legal evidence, non-discrimination, data protection, freedom of expression, and copyright, questioning whether existing legal frameworks adequately balance advances in detection technologies with the protection of individual rights. As national strategies become increasingly vital amid geopolitical realities and fragmented global governance, the research advocates for a unified international approach grounded in UN Resolution 78/265 on safe, secure, and trustworthy AI. It calls for a collaborative framework that prioritizes interoperable technical standards and harmonized regulations. The paper critiques legal frameworks in the EU, US, UK, and China—jurisdictions selected for their global digital influence and divergent regulatory philosophies—and recommends developing robust, accessible, adaptable, and internationally interoperable tools to address evidentiary reliability, privacy, freedom of expression, copyright, and algorithmic bias. Specifically, it proposes enhanced technical standards; regulatory frameworks that support the adoption of explainable AI (XAI) and C2PA; and strengthened cross-sector collaboration to foster a trustworthy deepfake ecosystem.

1. The global deepfake threat: Challenges and the regulatory imperative

The alarming potential of deepfakes was starkly demonstrated in 2024 when a CEO was deceived into authorizing a $25.6 million transfer via an AI-generated video call.[1] Deepfakes—synthetically generated or manipulated media (video, audio, images, text) convincingly replicating reality—pose a growing threat to individuals, organizations, and democratic discourse.[2] This threat manifests in various forms.

Financially, savers in the UK, Europe, and Canada lost $35 million to celebrity deepfake scams,[3] while businesses saw losses nearing $450,000,[4]

E-mail address: f.romero-moreno@herts.ac.uk.

[1] World Economic Forum, 'This happens more frequently than people realize': Arup chief on the lessons learned from a $25m deepfake crime. https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/, 2025 (accessed 6 June 2025).

[2] See, e.g., Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending certain regulations ('the AI Act') [2024] OJ L 178/1, art 3(60).

[3] The Guardian, Revealed: the scammers who conned savers out of $35m using fake celebrity ads. https://www.theguardian.com/money/2025/mar/05/revealed-the-scammers-who-conned-savers-out-of-35m-using-fake-celebrity-ads, 2025 (accessed 6 June 2025).

[4] Regula, The impact of deepfake fraud: Risks, solutions, and global trends. https://regulaforensics.com/blog/impact-of-deepfakes-on-idv-regula-survey/, 2024 (accessed 6 June 2025).

https://doi.org/10.1016/j.clsr.2025.106162

Available online 23 June 2025

with financial sector losses exceeding $600,000.[5] Politically, deepfakes spread misinformation, including the deepfake Biden robocall aimed at voter suppression,[6] doctored videos of Ukrainian President Zelenskyy,[7] and viral fabrications falsely depicting Indian ministers apologizing amidst the India-Pakistan military conflict.[8]

Socially, deepfakes fuel the proliferation of non-consensual intimate imagery, disproportionately affecting women (99 % of victims).[9] Celebrities,[10] politicians,[11] and ordinary individuals are vulnerable, given the ease of creating 60 s pornographic videos quickly at minimal cost.[12] Deepfakes are also used for targeted harassment, often incorporating racial or gender-based slurs, as illustrated by the victimization of schoolgirls in Almendralejo (Spain),[13] and targeting female journalists.[14] Research indicates potential racial bias in deepfake technology may disproportionately affect communities of color through harassment and misidentification.[15] Alarmingly, AI-generated child sexual abuse material quadrupled between 2023 and 2024,[16] and increased sextortion cases.[17]

Projected financial losses alone are expected to surge from $12.3 billion in 2023 to a staggering $40 billion by 2027.[18] This is amplified by the 223 % surge in dark web trading of deepfake creation tools,[19] some costing as little as $20,[20] making these deceptive technologies readily accessible. The World Economic Forum has warned that cyber insecurity, including deepfakes, poses a long-term global risk to supply chains, financial stability, and democratic systems.[21] These detection challenges necessitate an evolution in technical and regulatory frameworks.

Beyond these immediate harms, deepfakes are powered by sophisticated generative AI tools like DALL-E or Stable Diffusion. These tools learn and replicate complex patterns in media, contributing to deepfakes' increasingly realistic and difficult-to-detect nature.[22] This inherent duality is evident in examples ranging from viral celebrity deepfakes[23] to multi-million dollar fraud,[24] which contrasts sharply with the use of deepfake "wrappers" protecting LGBTQ+ activists in the HBO documentary *Welcome to Chechnya*.[25] This tension underscores the complex challenge of balancing innovation with safeguards.

While current literature extensively examines the technical effectiveness of deepfake detection tools—from single to multi-modal data analysis (images, audio, video, text) using machine learning, high-

[5] Regula, Deepfake fraud costs the financial sector an average of $600,000 for each company. https://regulaforensics.com/news/deepfake-fraud-costs/, 2024 (accessed 6 June 2025).

[6] Federal Communications Commission, FCC proposes $6 million fine for illegal robocalls that used Biden deepfake generative AI voice message. https://docs.fcc.gov/public/attachments/DOC-402762A1.pdf, 2024, (accessed 6 June 2025).

[7] M. Boháček, H. Farid, 2022. Protecting president Zelenskyy against deep fakes. arXiv. arXiv:2206.12043v1. https://doi.org/10.48550/arXiv.2206.12043.

[8] Boom Fact Check, AI videos of PM Modi, Amit Shah & Jaishankar apologizing to Pakistan viral. https://www.boomlive.in/fact-check/narendra-modi-jaishankar-amit-shah-india-pakistan-operation-sindoor-pahalgam-deepfakes-ai-28534, 2025 (accessed 6 June 2025).

[9] C. Yavuz, 2025. Adverse human rights impacts of dissemination of nonconsensual sexual deepfakes in the framework of European Convention on Human Rights: a victim-centered perspective. Computer Law & Security Review 56, 106108. https://doi.org/10.1016/j.clsr.2025.106108.

[10] J. Sturges, 2024. Taylor Swift, deepfakes, and the First Amendment: changing the legal landscape for victims of non-consensual artificial pornography, Georgetown Journal of Gender and the Law 25(2) (2024) 1-11. https://www.law.georgetown.edu/gender-journal/online/volume-xxv-online/taylor-swift-deepfakes-and-the-first-amendment-changing-the-legal-landscape-for-victims-of-non-consensual-artificial-pornography/.

[11] Politico, Italy's Giorgia Meloni called to testify in deepfake porn case. https://www.politico.eu/article/italian-pm-giorgia-meloni-called-to-testify-in-deepfake-porn-case/, 2024, (accessed 6 June 2025).

[12] Security Hero, 2023 state of deepfakes: realities, threats, and impact. https://www.securityhero.io/state-of-deepfakes/, 2023 (accessed 6 June 2025).

[13] A. M. Narvali, J. A. Skorburg, M. J. Goldenberg, Cyberbullying girls with pornographic deepfakes is a form of misogyny, The Conversation. https://theconversation.com/cyberbullying-girls-with-pornographic-deepfakes-is-a-form-of-misogyny-217182, 2023 (accessed 6 June 2025).

[14] Women Press Freedom, UK: Women Press Freedom Condemns deepfake attacks on Cathy Newman as part of a growing trend against journalists. https://www.womeninjournalism.org/threats-all/uk-women-press-freedom-condemns-deepfake-attacks-on-cathy-newman-as-part-of-a-growing-trend-against-journalists, 2024 (accessed 6 June 2025).

[15] University at Buffalo, Study: new deepfake detector designed to be less biased. https://www.buffalo.edu/news/releases/2024/01/new-deepfake-detector-designed-to-less-biased.html, 2024 (accessed 6 June 2025); S. Overton, Testimony before the Subcommittee on Cybersecurity, Information Technology, and Government Innovation, US House Committee on Oversight and Accountability: "Advances in Deepfake Technology." https://oversight.house.gov/wp-content/uploads/2023/11/Overton-Testimony-on-Advances-in-Deepfake-Technology-11-8-23-1.pdf, 2023 (accessed 6 June 2025).

[16] Internet Watch Foundation, New AI child sexual abuse laws announced following IWF campaign. https://www.iwf.org.uk/news-media/news/new-ai-child-sexual-abuse-laws-announced-following-iwf-campaign/, 2025 (accessed 6 June 2025).

[17] National Crime Agency, NCA issues urgent warning about 'sextortion'. https://www.nationalcrimeagency.gov.uk/news/nca-issues-urgent-warning-about-sextortion, 2024 (accessed 6 June 2025).

[18] Deloitte, Generative AI is expected to magnify the risk of deepfakes and other fraud in banking. https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deepfake-banking-fraud-risk-on-the-rise.html, 2024 (accessed 6 June 2025).

[19] Accenture, Beyond the illusion—unmasking the real threats of deepfakes. https://www.accenture.com/us-en/blogs/security/beyond-illusion-unmasking-real-threats-deepfakes, 2024 (accessed 6 June 2025).

[20] Bloomberg, Deepfake imposter scams are driving a new way of fraud. https://www.bloomberg.com/news/articles/2023-08-21/money-scams-deepfakes-ai-will-drive-10-trillion-in-financial-fraud-and-crime?embedded-checkout=true, 2023 (accessed 6 June 2025).

[21] World Economic Forum, Global Risks Report 2024. https://www.weforum.org/publications/global-risks-report-2024/, 2024 (accessed 6 June 2025).

[22] F.-A. Croitoru, A.-I. Hiji, V. Hondru, N.C. Ristea, P. Irofti, M. Popescu, C. Rusu, R.T. Ionescu, F.S. Khan, M. Shah, 2024. Deepfake media generation and detection in the generative AI era: a survey and outlook. arXiv. arXiv:2411.19537. https://doi.org/10.48550/arXiv.2411.19537.

[23] B. Kira, 2024. When non-consensual intimate deepfakes go viral: the insufficiency of the UK Online Safety Act. Computer Law & Security Review. 54, 106024. https://doi.org/10.1016/j.clsr.2024.106024.

[24] US Department of Homeland Security, Increasing threats of deepfake identities. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf, 2023 (accessed 6 June 2025).

[25] World Intellectual Property Organization, Artificial intelligence: deepfakes in the entertainment industry. https://www.wipo.int/web/wipo-magazine/articles/artificial-intelligence-deepfakes-in-the-entertainment-industry-42620, 2022 (accessed 6 June 2025).

lighting a constant "arms race"—a critical research gap exists.[26] The comprehensive analysis of how diverse global regulations address these very detection technologies remains limited.[27] This research bridges this gap by also investigating such legal landscape and its impact on developing and deploying effective deepfake detection systems.

The paper examines deepfake detection tools and analyses the role of content provenance, as defined by the Coalition for Content Provenance and Authenticity (C2PA) standards.[28] It further evaluates the current legal landscape in key regions—the EU, US, UK, and China—and their respective frameworks for addressing deepfake detection challenges. These influential regions offer critical insights into global approaches to this evolving threat due to their significant digital influence and contrasting regulatory philosophies.[29]

The EU's proactive digital regulations, including the AI Act (AIA mandating transparency),[30] the General Data Protection Regulation (GDPR data governance),[31] and the Digital Services Act (DSA platform accountability),[32] offer a multi-layered approach. In contrast, the US lacks a unified federal framework, resulting in varied state-level initiatives.[33] The UK's Online Safety Act 2023 (OSA) places responsibility on platforms,[34] while China's Deep Synthesis Provisions employ a top-down regulatory model with mandatory labelling and algorithm review.[35]

The central argument of this study is that effectively addressing the deepfake threat, while safeguarding fundamental rights globally, demands a comprehensive, multi-faceted approach integrating technical detection, ethical considerations, adaptive governance, and clear accountability across the deepfake ecosystem.

To explore this central argument, this paper proceeds with the following roadmap: Section 2 reviews deepfake detection methods and techniques, including their effectiveness, limitations, and relevant technical standards like C2PA. It highlights the ongoing "arms race"[36] and proposes solutions related to C2PA implementation and explainable AI (XAI). Section 3 examines the fragmented legal landscape in the EU, US, UK, and China, analyzing tensions and divergences hindering a unified global response. Section 4 argues for national action strategies due to the current geopolitical climate and the fragmented global frameworks. Section 5 assesses deepfake detection and regulatory pathways through the framework of the UN Resolution 78/265 on safe, secure, and trustworthy AI.[37] Section 6 concludes by proposing an integrated framework fostering a trustworthy digital ecosystem against evolving deepfake threats, outlining stakeholder roles and responsibilities.

2. Deepfake detection: Technical approaches and their limitations

2.1. Core concepts and metrics

In deepfake detection, precision (correctly identified deepfakes among all predictions), recall (correctly identified actual deepfakes),

[26] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: a survey of face manipulation and fake detection, Information Fusion 64 (2020) 131-148. https://doi.org/10.1016/j.inffus.2020.06.014; S. Agarwal, H. Farid, O. Fried, M. Agrawala, Detecting deep-fake videos from Phoneme-Viseme Mismatches, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 2814-2822. https://doi.org/10.1109/CVPRW50498.2020.00338; R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, S. Parkinson, A survey on the detection and impacts of deepfakes in visual, audio, and textual formats, IEEE Access 11 (2023) 144497-144529. https://doi.org/10.1109/ACCESS.2023.3344653; H. Khalid, M. Kim, S. Tariq, S.S. Woo, 2021. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. arXiv. arXiv:2109.02993. https://doi.org/10.48550/arXiv.2109.02993; T.T. Nguyen, Q.V.H. Nguyen, D.T. Nguyen, D.T. Nguyen, T. Huynh-The, S. Nahavandi, T.T. Nguyen, Q.-V. Pham, C.M. Nguyen, 2022. Deep learning for deepfakes creation and detection: a survey. Computer Vision and Image Understanding. 223, 103525. https://doi.org/10.1016/j.cviu.2022.103525; D. K. Citron, R. Chesney, Deepfakes and the new disinformation war: the coming age of post-truth geopolitics, Foreign Affairs (2018). https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war; G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, M. Prasad, 2024. A comprehensive review of deepfake detection using advanced machine learning and fusion methods. Electronics. 13, 95. https://doi.org/10.3390/electronics13010095; F. Abbas, A. Taeihagh, 2024. Unmasking deepfakes: a systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications 252, Part B, 124260. https://doi.org/10.1016/j.eswa.2024.124260; A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, 2019. FaceForensics++: learning to detect manipulated facial images. Computer Science, Computer Vision and Patter Recognition. arXiv. arXiv:1901.08971. https://doi.org/10.48550/arXiv.1901.08971.

[27] M.-P. Sandoval, M. de Almeida Vau, J. Solaas, L. Rodrigues, 2024. Threat of deepfakes to the criminal justice system: a systematic review. Crime Science. 13, 41. https://doi.org/10.1186/s40163-024-00239-1; A.P. Singh, Legal implications of deepfake technology in Criminal Law, International Journal of Law Management & Humanities 8(1) (2025) 1645-1661. https://doij.org/10.10000/IJLMH.119051; B. van der Sloot, Y. Wagensveld, 2022. Deepfakes: regulatory challenges for the synthetic society. Computer Law & Security Review. 46, 105716. https://doi.org/10.1016/j.clsr.2022.105716; M. Labuz, Deep fakes and the Artificial Intelligence Act—an important signal or a missed opportunity?, Policy & Internet 16(4) (2024) 1-18. https://onlinelibrary.wiley.com/doi/10.1002/poi3.406; M. Labuz, A teleological interpretation of the definition of deepfakes in the EU Artificial Intelligence Act—a purpose-based approach to potential problems with the word "existing", Policy & Internet 17(1) (2025) 1-14. https://onlinelibrary.wiley.com/doi/10.1002/poi3.435.

[28] C2PA. https://c2pa.org/, founded in 2021 (accessed 6 June 2025).

[29] J. Kazaz, Regulating deepfakes: global approaches to combatting AI-driven manipulation, GLOBSEC Policy Paper (2024) 1-7. https://www.globsec.org/sites/default/files/2024-12/Regulating%20Deepfakes%20-%20Global%20Approaches%20to%20Combatting%20AI-Driven%20Manipulation%20policy%20paper%20ver4%20web.pdf.

[30] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending certain regulations ('the AI Act') [2024] OJ L 178/1.

[31] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

[32] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1.

[33] See, e.g., California SB 926, 2023-2024 Reg Sess, ch 289 (2024); Texas SB 751, 89th Reg Sess (2025); Florida Laws 2022, ch 2022-212; Louisiana Acts 2023, No 175; California AB 2655, 2023-2024 Reg Sess, ch 261 (2024); California SB 942, 2023-2024 Reg Sess, ch 291 (2024).

[34] Online Safety Act 2023.

[35] Provisions on the Administration of Deep Synthesis Internet Information Services (Order No 12 of the Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security, 25 November 2022). <http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm> accessed 6 June 2025.

[36] L. Laurier, A. Giulietta, A. Octavia, M. Cleti, 2024. The cat and mouse game: the ongoing arms race between diffusion models and detection methods. arXiv. arXiv:2410.18866v1. https://doi.org/10.48550/arXiv.2410.18866.

[37] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

Table 1
Established deepfake detection methods.

Method Category	Description	Key Features	Challenges
Artifact-based	Analyzes images and videos for creation inconsistencies	Detects visual and inter-modal artifacts (e.g., blurring, lip-sync errors)	"Arms race" with creators, artifacts often short-lived
Behavioral Biometrics	Analyzes user behavior for anomalies	Measures typing speed, mouse movements, touchscreen pressure	Vulnerable to adaptive attacks, privacy concerns, explainability issues
Physiological Signals	Detects anomalies via biological cues	Analyzes physiological cues, such as heart rate, blinking, pupil dilation	Needs high-quality video, personal variability, susceptible to circumvention
Deep Learning	Uses deep learning to detect manipulation patterns	Applies CNNs, RNNs, GANs to identify deepfake traits	High data needs, lack of transparency, evolving threats, limited explainability
Hybrid Multi-Modal	Combines various analysis methods for greater accuracy	Integrates audio, video, image, and text analysis	Data fusion complexity, high computational cost, robustness and bias concerns

and F1-score (a balance of both) are crucial metrics. Their optimal balance depends on the application and the potential harms of misclassification.[38]

Deepfake detection approaches fall into two categories: those applied during creation and those used upon receipt. At creation, initiatives like embedding provenance (e.g., labelling via the Coalition for Content Provenance and Authenticity - C2PA) and watermarks aim to establish authenticity from the source[39] (as shown in Table 2). Upon receipt or sharing, methods such as AI-powered detection analyze the media itself[40] (as shown in Table 1).

Traditional detection methods include artifact-based methods (scrutinizing inconsistencies),[41] behavioral biometrics (e.g., typing speed),[42] physiological signal analysis (e.g., heart rate),[43] and deep learning-based methods (identifying manipulation patterns).[44] Multimodal hybrid approaches combine these techniques, analyzing video, audio, images, and text.[45] Beyond these, promising solutions include liveness detection,[46] Zero-Knowledge Biometrics (ZKB),[47] blockchain-based verification,[48] quantum computing,[49] and adversarial training-based methods.[50]

The challenge lies in deepfakes' constant evolution. For instance, underfitting (models too simplistic) fails to capture nuanced manipulations in financial fraud or non-consensual content, leading to low

Table 2
Methods for identifying deepfakes at creation.

C2PA-Enabled Method	Description	Limitations
Provenance (Labelling)	Metadata labels (e.g. "digital nutrition label")	Inconsistent standards, complex interpretation, over-reliance, limited provenance scope, privacy concerns, metadata stripping, low adoption
Digital Watermarks	Embedded visible/ invisible data	Vulnerable to removal/manipulation, risk of false labeling ("liar's dividend"), lack of standards, limited implementation

[38] See, e.g., R. Sunil, P. Mer, A. Diwan, R. Mahadeva, A. Sharma, 2025. Exploring autonomous methods for deepfake detection: a detailed survey on techniques and evaluation. Heliyon. 11, e42273. https://doi.org/10.1016/j.heliyon.2025.e42273.

[39] C2PA. https://c2pa.org/, founded in 2021 (accessed 6 June 2025).

[40] Digital Regulation Cooperation Forum, Tackling online fraud and scams: Ofcom and FCA collaboration. https://www.drcf.org.uk/publications/blogs/tackling-online-fraud-and-scams-ofcom-and-fca-collaboration/, 2024 (accessed 6 June 2025).

[41] See, e.g., F. Abbas, A. Taeihagh, 2024. Unmasking deepfakes: a systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications 252, Part B, 124260. https://doi.org/10.1016/j.eswa.2024.124260; S. Agarwal, H. Farid, O. Fried, M. Agrawala, Detecting deep-fake videos from Phoneme-Viseme Mismatches, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 2814-2822. https://doi.org/10.1109/CVPRW50498.2020.00338; Deepware, Frequently asked questions. https://deepware.ai/faq/ (accessed 6 June 2025).

[42] See, e.g., M. Ghilom, S. Latifi, 2024. The role of machine learning in advanced biometric systems. Electronics. 13, 2667. https://doi.org/10.3390/electronics13132667; BioCatch, Undetectable scams: deepfakes & AI change the game. https://www.biocatch.com/blog/undetectable-scams-deepfakes, 2025 (accessed 6 June 2025); BioCatch, From phishing to deepfakes: tackling identity fraud and social engineering in the Middle East. https://www.biocatch.com/blog/tackling-identity-fraud-middle-east, 2024 (accessed 6 June 2025).

[43] See, e.g., J. Hernandez-Ortega, R. Tolosana, J. Fierrez, A. Morales, 2020. DeepfakesOn-Phys: deepfakes detection based on heart rate estimation. arXiv. arXiv:2010.00400. https://doi.org/10.48550/arXiv.2010.00400; T. Jung, S. Kim, K. Kim, DeepVision: deepfakes detection using human eye blinking pattern, IEEE Access 8 (2020) 83144-83154. https://doi.org/10.1109/ACCESS.2020.2988660; Intel's FakeCatcher, The word's first-real time deepfake detector. https://download.intel.com/newsroom/2022/new-technologies/FakeCatcher-Infographic.pdf, 2022 (accessed 6 June 2025).

[44] See, e.g., T.T. Nguyen, Q.V.H. Nguyen, D.T. Nguyen, D.T. Nguyen, T. Huynh-The, S. Nahavandi, T.T. Nguyen, Q.-V. Pham, C.M. Nguyen, 2022. Deep learning for deepfakes creation and detection: a survey. Computer Vision and Image Understanding. 223, 103525. https://doi.org/10.1016/j.cviu.2022.103525; Sentinel, Defending against deepfakes and information warfare. https://thesentinel.ai/, 2020 (accessed 6 June 2025).

[45] See, e.g., D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, & S. Tubaro, 2023. A robust approach to multimodal deepfake detection. Journal of Imaging. 9, 122. https://doi.org/10.3390/jimaging9060122; Reality Defender, https://www.realitydefender.com/, 2024 (accessed 6 June 2025).

[46] See, e.g., S. Khade, S. Ahirrao, S. Phansalkar, K. Kotecha, S. Gite, S.D. Thepade, 2021. Iris liveness detection for biometric authentication: a systematic literature review and future directions. Inventions. 6, 65. https://doi.org/10.3390/inventions6040065; Oz Forensics, Passive and active liveness. https://doc.ozforensics.com/oz-knowledge/general/oz-platform/passive-and-active-liveness, 2025 (accessed 6 June 2025).

[47] See, e.g., Keyless, Zero-Knowledge Biometrics™ the future of authentication. https://26689385.fs1.hubspotusercontent-eu1.net/hubfs/26689385/%5B2023%5D%20Downloadable%20Content/Keyless_Zero_Knowledge_Biometrics.pdf, 2023 (accessed 6 June 2025).

[48] See, e.g., A. Heidari, N.J. Navimipour, H. Dag, S. Talebi, M. Unal, A novel blockchain-based deepfake detection method using federated and deep learning models, Cognitive Computation 16 (2024) 1073-1091. https://doi.org/10.1007/s12559-024-10255-7; Weverify, Deepfake detector. https://weverify.eu/tools/deepfake-detector/, founded in 2020 (accessed 6 June 2025).

[49] See, e.g., C—H.A. Lin, C-Y. Liu, S.Y-C. Chen, K-C. Chen, 2024. Quantum-Trained Convolutional Neural Network for deepfake audio detection. arXiv. arXiv:2410.09250v1. https://doi.org/10.48550/arXiv.2410.09250; B. Eray Katı, E. Uğur Küçüksille, G. Sarıman, 2025. Enhancing deepfake detection through Quantum Transfer Learning and Class-Attention Vision Transformer architecture. Applied Science. 15, 525. https://doi.org/10.3390/app15020525.

[50] See, e.g., P. Neekhara, B. Dolhansky, J. Bitton and C. C. Ferrer, Adversarial threats to deepfake detection: a practical perspective, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA (2021) 923-932. https://ieeexplore.ieee.org/document/9522903; Mindgard, Bypassing AI-driven deepfake detection via evasion attacks. https://mindgard.ai/blog/bypassing-ai-driven-deepfake-detection-via-evasion-attacks, 2025 (accessed 6 June 2025).

precision and recall.[51] This means it produces both false negatives (missed deepfakes) and false positives (misidentified genuine content).[52] Conversely, overfitting (models memorizing training noise) results in poor real-world generalization, potentially misclassifying genuine content (e.g., due to lighting differences) in financial scams or political smear campaigns, leading to a high false positive rate and impacting precision.[53]

The optimal balance varies by application. For election interference, prioritizing recall is vital to avoid missing malicious deepfakes, even if it means more false positives. However, in high-volume social media content moderation, precision is often prioritized to minimize wrongful censorship of potentially non-consensual deepfakes.[54] Combining multiple approaches is generally the most effective strategy to improve accuracy and robustness in this evolving landscape.

2.2. Established deepfake detection methods and their limitations

2.2.1. Artifact-based methods

Artifact-based deepfake detection seeks to identify telltale inconsistencies introduced during the manipulation process, a crucial first line of defence against various malicious deepfake threats. These inconsistencies range from readily apparent visual anomalies, such as blurring and checkerboard patterns often found in less sophisticated face-swap deepfakes or manipulated identification documents used in financial fraud,[55] to more subtle inter-modal discrepancies like phoneme-viseme mismatches, crucial for detecting inconsistent lip-sync and speech manipulations in political misinformation videos.[56]

Early detection techniques relying on pixel-level and frequency domain analysis, while capable of identifying basic artifacts in initial deepfake iterations, are now easily circumvented by advanced generative models employing seamless blending and sophisticated rendering.[57] Contemporary AI-driven approaches, including Deepware Scanner's analysis of complex visual cues in high-quality forgeries,[58] and attention-based networks like BiG-Arts[59] and LAA-Net designed to pinpoint subtle spatial and temporal anomalies, represent advancements.[60]

However, the fundamental "arms race" persists. As deepfake creators improve their evasion techniques—including adversarial methods like using photorealistic imagery without obvious visual flaws or synthesizing accurate lip movements to evade audio-visual mismatch detection—the reliability of artifact-based methods decreases.[61] Post-processing techniques like noise addition and re-compression further serve as effective evasion tactics, obscuring detectable inconsistencies across a spectrum of harmful deepfakes.[62]

2.2.2. Behavioral and physiological biometrics for detection

Deepfake detection employs behavioral biometrics and physiological signal analysis, each with inherent limitations against specific threats. Behavioral biometrics, as seen in systems like BioCatch,[63] analyses typing speed, mouse movements, and touchscreen pressure to detect anomalies indicative of financial fraud, like an imposter using a deepfake-controlled account.[64] However, adaptive mimicry, where those spreading political misinformation via compromised accounts learn legitimate behavior, limits its efficacy.[65] Extensive data collection for robust analysis also raises privacy concerns (e.g., detailed user interaction patterns, device information, geolocation, transactional history),[66] particularly when monitoring individuals potentially involved in non-consensual deepfake distribution. The challenge of distinguishing genuine anomalies from normal user variation leads to reliability issues.[67]

Physiological analysis examines subtle cues like heart rate and "blood flow" to theoretically detect deepfake imposters in financial scams whose physiological signals might appear unnatural under scrutiny.[68] Examples of physiological signals used include heart rate,[69]

[51] A.H. Soudy, O. Sayed, H. Tag-Elser, R. Mahmoud, Deepfake detection using convolutional vision transformers and convolutional neural networks, Neural Computing and Applications 36 (2024) 19759-19775. https://doi.org/10.1007/s00521-024-10181-7.

[52] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, A. Sharma, 2025. Exploring autonomous methods for deepfake detection: a detailed survey on techniques and evaluation. Heliyon. 11, e42273. https://doi.org/10.1016/j.heliyon.2025.e42273.

[53] A.H. Soudy, O. Sayed, H. Tag-Elser, R. Mahmoud, Deepfake detection using convolutional vision transformers and convolutional neural networks, Neural Computing and Applications 36 (2024) 19759-19775. https://doi.org/10.1007/s00521-024-10181-7.

[54] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, A. Sharma, 2025. Exploring autonomous methods for deepfake detection: a detailed survey on techniques and evaluation. Heliyon. 11, e42273. https://doi.org/10.1016/j.heliyon.2025.e42273.

[55] F. Abbas, A. Taeihagh, 2024. Unmasking deepfakes: a systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications 252, Part B, 124260. https://doi.org/10.1016/j.eswa.2024.124260.

[56] S. Agarwal, H. Farid, O. Fried, M. Agrawala, Detecting deep-fake videos from Phoneme-Viseme Mismatches, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 2814-2822. https://doi.org/10.1109/CVPRW50498.2020.00338.

[57] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, 2020. Leveraging frequency analysis for deep fake image recognition. Computer Science, Computer Vision and Pattern Recognition. arXiv. arXiv:2003.08685v3. https://doi.org/10.48550/arXiv.2003.08685.

[58] Deepware, Frequently asked questions. https://deepware.ai/faq/ (accessed 6 June 2025).

[59] H. Chen, Y. Li, D. Lin, B. Li, J. Wu, 2023. Watching the big artifacts: exposing deepfake videos via bi-granularity artifacts. Pattern Recognition. 135, 109179. https://doi.org/10.1016/j.patcog.2022.109179.

[60] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, D. Aouada, LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. arXiv. arXiv:2401.13856(v2). https://doi.org/10.48550/arXiv.2401.13856.

[61] Y. Chen, Y. Yu, R. Ni, H. Li, W. Wang, Y. Zhao, 2025. NPVForensics: learning VA correlations in non-critical phoneme–viseme regions for deepfake detection, Image and Vision Computing 156, 105461. https://doi.org/10.1016/j.imavis.2025.105461.

[62] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, M. Prasad, 2024. A comprehensive review of deepfake detection using advanced machine learning and fusion methods. Electronics. 13, 95. https://doi.org/10.3390/electronics13010095.

[63] BioCatch, Preparing smaller financial institutions for deepfakes and other AI-powered attacks. https://www.biocatch.com/blog/preparing-smaller-financial-institutions-for-ai-deepfake-attacks, 2024 (accessed 6 June 2025).

[64] BioCatch, Undetectable scams: deepfakes & AI change the game. https://www.biocatch.com/blog/undetectable-scams-deepfakes, 2025 (accessed 6 June 2025); BioCatch, From phishing to deepfakes: tackling identity fraud and social engineering in the Middle East. https://www.biocatch.com/blog/tackling-identity-fraud-middle-east, 2024 (accessed 6 June 2025).

[65] Biometric Update, Deepfake detection advancing with multi-signal approach. https://www.biometricupdate.com/202412/deepfake-detection-advancing-with-multi-signal-approach, 2024 (accessed 6 June 2025).

[66] BioCatch, Why BioCatch. https://www.biocatch.com/why-biocatch?hsCtaTracking=a73571b6-f408-4a5f-981f-c9976932010a%7C22182ece-a206-47b4-bfd0-513b8cc46cea, 2024 (accessed 6 June 2025).

[67] M. Ghilom, S. Latifi, 2024. The role of machine learning in advanced biometric systems. Electronics. 13, 2667. https://doi.org/10.3390/electronics13132667.

[68] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, A. Morales, 2020. DeepfakesOn-Phys: deepfakes detection based on heart rate estimation. arXiv. arXiv:2010.00400. https://doi.org/10.48550/arXiv.2010.00400.

blinking,[70] breathing,[71] and pupil dilation.[72] Intel's FakeCatcher, which claims 96 % accuracy (a figure not independently peer-reviewed), exemplifies this approach.[73] However, synthesized realistic signals from advanced deepfake generators pose a future challenge, potentially impacting political deepfake detection.[74] Reliance on high-quality video restricts real-world use against low-resolution non-consensual deepfakes.[75] Individual physiological variations further complicate accurate analysis, potentially causing false alarms.[76]

2.2.3. Deep learning approaches to detection

Deep learning forms the bedrock of many contemporary deepfake detection systems, with firms like Sentinel[77] and Attestive[78] utilizing sophisticated architectures to analyze facial and vocal data. Convolutional Neural Networks (CNNs) excel at identifying frame-by-frame visual anomalies crucial in detecting manipulated videos used for financial fraud, while Recurrent Neural Networks (RNNs) are adept at capturing temporal inconsistencies, like lip-sync errors often present in political mis-information videos. Ironically, Generative Adversarial Networks (GANs), the very technology behind many advanced deepfakes, including those used to generate non-consensual imagery, are also employed in detection to identify subtle, AI-generated artifacts.[79]

However, the effectiveness of these methods is significantly impacted by insufficient or biased training data, leading to often significant poor generalization against unseen deepfakes across diverse applications.[80] Furthermore, these models exhibit vulnerability to adversarial attacks, subtle perturbations designed to deceive detection systems,[81] even high-accuracy ones like those claimed by Sensity AI (e.g., 98 %).[82] These figures reported by vendors and not peer-reviewed often reflect curated datasets, not real-world conditions.[83] The inherent "black box" nature of deep learning hinders transparency and error diagnosis, complicating efforts posing challenges for legal admissibility.[84] Evolving deepfake generation requires constant model adaptation against new threats, making explainability and robustness persistent critical limitations.

2.2.4. Hybrid multimodal detection

Integrating audio, video, image, and text, hybrid multimodal analysis also offers a significantly more robust and nuanced defence against sophisticated deepfakes.[85] Platforms like Reality Defender[86] and Sensity AI[87] use this to detect inconsistencies, such as synthesized voices with manipulated video in financial fraud schemes or nonsensical text with consistent visuals in political misinformation campaigns. Reality Defender concurrently analyses these modalities for voice clones, video manipulations, synthetic images, and AI-generated text via a probabilistic, watermark-independent method.[88] Sensity AI's multi-layered platform similarly uses advanced AI and machine learning for rapid assessment of video, images, audio, and identities to combat synthetic media misuse.[89] This approach overcomes unimodal limitations like subtle audio mismatches in non-consensual deepfakes that might otherwise go unnoticed.[90]

However, adversarial attacks targeting the data fusion process still pose a significant hurdle.[91] Reliance on accurate audio-visual synchronization is also vulnerable to subtle desynchronization within advanced deepfakes used for financial scams.[92] Computational cost severely limits real-time application against rapidly spreading political misinformation.[93] Another way to undermine detection is the threat of missing modalities, such as omitting video with fraudulent audio.[94] Dataset bias

[70] T. Jung, S. Kim, K. Kim, DeepVision: Deepfakes detection using human eye blinking pattern, IEEE Access. 8 (2020) 83144-83154. https://doi.org/10.1109/ACCESS.2020.2988660.

[71] T.-P. Doan, L. Nguyen-Vu, S. Jung, K. Hong, BTS-E: audio deepfake detection using breathing-talking-silence encoder, ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2023) 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095927.

[72] K. Patil, S. Kale, J. Dhokey, 2023. Deepfake detection using biological features: a survey. arXiv. arXiv:2301.05819. https://doi.org/10.48550/arXiv.2301.05819.

[73] Intel's FakeCatcher, The word's first-real time deepfake detector. https://download.intel.com/newsroom/2022/new-technologies/FakeCatcher-Infographic.pdf, 2022 (accessed 6 June 2025).

[74] X. Xiong, P. Patel, Q. Fan, A. Wadhwa, S. Selvam, X. Guo, L. Qi, X. Liu, R. Sengupta, 2025. TalkingHeadBench: a multi-modal benchmark & analysis of talking-head deepfake detection. arXiv. arXiv:2505.24866v1. https://doi.org/10.48550/arXiv.2505.24866.

[75] Intel's FakeCatcher, The word's first-real time deepfake detector. https://download.intel.com/newsroom/2022/new-technologies/FakeCatcher-Infographic.pdf, 2022 (accessed 6 June 2025).

[76] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, A. Morales, 2020. DeepfakesOn-Phys: deepfakes detection based on heart rate estimation. arXiv. arXiv:2010.00400. https://doi.org/10.48550/arXiv.2010.00400

[77] Sentinel, Defending against deepfakes and information warfare. https://thesentinel.ai/, 2020 (accessed 6 June 2025).

[78] Attestive, Deepfake video detection software. https://attestiv.com/deepfake-video-detection-software/, founded in 2018 (accessed 6 June 2025).

[79] T.T. Nguyen, Q.V.H. Nguyen, D.T. Nguyen, D.T. Nguyen, T. Huynh-The, S. Nahavandi, T.T. Nguyen, Q.-V. Pham, C.M. Nguyen, 2022. Deep learning for deepfakes creation and detection: a survey. Computer Vision and Image Understanding. 223, 103525. https://doi.org/10.1016/j.cviu.2022.103525.

[80] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, F. Xia, 2024. Deepfake video detection: challenges and opportunities. Artificial Intelligence Review. 57, 159. https://doi.org/10.1007/s10462-024-10810-6.

[81] Q. U. Ain, A. Javed, A. Irtaza, 2025. DeepEvader: An evasion tool for exposing the vulnerability of deepfake detectors using transferable facial distraction blackbox attack. Engineering Applications of Artificial Intelligence. 145, 110276. https://doi.org/10.1016/j.engappai.2025.110276.

[82] Sensity AI, All-in-one deepfake detection. https://sensity.ai/, founded 2018 (accessed 6 June 2025).

[83] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

[84] Q. U. Ain, A. Javed, A. Irtaza, 2025. DeepEvader: An evasion tool for exposing the vulnerability of deepfake detectors using transferable facial distraction blackbox attack. Engineering Applications of Artificial Intelligence. 145, 110276. https://doi.org/10.1016/j.engappai.2025.110276.

[85] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, & S. Tubaro, 2023. A robust approach to multimodal deepfake detection. Journal of Imaging. 9, 122. https://doi.org/10.3390/jimaging9060122.

[86] Reality Defender, https://www.realitydefender.com/, 2024 (accessed 6 June 2025).

[87] Sensity AI, All-in-one deepfake detection. https://sensity.ai/, founded 2018 (accessed 6 June 2025).

[88] Reality Defender, https://www.realitydefender.com/, 2024 (accessed 6 June 2025).

[89] Sensity AI, All-in-one deepfake detection. https://sensity.ai/, founded 2018 (accessed 6 June 2025).

[90] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, & S. Tubaro, 2023. A robust approach to multimodal deepfake detection. Journal of Imaging. 9, 122. https://doi.org/10.3390/jimaging9060122.

[91] S. Li, and H. Tang, 2024. Multimodal alignment and fusion: a survey. arXiv. arXiv:2411.17040v1. https://doi.org/10.48550/arXiv.2411.17040.

[92] J. Voas, W-C. Tseng, J. Stuedemann, L. Berry, X. Hu, D. Harwath, and P. Peng, 2024. Temporally streaming audio-visual synchronization for real-world videos. arXiv. arXiv:2402.04217. https://doi.org/10.1109/WACV61041.2025.00490.

[93] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, 2024. Deepfake video detection: challenges and opportunities. Artificial Intelligence Review. 57, 159. https://doi.org/10.1007/s10462-024-10810-6.

[94] C. Yu, P. Chen, J. Tian, J. Liu, J. Dai, X. Wang, Y. Chai, S. Jia, S. Lyu, J. Han, 2023. A unified framework for modality-agnostic deepfakes detection. arXiv. arXiv:2307.14491v2. https://doi.org/10.48550/arXiv.2307.14491.

and lack of explainability further complicate fairness and accountability.[95] Generalization to novel deepfakes remains a key challenge for these integrated systems.[96]

2.3. Emerging deepfake detection approaches

2.3.1. Liveness detection and Zero-Knowledge Biometrics (ZKB)

Liveness detection and Zero-Knowledge Biometrics (ZKB) offer distinct approaches to mitigating deepfake threats in biometric authentication. Liveness detection tools, like Facia, ensure interaction originates from a live person ("liveness" of a biometric sample), not a deepfake, thus preventing deepfake-based access to financial accounts.[97] However, the ability of increasingly realistic deepfakes to mimic "active liveness" (e.g., blinking for access control) and "passive liveness" (e.g., facial analysis for identity verification) poses a significant risk,[98] thereby potentially enabling non-consensual deepfake impersonations. While companies like Oz Forensics claim 100 % accuracy in preventing deepfakes,[99] these vendor-reported metrics often lack independent verification and peer-review. Furthermore, adversarial deepfakes exploiting specific algorithm vulnerabilities are a critical threat across applications.[100]

Keyless's ZKB prioritizes privacy by eliminating raw biometric data storage (e.g., faces or voices), thus mitigating the threat of stolen templates used for unauthorized access to secure systems.[101] However, while enhancing security, decentralized biometrics' ability to directly counter real-time deepfake presentation attacks during authentication for fraudulent transactions depends on the robustness of integrated liveness mechanisms.[102] The complexity of Secure Multi-Party Computation and reliance on infrastructure trust[103] also present vulnerabilities, potentially affecting reliability in critical applications. Both approaches require balancing security, user experience, and data protection as deepfake threats evolve.

2.3.2. Blockchain and quantum computing in deepfake mitigation

While not direct detection methods, blockchain and quantum computing offer distinct, albeit currently limited, techniques to bolster the fight against deepfakes. Blockchain technology, as implemented by tools such as WeVerify, primarily enhances media authenticity by establishing an immutable record of content origin and modifications,[104] theoretically aiding in tracing deepfakes used in financial fraud by impersonating company officials. However, its effectiveness against novel deepfakes deployed in political misinformation campaigns, particularly if the manipulation precedes blockchain recording, often remains limited.[105] Its core vulnerability lies in reliance on initial AI detection linked to the chain, which can be circumvented.[106]

Quantum Machine Learning (QML) presents a promising, albeit long-term, prospect for advanced deepfake analysis. Its potential to overcome computational bottlenecks could be invaluable for detecting subtle cues in sophisticated financial fraud deepfakes, hinging on minute inconsistencies.[107] Techniques like QT-CNN hold promise.[108] However, the complexity and real-world use of QML models against diverse deepfake techniques from subtle lip-sync alterations in political videos to intricate facial replacements in non-consensual media are heavily restricted by current quantum hardware limitations.[109] Effectively encoding classical media data into quantum states[110] and the current lack of explainability in QML models further impede their near-term utility in the broader deepfake detection landscape.[111]

2.3.3. Adversarial training for robustness

Adversarial training strengthens deepfake detection models against the critical threat of "adversarial perturbations",[112] often subtle manipulations designed to evade detection in applications like financial fraud CEO impersonations. It seeks to enhance robustness against attacks maintaining visual plausibility while deceiving detectors. Approaches

[95] M. Casu, L. Guarnera, P. Caponnetto, S. Battiato, 2024. GenAI mirage: the impostor bias and the deepfake detection challenge in the era of artificial illusions. Forensic Science International: Digital Investigation 50, 301795. https://doi.org/10.1016/j.fsidi.2024.301795.

[96] B. Li, J. Sun, C.M. Poskitt, X. Wang, 2024. How generalizable are deepfake image detectors? An empirical study. arXiv. arXiv:2308.04177v2. https://doi.org/10.48550/arXiv.2308.04177.

[97] Facia, Advanced deepfake detection deepfakes aren't real, but they are a reality. https://facia.ai/features/deepfake-detection/, founded in 2022 (accessed 6 June 2025).

[98] Oz Forensics, Passive and active liveness. https://doc.ozforensics.com/oz-knowledge/general/oz-platform/passive-and-active-liveness, 2025 (accessed 6 June 2025); Oz Forensics, Liveness detection for face recognition. https://ozforensics.com/our_solutions/oz_liveness, 2024 (accessed 6 June 2025).

[99] Oz Forensics, Face liveness detection and biometric software effectively prevent deepfake and spoofing attacks. https://ozforensics.com/#main_window, 2025 (accessed 6 June 2025).

[100] S. Khade, S. Ahirrao, S. Phansalkar, K. Kotecha, S. Gite, S.D. Thepade, 2021. Iris liveness detection for biometric authentication: a systematic literature review and future directions. Inventions. 6, 65. https://doi.org/10.3390/inventions6040065; M.R. Hasan, S.M.H. Mahmud, X.Y. Li, 2019. Face anti-spoofing using texture-based techniques and filtering methods. Journal of Physics: Conference Series. 1229, 012044. https://doi.org/10.1088/1742-6596/1229/1/012044.

[101] Keyless, 8 considerations for banks implementing biometric authentication White Paper. https://site.keyless.io/hubfs/Content/White%20Papers/2024%20Keyless%20-%20Banking%20White%20Paper.pdf, 2024 (accessed 6 June 2025).

[102] S.M. Abdullahi, S. Sun, B. Wang, N. Wei, H. Wang, 2024. Biometric template attacks and recent protection mechanisms: a survey. Information Fusion. 103, 102144. https://doi.org/10.1016/j.inffus.2023.102144.

[103] Keyless, Zero-Knowledge Biometrics™ the future of authentication. https://26689385.fs1.hubspotusercontent-eu1.net/hubfs/26689385/%5B2023%5D%20Downloadable%20Content/Keyless_Zero_Knowledge_Biometrics.pdf, 2023 (accessed 6 June 2025).

[104] Weverify, Deepfake detector. https://weverify.eu/tools/deepfake-detector/, founded in 2020 (accessed 6 June 2025); A. Heidari, N.J. Navimipour, H. Dag, S. Talebi, M. Unal, A novel blockchain-based deepfake detection method using federated and deep learning models, Cognitive Computation 16 (2024) 1073-1091. https://doi.org/10.1007/s12559-024-10255-7.

[105] M.M. Rashid, S-H. Lee, K-R. Kwon, Blockchain technology for combating deepfake and protect video/image integrity, Journal of Korea Multimedia Society 24(8) (2021) 1044-1058. https://doi.org/10.9717/kmms.2021.24.8.1044.

[106] F. Abbas, A. Taeihagh, 2024. Unmasking deepfakes: a systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications 252, Part B, 124260. https://doi.org/10.1016/j.eswa.2024.124260.

[107] C—H.A. Lin, C-Y. Liu, S.Y-C. Chen, K-C. Chen, 2024. Quantum-Trained Convolutional Neural Network for deepfake audio detection. arXiv. arXiv: 2410.09250v1. https://doi.org/10.48550/arXiv.2410.09250.

[108] B. Eray Katı, E. Uğur Küçüksille, G. Sarıman, 2025. Enhancing deepfake detection through Quantum Transfer Learning and Class-Attention Vision Transformer architecture. Applied Science. 15, 525. https://doi.org/10.3390/app15020525.

[109] M. Rath, H. Date, 2024. Quantum data encoding: a comparative analysis of classical-to-quantum mapping techniques and their impact on machine learning accuracy. EPJ Quantum Technology. 11, 72. https://doi.org/10.1140/epjqt/s40507-024-00285-3.

[110] M. Schuld, F. Petruccione, Supervised learning with quantum computers, Springer, Berlin, 2018.

[111] A. Kottahachchi Kankanamge Don, I. Khalil, 2025. QRLaXAI: quantum representation learning and explainable AI. Quantum Machine Intelligence. 7, 24. https://doi.org/10.1007/s42484-025-00253-9.

[112] P. Neekhara, B. Dolhansky, J. Bitton and C. C. Ferrer, Adversarial threats to deepfake detection: a practical perspective, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA (2021) 923-932. https://ieeexplore.ieee.org/document/9522903.

such as Adversarial Feature Similarity Learning aim at deepfakes used in political misinformation by optimizing resistance to subtle expression or lip movement alterations.[113] Mindgard AI's red teaming platform simulates real-world attack scenarios, including various evasion attack vectors relevant to non-consensual deepfakes, such as subtle morphing techniques designed to make the manipulation less obvious to automated detection.[114]

However, high computational cost still hinders rapid deployment against evolving application-specific attacks.[115] Insufficient dataset diversity leaves models vulnerable to novel manipulations across different domains.[116] Generalization to unseen adversarial attacks remains a key challenge; a model robust against warping in political deepfakes might fail against color manipulation in financial fraud.[117] The trade-off between robustness and accuracy on clean data is critical,[118] especially in non-consensual content detection where false positives have severe consequences. Measuring true robustness against evolving threats is also difficult.[119] Thus, adversarial training is most effective when integrated with other defence mechanisms and continuously evaluated against attack strategies relevant to specific deepfake applications.[120]

2.4. Content provenance: C2PA as a key approach

2.4.1. C2PA's role in deepfake detection

To combat harmful deepfakes, the Coalition for Content Provenance and Authenticity (C2PA) sets a new standard for online trust. C2PA records verifiable content creation and modification information, including AI tool use like deepfake generators.[121] This helps distinguish authentic media from deepfakes used in financial fraud (e.g., manipulated financial endorsements), political misinformation (e.g., AI-generated campaign videos), and non-consensual deepfakes (e.g., AI manipulated images). By providing clear provenance, C2PA labelling can reveal AI use or inconsistencies in content history, while complementary techniques like watermarking help trace origin and verify authenticity more robustly.[122]

C2PA's core principles including privacy, accessibility, interoperability, and security[123] align with global efforts like UN Resolution 78/265[124] and the EU AI Act (AIA).[125] These advocate for tools to identify AI content and empower users, potentially reducing deepfake impact on election integrity and personal reputation.

However, challenges remain. Manipulating provenance or watermarks is a risk for sophisticated actors seeking financial gain or spreading malicious political narratives.[126] Widespread adoption requires standardization[127] and interoperability across platforms,[128] ensuring consistent, verifiable C2PA information even on social media.

Industry leaders like Google are already integrating C2PA, recognizing its potential.[129] Continued development and collaboration are essential for long-term effectiveness in mitigating deepfake harms, from economic stability to individual safety.

2.4.2. C2PA labelling

C2PA leverages data provenance through "manifests"—digital "nutrition labels" attached to content.[130] These provide verifiable information on origin, history, and modifications, including generative AI use that might create harmful deepfakes.[131] Accessible via the "Content Credentials" icon, this history fosters a trustworthy online environment.[132]

Provenance is crucial for mitigating deepfakes. It helps raise red flags in financial fraud (e.g., AI-altered financial documents) and undermine credibility in political misinformation by tracing manipulated media to its AI origin. Google's "About this image" and YouTube's labels exemplify this context.[133] Provenance data can also indicate AI involvement

[113] S. Khan, J-C. Chen, W-H. Liao, C-S. Chen, 2024. Adversarially robust deepfake detection via Adversarial Feature Similarity Learning. arXiv. arXiv:2403.08806. https://doi.org/10.48550/arXiv.2403.08806.

[114] Mindgard, Bypassing AI-driven deepfake detection via evasion attacks. https://mindgard.ai/blog/bypassing-ai-driven-deepfake-detection-via-evasion-attacks, 2025 (accessed 6 June 2025).

[115] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, 2018. Towards deep learning models resistant to adversarial attacks. Statistics Machine Learning. arXiv. arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083.

[116] L. Jiang, R. Li, W. Wu, C. Qian, C.C. Loy, 2020. DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection. Computer Science, Computer Vision and Patter Recognition. arXiv. arXiv:2001.03024. https://doi.org/10.48550/arXiv.2001.03024; A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, 2019. FaceForensics++: learning to detect manipulated facial images. Computer Science, Computer Vision and Patter Recognition. arXiv. arXiv:1901.08971. https://doi.org/10.48550/arXiv.1901.08971.

[117] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, Z. Ge, 2023. Implicit identity leakage: the stumbling block to improving deepfake detection generalization, Computer Science, Computer Vision and Patter Recognition. arXiv. arXiv:2210.14457. https://doi.org/10.48550/arXiv.2210.14457.

[118] H. Zhang, Y. Yu, J. Jiao, E.P. Xing, L. El Ghaoui, M.I. Jordan, 2019. Theoretically principled trade-off between robustness and accuracy. Computer Science, Machine Learning. arXiv. arXiv:1901.08573. https://doi.org/10.48550/arXiv.1901.08573; A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, 2018. Towards deep learning models resistant to adversarial attacks. Statistics Machine Learning. arXiv. arXiv:1706.06083 https://doi.org/10.48550/arXiv.1706.06083.

[119] N. Carlini, D. Wagner, 2017. Adversarial examples are not easily detected: bypassing ten detection methods. Computer Science, Machine Learning. arXiv. arXiv:1705.07263. https://doi.org/10.48550/arXiv.1705.07263.

[120] S. Khan, J-C. Chen, W-H. Liao, C-S. Chen, 2024. Adversarially robust deepfake detection via Adversarial Feature Similarity Learning. arXiv. arXiv:2403.08806. https://doi.org/10.48550/arXiv.2403.08806.

[121] C2PA, Content Credentials. https://c2pa.org/post/contentcredentials/, 2024 (accessed 6 June 2025); CR, Content Credentials. https://contentcredentials.org/, 2024 (accessed 6 June 2025).

[122] C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).

[123] C2PA, Guiding Principles. https://c2pa.org/principles/, 2024 (accessed 6 June 2025).

[124] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[125] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending certain regulations ('the AI Act') [2024] OJ L 178/1, art 50(1)-(7), Recs 133-137.

[126] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

[127] Brookings, Detecting AI fingerprints: a guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/, 2024 (accessed 6 June 2025).

[128] US Department of Homeland Security, S&T Digital forgeries report technology landscape threat assessment January 24, 2023. https://www.dhs.gov/sites/default/files/2023-06/23_0630_st_digital_forgeries_report_signed.pdf, 2023 (accessed 6 June 2025).

[129] Google, How we're increasing transparency for gen AI content with the C2PA. https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/, 2024 (accessed 6 June 2025).

[130] C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).

[131] C2PA, Content Credentials. https://c2pa.org/post/contentcredentials/, 2024 (accessed 6 June 2025).

[132] CR, Content Credentials. https://contentcredentials.org/, 2024 (accessed 6 June 2025).

[133] Google, How we're increasing transparency for gen AI content with the C2PA. https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/, 2024 (accessed 6 June 2025); Google, More transparency for AI edits in Google Photos. https://blog.google/products/photos/ai-editing-transparency/, 2024 (accessed 6 June 2025).

in non-consensual deepfakes.

However, C2PA labelling faces challenges. Inconsistent standardization across platforms hinders deepfake identification.[134] Interpreting complex provenance data, such as timestamps and authorship, can be difficult for average users,[135] especially during rapid news cycles or financial scams. Over-reliance on labels as the sole authenticity indicator risks overlooking sophisticated deepfakes with manipulated or omitted data.[136] Scope is limited to provenance, potentially missing malicious intent[137] or misleading context, even with accurate labelling.[138] Privacy concerns exist regarding sensitive data collection,[139] and managing vast online media or "stripped" files impacts accuracy,[140] allowing fraudulent deepfakes to circulate undetected. Despite limitations, C2PA labelling is vital for understanding digital content by promoting transparency and empowering users.

2.4.3. *Digital watermarking: Enhancing content integrity*

Digital watermarking, while distinct from C2PA, complements it by embedding imperceptible (or visible) markers directly into content, offering robust detection for AI-generated media and enhancing provenance durability.[141] For example, Google's SynthID Detector identifies AI-generated images, audio, video, and text via these watermarks, tackling harmful deepfakes like financial fraud and political misinformation.[142] SynthID is designed to integrate with C2PA's Content Credentials through "soft bindings," which allow watermarks to recover or reference associated provenance even if metadata is detached.[143]

However, problems persist. Watermarks can be removed, and the "liar's dividend" phenomenon, where authentic content is falsely claimed as AI-generated, further erodes trust.[144] The rapid rise of deepfakes, particularly in audio scams and non-consensual visual manipulations, demands continuous advancements in watermarking and detection.[145] Furthermore, unlike open-source initiatives like Google DeepMind's SynthID Text,[146] which fosters collaborative development, proprietary tools like Steg.AI[147] hinder standardization. Limited training data creates vulnerabilities, as a detector trained on one AI model (e.g., Gemini) may fail on another (e.g., ChatGPT), highlighting a need for broader interoperability.[148]

Ultimately, C2PA's provenance provides crucial historical context, while watermarking offers a persistent, on-content link to that origin, enhancing provenance data resilience. Inconsistencies between provenance and watermark detection trigger scrutiny, boosting overall reliability and deepfake identification.

2.5. *Overarching technical challenges and ethical considerations*

The deployment of deepfake detection tools faces significant hurdles related to their technical limitations and broader societal impacts.

2.5.1. *Challenges of legal admissibility*

Deepfake detection tools face significant legal admissibility challenges. Vendor-reported high accuracy rates (e.g., Intel's FakeCatcher at 96 %,[149] Sensity AI at 98 %,[150] Oz Forensics at 100 %)[151] frequently lack independent validation and peer-review, raising concerns about their reliability and admissibility in court. The US Supreme Court's *Daubert*[152] standard mandates that scientific evidence be reliable and relevant, considering testability, peer-review, error rates, and general scientific acceptance. Federal Rule of Evidence 901(a) requires authentication, necessitating robust proof of the detection method's accuracy beyond mere demonstration of manipulation.[153] Cases like *US v Reffitt*[154] and

[134] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025).

[135] Hacker Factor, C2PA's Time Warp. https://www.hackerfactor.com/blog/index.php?/archives/1023-C2PAs-Time-Warp.html, 2025 (accessed 6 June 2025).

[136] Google, Determining trustworthiness through context and provenance. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/determining_trustworthiness_en.pdf, 2024 (accessed 6 June 2025).

[137] Ibid.

[138] L. Fazio, Out-of-context photos are a powerful low-tech form of misinformation, The Conversation. https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959, 2020 (accessed 6 June 2025).

[139] Brookings, Detecting AI fingerprints: a guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/, 2024 (accessed 6 June 2025); WITNESS, Tomorrow's great digital divide: content with or without provenance. https://blog.witness.org/2025/03/tomorrows-great-digital-divide/, 2025 (accessed 6 June 2025); C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).

[140] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

[141] C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).

[142] Google, SynthID Detector — a new portal to help identify AI-generated content. https://blog.google/technology/ai/google-synthid-ai-content-detector/ introduced in May 2025 (accessed 6 June 2025); Google DeepMind's SynthID, Identifying AI-generated content with SynthID. https://deepmind.google/technologies/synthid/, introduced in August 2023 (accessed 6 June 2025).

[143] C2PA, Technical Specification - Soft Bindings. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html#_soft_bindings, 2024 (accessed 6 June 2025).

[144] Brookings, Detecting AI fingerprints: a guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/, 2024 (accessed 6 June 2025).

[145] Ibid.

[146] Google AI for developers, SynthID: tools for watermarking and detecting LLM-generated text. https://ai.google.dev/responsible/docs/safeguards/synthid, introduced in October 2024 (accessed 6 June 2025).

[147] Steg.AI, Forensic watermarking for digital media. https://steg.ai/, founded in 2019 (accessed 6 June 2025).

[148] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

[149] Intel's FakeCatcher, The word's first-real time deepfake detector. https://download.intel.com/newsroom/2022/new-technologies/FakeCatcher-Infographic.pdf, 2022 (accessed 6 June 2025).

[150] Sensity AI, All-in-one deepfake detection. https://sensity.ai/, founded 2018 (accessed 6 June 2025).

[151] Oz Forensics, Face liveness detection and biometric software effectively prevent deepfake and spoofing attacks. https://ozforensics.com/#main_window, 2025 (accessed 6 June 2025).

[152] *Daubert v Merrell Dow Pharmaceuticals Inc.*, 509 US 579, 589–595 (1993). Under this standard, which draws from Federal Rule of Evidence 702, scientific evidence (e.g., deepfake detection methodologies) must meet criteria including testability (e.g., robust datasets and validation protocols), peer review, known error rates (e.g., performance metrics and statistical analysis), and general acceptance in the scientific community.

[153] US Federal Rule of Evidence 901(a).

[154] *United States v Reffitt* 602 F Supp 3d 85 (DDC 2022). This case, addressing 'innuendo about altered videos', involved disputed video evidence sometimes requiring forensic verification. It underscored jury verdicts' reliance on credible evidence, stressing the legal need to verify complex digital evidence (potentially from vendor tools) and ensure robust authentication against mere manipulation claims.

Huang v Tesla[155] underscore the critical need for robust, independently verifiable evidence, especially from vendors. Furthermore, the UK's *Emotional Perception v Comptroller*[156] case demonstrates how AI opacity complicates understanding decision-making, impacting legal transparency and explainability. Limited human accuracy in audio deepfake detection (73 %) compounds the challenge,[157] reinforcing that admissibility requires independently verified reliability and transparent disclosure, a standard current vendor-driven metrics often fail to provide.

2.5.2. Bias in deepfake detection

Beyond technical challenges, bias is a major concern in deepfake detection. While some tools, like HyperVerge, claim race, age, and gender agnosticism,[158] real-world performance often reveals disparities. Similar to other AI applications such as facial recognition,[159] deepfake detection used in non-consensual instances raises questions about fairness and discrimination. For example, in the UK's *R (Bridges)* case involving police use of facial recognition, the software correctly identified only 34 % of men and just 18 % of women, with significantly higher false positives for women (82 % vs 66 %).[160] A study using the FaceForensic++ dataset and the Xception algorithm further revealed significant disparities: Black men were misclassified as fake 39.1 % of the time, compared to 15.6 % for white women.[161] Biased training data can cause models to misclassify deepfakes based on protected characteristics, leading to unfair outcomes. Overfitting and underfitting of biased datasets compound the difficulties of creating reliable, unbiased deepfake detection,[162] which is essential to mitigate societal inequalities.

2.5.3. Data protection and biometric data in detection

Deepfake detection tools like BioCatch,[163] Attestive,[164] and Sentinel,[165] which collect and analyze biometric data (e.g., keystroke dynamics, facial expressions), often struggle to adhere to stringent data protection frameworks such as the GDPR. The GDPR mandates lawful processing, data protection by design and default, and Data Protection Impact Assessments.[166] This starkly contrasts with privacy-prioritizing solutions like Keyless' Zero-Knowledge Biometrics™, which eliminate raw biometric data storage.[167] Emphasizing less restrictive data processing, the strict necessity and proportionality principles highlighted by the Court of Justice of the EU (CJEU) in cases like *KNLTB*[168] and *HTB Neunte Immobilien Portfolio*[169] demand that deepfake detection methods, especially those involving biometric data, prioritize Privacy-Enhancing Technologies (PETs).[170] Global perspectives, seen in China's Deep Synthesis Provisions (requiring consent for biometric information)[171] and the principles from the Chinese *Guo Bing* case,[172] also underscore the importance of robust legal frameworks for biometric data use.

2.5.4. Free speech and copyright tension in deepfake detection

Accurate, rapid deepfake detection is challenging because platform pressure to remove harmful content (e.g., non-consensual pornography, election disinformation) risks over-removing legitimate content. This infringement of free expression is a key concern – seen, for instance, in the US *Kohls v Bonta*[173] case regarding political deepfakes – and creates challenges under intermediary liability frameworks like the EU Digital Services Act[174] and the UK Online Safety Act.[175] The reliance on imperfect detection methods can lead to censorship of protected speech (a risk highlighted by over-blocking issues in *Cartier v BT*),[176] underscoring the need for flexible technical measures as noted by CJEU *UPC Telekabel*.[177] Separately, using copyrighted material in training datasets

[155] *Huang v Tesla Inc* (Cal Super Ct, 26 April 2019) 19CV346663, [43]. This case demonstrates that proving technical defects requires diverse post-incident data, such as accident reports and regulatory findings. Analogously, deepfake detection evidence must move beyond mere allegations, relying instead on robust forensic analysis, validated methodologies, and verifiable data to be admissible.

[156] *Comptroller General of Patents, Designs and Trade Marks v Emotional Perception AI Ltd* [2024] EWCA Civ 825 [51], [58], [79].

[157] K. T. Mai, S. Bray, T. Davies, L. D. Griffin, 2023. Warning: humans cannot reliably detect speech deepfakes, PLoS ONE. 18, e0285333. https://doi.org/10.1371/journal.pone.0285333.

[158] HyperVerge. https://hyperverge.co/, (accessed 6 June 2025).

[159] For discussion on facial recognition technology's implications for discrimination, fairness, and data protection, including GDPR and contemporary use cases, see F. Romero-Moreno, Facial recognition technology: how it's being used in Ukraine and why it's still so controversial, The Conversation. https://theconversation.com/facial-recognition-technology-how-its-being-used-in-ukraine-and-why-its-still-so-controversial-183171, 2022 (accessed 6 June 2025); F. Romero-Moreno, AI facial recognition and biometric detection: balancing consumer rights and corporate interests, In 2021 International Carnahan Conference on Security Technology (ICCST) IEEE (2021) 1-5. https://ieeexplore.ieee.org/document/9717403.

[160] *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police* [2020] EWCA Civ 1058 [188].

[161] Y. Ju, S. Hu, S. Jia, G. H. Chen, S. Lyu, 2023. Improving fairness in deepfake detection. arXiv. arXiv:2306.16635v3. https://doi.org/10.48550/arXiv.2306.16635.

[162] Y. Xu, P. Terhörst, K. Raja, M. Pedersen, 2024. Analyzing fairness in deepfake detection with massively annotated databases, arXiv. arXiv:2208.05845. https://arxiv.org/abs/2208.05845.

[163] BioCatch, Why BioCatch. https://www.biocatch.com/why-biocatch?hsCtaTracking=a73571b6-f408-4a5f-981f-c9976932010a%7C22182ece-a206-47b4-bfd0-513b8cc46cea, 2024 (accessed 6 June 2025).

[164] Attestive, Deepfake video detection software. https://attestiv.com/deepfake-video-detection-software/, founded in 2018 (accessed 6 June 2025).

[165] Sentinel, Defending against deepfakes and information warfare. https://thesentinel.ai/, 2020 (accessed 6 June 2025).

[166] See GDPR, arts 5(1)(a), 6 (regarding lawful processing); art 25 (regarding data protection by design and default); and art 35 (regarding Data Protection Impact Assessments).

[167] Keyless, 8 considerations for banks implementing biometric authentication White Paper. https://site.keyless.io/hubfs/Content/White%20Papers/2024%20Keyless%20-%20Banking%20White%20Paper.pdf, 2024 (accessed 6 June 2025).

[168] Case C-621/22 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECLI:EU:C:2024:857 [42], [51], [57], [58].

[169] Joined Cases C-17/22 and C-18/22 *HTB Neunte Immobilien Portfolio geschlossene Investment UG & Co. KG and Ökorenta Neue Energien Ökostabil IV geschlossene Investment GmbH & Co. KG v Müller Rechtsanwaltsgesellschaft mbH and Others* [2024] EU:C:2024:738 [51], [59], [73], [74], [76], [78].

[170] Information Commissioner's Office, Privacy-enhancing technologies (PETs). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/, 2023 (accessed 6 June 2025).

[171] Provisions on the Administration of Deep Synthesis Internet Information Services (Order No 12 of the Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security, 25 November 2022). <http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm> accessed 6 June 2025.

[172] Guo Bing v Hangzhou Wildlife World Co Ltd (Hangzhou Fuyang District People's Court, (2019) Zhe 0111 Min Chu No 6971, 20 November 2020) https://www.chinajusticeobserver.com/law/x/2019-zhe-jiang-0001-min-chu-no-6971 accessed 6 June 2025.

[173] *Kohls v Bonta*, 2:24-cv-02527 (ED Cal 2024).

[174] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1.

[175] Online Safety Act 2023.

[176] *Cartier International AG v British Telecommunications Plc* [2018] UKSC 28 [5].

[177] Case C-314/12 *UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH and Wega Filmproduktionsgesellschaft mbH*, [2014] EU:C:2014:192 [52].

implicates intellectual property rights,[178] raising fair use and parody questions that echo CJEU cases like *SABAM v Scarlet*[179] and *Netlog*.[180] Consequently, as stressed in the Advocate General's opinion in *Poland v Council and Parliament*,[181] minimizing "false positives" in detection methods is crucial to balance the fight against damaging deepfakes with preserving free speech and copyright protection.

2.5.5. Real-world deployment challenges and the detection divide

Deepfake detection faces an evolving challenge as rapid advancements make telltale signs (e.g., lighting inconsistencies, unnatural blinking) less apparent for both human and automated analysis.[182] A "deepfake divide"[183] exists due to the inaccessibility of sophisticated detection tools to the public (cost, complexity, or limited availability).[184] Issues like "stripped" files lacking metadata hindering analysis, particularly given training data limitations.[185] Interoperability is a key concern: classifiers for specific platforms (e.g., ElevenLabs) often fail with deepfakes from other tools,[186] and even broader tools like Microsoft's Video Authenticator can struggle across different AI models.[187] Alarmingly, these classifiers are themselves vulnerable to manipulation.[188] The absence of standardized watermarks further hinders coordinated detection.[189] This complex landscape necessitates robust, adaptable deepfake detection mechanisms that, per CJEU jurisprudence (notably the *UPC Telekabel*[190] ruling), must be "sufficiently effective" to genuinely protect against harmful deepfakes by preventing or significantly discouraging their dissemination.

2.6. Ensuring trustworthy deepfake detection: XAI, C2PA, and human rights

The pervasive rise of deepfakes, created for malicious purposes ranging from financial fraud to democratic manipulation and online harassment, poses a significant threat by eroding public trust and undermining human rights. Central to combating this evolving threat are explainable AI (XAI)[191] and the Coalition for Content Provenance and Authenticity (C2PA),[192] which integrates provenance labelling and digital watermarking technologies. These crucial technical tools tackle fundamental problems: ensuring the legal admissibility of manipulated media through XAI's transparency, combating detection bias for fairer outcomes via XAI, establishing robust data protection frameworks, mitigating over-removal of legitimate content through C2PA's verifiable origins and embedded watermarks, and bridging the "deepfake divide"[193] to ensure wider access to detection tools. Given the emphasis in leading legal frameworks such as the EU on trustworthy, accountable, and rights-respecting AI systems, a holistic and integrated approach to XAI is key, ensuring alignment of explanation methods with these demands.[194]

2.6.1. XAI for deepfake detection and its challenges

XAI is paramount for trustworthy deepfake detection, making AI decision-making transparent and understandable,[195] vital for legal reliability and reducing erroneous removals via informed human scrutiny. This transparency directly aligns with GDPR's principles of fairness and transparency, ensuring individuals can comprehend how AI assesses media.[196] By pinpointing salient features like inconsistent lip-sync and audio artifacts,[197] offering visual/textual explanations such as heatmaps,[198] and generating natural language interpretations of AI

[178] B. Zi, M. Chang, J. Chen, X. Ma, Y.-G. Jiang, 2024. WildDeepfake: a challenging real-world dataset for deepfake detection. arXiv. arXiv:2101.01456v2. https://arxiv.org/abs/2101.01456; A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, Face Forensics++: learning to detect manipulated facial images, Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 1-11. https://ieeexplore.ieee.org/document/9010912.

[179] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [52].

[180] Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [50].

[181] AG opinion in Case C-401/19 *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [AG 214].

[182] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: a survey of face manipulation and fake detection, Information Fusion 64 (2020) 131-148. https://doi.org/10.1016/j.inffus.2020.06.014.

[183] D. B. Bitton, C. P. Hoffmann, A. Godulla, Deepfakes in the context of AI inequalities: analyzing disparities in knowledge and attitudes, Information, Communication & Society 28(2) (2025) 295-315. https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2420037.

[184] WITNESS, Tomorrow's great digital divide: content with or without provenance. https://blog.witness.org/2025/03/tomorrows-great-digital-divide/, 2025 (accessed 6 June 2025).

[185] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

[186] Ibid.

[187] US Department of Homeland Security, S&T Digital forgeries report technology landscape threat assessment January 24, 2023. https://www.dhs.gov/sites/default/files/2023-06/23_0630_st_digital_forgeries_report_signed.pdf, 2023 (accessed 6 June 2025).

[188] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

[189] Brookings, Detecting AI fingerprints: a guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/, 2024 (accessed 6 June 2025).

[190] Case C-314/12 *UPC Telekabel Wien GmbH v Constantin Film Verleih GmbH and Wega Filmproduktionsgesellschaft mbH*, [2014] EU:C:2014:192 [62].

[191] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, 2023. Explainable Artificial Intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Information Fusion. 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805.

[192] C2PA. https://c2pa.org/, founded in 2021 (accessed 6 June 2025).

[193] D. B. Bitton, C. P. Hoffmann, A. Godulla, Deepfakes in the context of AI inequalities: analyzing disparities in knowledge and attitudes, Information, Communication & Society 28(2) (2025) 295-315. https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2420037.

[194] A. Bringas Colmenarejo, L. State, G. Comandé, How should an explanation be? A mapping of technical and legal desiderata of explanations for machine learning models, International Review of Law, Computers & Technology (2025) 1-32. https://doi.org/10.1080/13600869.2025.2497633.

[195] M. T. Ribeiro, S. Singh, C. Guestrin, 'Why should I trust you?': explaining the predictions of any classifier, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016) 1135-1144. https://doi.org/10.1145/2939672.2939778; R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, 2018. A survey of methods for explaining black box models. ACM Computing Surveys. 51, 93. https://doi.org/10.1145/3236009; W. Ge, J. Patino, M. Todisco, N. Evans, 2024. Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations. arXiv. arXiv:2110.03309v2. https://doi.org/10.48550/arXiv.2110.03309.

[196] XAI's contribution to transparency directly supports the GDPR's principles of fairness and transparency in data processing (GDPR, art 5(1)(a); art 12). This enables individuals to comprehend how AI systems assess media, which is crucial for exercising rights to information and access concerning automated processing (arts 13(2)(f), 14(2)(g), 15(1)(h); see also art 22 and recs 60, 63, 71).

[197] S.K. Datta, S. Jia, S. Lyu, 2024. Exposing lip-syncing deepfakes from mouth inconsistencies. arXiv. arXiv:2401.10113v2. https://doi.org/10.48550/arXiv.2401.10113.

[198] I. Ul Haq, K.M. Malik, K. Muhammad, 2024. Multimodal neurosymbolic approach for explainable deepfake detection. ACM Transactions on Multimedia Computing Communications and Applications. 20, 341. https://doi.org/10.1145/3624748.

reasoning,[199] XAI enables human scrutiny to correct potential errors in deepfake detection. For instance, XAI allows human reviewers to override false positives by explaining its detection reasoning, such as flagging subtle lighting variations, preventing over-removal of content that might be part of an original artistic style.[200] This explainability facilitates improvements in precision, recall, and F1-score,[201] as shown by companies like DuckDuckGoose[202] and Reality Defender,[203] potentially meeting US *Daubert*[204] evidential standards.

Furthermore, XAI aids in identifying and mitigating bias within training data,[205] thus promoting fairer, non-discriminatory outcomes in sensitive cases like non-consensual deepfakes and upholding human rights (e.g., European Convention on Human Rights (ECHR)[206] Article 14, EU Charter of Fundamental Rights (EU Charter)[207] Articles 20–23). However, deepfake detection tools face copyright infringement concerns if trained on copyrighted material without consent, attribution or compensation.[208] XAI transparency also enhances accountability in data processing, supporting users' right to understand their data usage.[209]

The CJEU's emphasis on strict necessity and proportionality principles necessitates that deepfake detection, particularly when it involves biometric data, utilizes the least restrictive data processing means.[210] This includes data minimization, purpose limitation, and security, frequently achieved using PETs.[211] Despite its promise, XAI faces several challenges: subjective interpretability[212] (especially in complex fraud), lack of standardization hindering cross-platform comparisons[213] (e.g., political content analysis), vulnerability to adversarial attacks enabling sophisticated forgery evasion,[214] and the nascent state of audio XAI, which is critical for detecting voice cloning in scams due to complex audio data and absent standardized features.[215]

2.6.2. C2PA: Challenges and regulatory imperatives

While XAI focuses on the transparency and reliability of detection mechanisms, C2PA addresses the authenticity of the content itself, providing verifiable information about its origin and modifications.[216] This empowers individuals to assess content reliability and respects fundamental rights to freedom of expression and access to information (ECHR Article 10, EU Charter Article 11). For instance, C2PA can verify the origin of news videos, providing strong evidence of authenticity aiding in identifying deepfakes and establishing legal validity.[217] This strengthens the rule of law (ECHR Article 6, EU Charter Article 47) and significantly contributes to reducing over-removal by providing verifiable provenance. Moreover, C2PA's focus on verifiable origins supports data protection principles by establishing data legitimacy, and privacy-preserving techniques align with GDPR's data minimization.[218]

However, C2PA faces challenges such as the potential for metadata tampering[219] and the significant hurdle of achieving widespread adop-

[199] G. Huang, Y. Li, S. Jameel, Y. Long, G. Papanastasiou, From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?, Computational and Structural Biotechnology Journal 24 (2024) 362-373. https://doi.org/10.1016/j.csbj.2024.05.004.

[200] R. Uma Maheshwari, B. Paulchamy, Securing online integrity: a hybrid approach to deepfake detection and removal using explainable AI and adversarial robustness training, Automatika 65(4) (2024) 1517–1532. https://doi.org/10.1080/00051144.2024.2400640.

[201] N. Mansoor, A. I. Iliev, 2025. Explainable AI for deepfake detection. Applied Sciences. 15, 725. https://doi.org/10.3390/app15020725.

[202] DuckDuckGoose AI, https://www.duckduckgoose.ai/, founded in 2020 (accessed 6 June 2025).

[203] Reality Defender, Visual deepfake detection explainability. https://www.realitydefender.com/blog/visual-deepfake-detection-explainability, 2024 (accessed 6 June 2025).

[204] *Daubert v Merrell Dow Pharmaceuticals Inc.,* 509 US 579, 589–595 (1993).

[205] B. van Stein, D. Vermetten, F. Caraffini, A.V. Kononova, 2023. Deep-BIAS: detecting structural bias using explainable AI. arXiv. arXiv:2304.01869. https://doi.org/10.48550/arXiv.2304.01869.

[206] Convention for the Protection of Human Rights and Fundamental Freedoms (adopted 4 November 1950, entered into force 3 September 1953) 213 UNTS 221 (European Convention on Human Rights, as amended) (ECHR).

[207] Charter of Fundamental Rights of the European Union [2012] OJ C 326/391 (CFR).

[208] J. Collomosse, A. Parsons, To authenticity, and beyond! Building safe and fair generative AI upon the three pillars of provenance, IEEE Computer Graphics and Applications 44(3) (2024) 82–90. https://dl.acm.org/doi/10.1109/MCG.2024.3380168.

[209] XAI's enhancement of transparency contributes to accountability in data processing (GDPR, arts 5(2), 24) and supports users' rights to information regarding their data usage, including the right of access and to understand the logic of processing (arts 12, 13, 14, 15; see also recs 60, 63).

[210] Case C-621/22 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECLI:EU:C:2024:857 [42], [51], [57], [58]; Joined Cases C-17/22 and C-18/22 *HTB Neunte Immobilien Portfolio geschlossene Investment UG & Co. KG and Ökorenta Neue Energien Ökostabil IV geschlossene Investment GmbH & Co. KG v Müller Rechtsanwaltsgesellschaft mbH and Others* [2024] EU:C:2024:738 [51], [59], [73], [74], [76], [78]; Case C-205/21 *Ministerstvo na vatreshnite raboti (Enregistrement de données biométriques et génétiques par la police)* [2023] ECLI:EU:C:2023:49 [126]-[128], [133].

[211] Information Commissioner's Office, Privacy-enhancing technologies (PETs). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/, 2023 (accessed 6 June 2025).

[212] F. Doshi-Velez, B. Kim, 2017. Towards a rigorous science of interpretable machine learning. arXiv. arXiv:1702.08608v2. https://doi.org/10.48550/arXiv.1702.08608.

[213] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M.B. Amin, B. Kang, Survey on explainable AI: from approaches, limitations and applications aspects, Human-Centric Intelligent Systems 3 (2023) 161-188. https://doi.org/10.1007/s44230-023-00038-y; L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, 2024. Explainable Artificial Intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Information Fusion. 106, 102301. https://doi.org/10.1016/j.inffus.2024.102301; O. Benchekroun, A. Rahimi, Q. Zhang, T. Kodliuk, 2020. The need for standardized Explainability. arXiv. arXiv:2010.11273v2. https://doi.org/10.48550/arXiv.2010.11273.

[214] R. Kozik, M. Ficco, A. Pawlicka, M. Pawlicki, F. Palmieri, M. Choraś, 2024. When explainability turns into a threat - using XAI to fool a fake news detection method. Computer & Security. 137, 103599. https://doi.org/10.1016/j.cose.2023.103599.

[215] S.-Y. Lim, D.-K. Chae, S.-C. Lee, 2022. Detecting deepfake voice using explainable deep learning techniques. Applied Sciences. 12, 3926. https://doi.org/10.3390/app12083926.

[216] C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).

[217] CMSWIRE, Fighting deepfakes with Content Credentials and C2PA. https://www.cmswire.com/digital-experience/fighting-deepfakes-with-content-credentials-and-c2pa/, 2024 (accessed 6 June 2025).

[218] Indeed, C2PA's focus on verifiable origins supports data protection principles by helping to establish data legitimacy, aligning with requirements for lawfulness, fairness, and transparency (GDPR, art 5(1)(a)), accuracy (art 5(1)(d)), and lawful basis for processing (art 6), thereby supporting accountability (arts 5(2), 24). Furthermore, privacy-preserving techniques are consistent with data minimization (art 5(1)(c)) and data protection by design and by default (art 25).

[219] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

tion.[220] CJEU rulings concerning *SCHUFA*[221] and *Dun & Bradstreet*[222] emphasize transparency and explainability in AI decisions, principles the European Data Protection Supervisor (EDPS) applies to deepfake detection tools,[223] aligning with the GDPR and respecting ECHR Article 8 and EU Charter Articles 7–8.

Legal frameworks must mandate design-stage trade-offs in deepfake detection, balancing accuracy with fundamental rights like privacy, fairness, and explainability, particularly for biometric data under necessity and proportionality principles.[224] XAI's transparency and C2PA's verifiable provenance are key to contesting erroneous classifications, mitigating over-removal, and protecting free expression, while prioritizing PETs to minimize sensitive biometric data processing.

The absence of harmonized international laws significantly hinders global efforts against cross-border disinformation. Sustained research and cross-sector collaboration are essential to mitigate deepfake harms, safeguard human rights, and ensure a digital environment where trust and authenticity can be reliably established.[225] These regulatory frameworks will be analyzed next.

3. The fragmented legal landscape of deepfake detection regulation

3.1. The EU Artificial Intelligence Act: A case of regulatory ambiguity for deepfake detection

3.1.1. A regulatory blind spot in risk-based AI

A critical contradiction is evident within the EU AI Act (AIA) concerning deepfake detection. While the Act defines deepfakes as synthetically generated or manipulated video, audio, or images, convincingly replicating real people, objects, places, or events (Article 3 (60)),[226] it neglects to establish a balance between the previously discussed deepfake detection methods and techniques and individual rights protection. Despite adopting a risk-based regulatory approach,[227] the Act lacks specific provisions for deepfake detection tools. This gap poses significant challenges, particularly when considering transparency obligations and the Act's "limited-risk" classification.

This is compounded by conflicting regulatory approaches to AI used in electoral disinformation. Recitals 120, 136 AIA suggest a "systemic-risk" categorization under the Digital Services Act (DSA), placing the onus on platforms. Conversely, Recital 62 AIA classifies such AI as "high-risk." This ambiguity creates legal uncertainty regarding precedence, specifically whether the AIA's "high-risk" classification supersedes the DSA's "systemic-risk" designation.[228] This could lead to forum shopping (e.g., exploiting lax environments for deepfake political propaganda) and inconsistent enforcement across EU Member States,[229] ultimately weakening protection against AI-driven electoral manipulation. A "systemic-risk" approach might prioritize platform-level detection, neglecting specific tool requirements, while "high-risk" compliance burdens could disproportionately impact smaller developers and stifle innovation.

The AIA's "unacceptable-risk" category (Article 5) prohibits manipulative AI systems, including deceptive AI (Article 5(1)(a)), which encompasses fraudulent deepfakes exploiting cognitive biases.[230] Recent cases, including the Almendralejo sextortion case,[231] a viral Pentagon explosion deepfake affecting US stock markets (both 2023),[232] and $35 million lost by over 6000 individuals across UK, Europe, and Canada since May 2022 to 2025,[233] underscore the devastating real-world impact of such deepfakes. This is particularly true when deepfakes target vulnerable individuals in emotionally manipulative contexts, such as romance scams and immersive virtual reality (Recital 29 AIA), where tools like Deep Nostalgia AI[234] demonstrate their potential for causing emotional distress and significant financial harm. While

[220] Brookings, Detecting AI fingerprints: a guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/, 2024 (accessed 6 June 2025).

[221] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [3], [56], [57], [59]; AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 54], [AG 57], [AG 58].

[222] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [38]-[76].

[223] European Data Protection Supervisor, Deepfake detection. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en, 2024 (accessed 6 June 2025).

[224] Information Commissioner's Office, Guidance on AI and data protection. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/, 2023 (accessed 6 June 2025).

[225] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025).

[226] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending certain regulations ('the AI Act') [2024] OJ L 178/1, art 3(60) defines 'deep fake' as 'AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful'. See also rec 134. For a detailed analysis of the AIA deepfake definition, see M. Labuz, Deep fakes and the Artificial Intelligence Act—an important signal or a missed opportunity?, Policy & Internet 16(4) (2024) 1-18. https://onlinelibrary.wiley.com/doi/10.1002/poi3.406; M. Labuz, A teleological interpretation of the definition of deepfakes in the EU Artificial Intelligence Act—a purpose-based approach to potential problems with the word "existing", Policy & Internet 17(1) (2025) 1-14. https://onlinelibrary.wiley.com/doi/10.1002/poi3.435.

[227] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending certain regulations ('the AI Act') [2024] OJ L 178/1, arts 5, 6, 50, Annexes I, III. The AI Act categorizes AI systems by risk: 'unacceptable risk' systems are banned (art 5); 'high risk' systems are regulated (art 6, Annexes I, III); and 'transparency risk' systems have transparency obligations (art 50). 'Minimal to no risk' systems are largely unregulated. See also Commission, 'Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 884 final (4 February 2025) p 1.

[228] F. Romero-Moreno, Generative AI and deepfakes: a human rights approach to tackling harmful content, International Review of Law, Computers & Technology 38(3) (2024) 297-326. https://doi.org/10.1080/13600869.2024.2324540.

[229] European Parliament, Criminal procedural laws across the European Union – a comparative analysis of selected main differences and the impact they have over the development of EU legislation. https://www.europarl.europa.eu/RegData/etudes/STUD/2018/604977/IPOL_STU(2018)604977_EN.pdf, 2018 (accessed 6 June 2025).

[230] For a detailed analysis of the AIA's "unacceptable-risk" category (Article 5), which prohibits manipulative AI systems, see M. Leiser, Psychological patterns and Article 5 of the AI Act: AI-powered deceptive design in the system architecture and the user interface, Journal of AI Law and Regulation 1(1) (2024) 5-23. https://doi.org/10.21552/aire/2024/1/4.

[231] A. M. Narvali, J. A. Skorburg, M. J. Goldenberg, Cyberbullying girls with pornographic deepfakes is a form of misogyny, The Conversation. https://theconversation.com/cyberbullying-girls-with-pornographic-deepfakes-is-a-form-of-misogyny-217182, 2023 (accessed 6 June 2025).

[232] Financial Times, Investors must beware deepfake market manipulation. https://www.ft.com/content/7b352945-9295-42f5-a5d1-a01edf48ba51, 2023 (accessed 6 June 2025).

[233] The Guardian, Revealed: the scammers who conned savers out of $35m using fake celebrity ads. https://www.theguardian.com/money/2025/mar/05/revealed-the-scammers-who-conned-savers-out-of-35m-using-fake-celebrity-ads, 2025 (accessed 6 June 2025).

[234] Deep Nostalgia AI, Revive your memories with Deep Nostalgia AI. https://deep-nostalgia-ai.com/, 2021 (accessed 6 June 2025).

Commission non-binding guidelines on prohibited AI practices suggest deceptive deepfakes *could* fall under this category,[235] the AIA should explicitly classify them as "unacceptable-risk" for legal certainty and to prevent malicious use.

The Commission guidelines rightly clarify the insufficiency of transparency measures alone in regulating deepfakes.[236] Labelling, while informative, does not inherently prevent manipulation, since users remain susceptible to cognitive biases as recognized in CJEU *Compass Banca*.[237] Therefore, these guidelines stress that the AIA's general prohibition (Article 5(1)(a)) allows banning even labelled AI if it causes significant harm.[238]

Critically, the Commission guidelines emphasize the need for clearer articulation of how the AIA's general prohibitions (Article 5(1)(a)) interact with specific provider and deployer transparency obligations (Article 50), especially regarding embedded design features and technical measures for detecting manipulated content.[239] This has significant implications for deepfake detection technologies, driving the industry toward greater transparency and explainability, as emphasized in the CJEU's *SCHUFA*[240] and *Dun & Bradstreet*[241] rulings. Integrating XAI for transparent decision-making with labelling and watermarking, such as C2PA standards for verifiable content origin, is crucial for consistently combating harmful deepfakes, including those used for financial fraud, political misinformation, and non-consensual content.

Furthermore, such Commission guidelines strengthen the AI Act's role by prohibiting harmful AI systems (Article 5(1)(a) and (b)), like sexually explicit deepfakes, due to their significant harm potential, even in non-criminal misuse.[242] However, fragmented EU national criminal laws, including varying definitions of what constitutes a deepfake or different thresholds for prosecution, could hinder cross-border AI crime investigations, challenging effective detection and attribution.[243]

3.1.2. *Deepfake transparency measures: Necessary but insufficient*

Article 50 of the AIA establishes transparency obligations for providers and deployers, aligning with CJEU transparency principles.[244] However, these obligations are insufficient to mitigate all data protection risks, conflicting with the EU's strong emphasis on individual GDPR privacy rights.

Article 50(1) requires providers of AI systems (e.g., Deepware) to notify users of interactions, unless obvious, such as when a user actively uploads content for analysis. However, this does not override GDPR obligations. Deepware's privacy policy potentially infringes several key GDPR provisions, including data minimization (Article 5(1)(c)), purpose limitation (Article 5(1)(b)), security (Article 32), data transfer restrictions (Chapter V), and data subject rights (Articles 15–22).[245]

Article 50(2) mandates provider machine-readable markings on artificially generated/manipulated content (e.g., Synthesia videos used in political campaigns)[246] disclosing its origin. While this applies to the output, input personal data remains subject to GDPR. Facial reconstructions (e.g., Reality Defender outputs used in fraud investigations)[247] also require marking. Generating synthetic data, even for model improvement, triggers GDPR obligations (Article 6), including data minimization/purpose limitation (Article 5), transparency, accuracy, security (Article 32), and data subject rights (Articles 15–22). Excluding exemptions (e.g., minor edits, law enforcement), processing synthetic data requires thorough GDPR analysis, including a lawful basis and appropriate safeguards. Failure to comply with this provider marking obligation risks violating both AIA Article 50(2) and potentially 5(1)(a) as the Commission guidelines suggest non-labelling could be considered a deceptive practice.[248]

Article 50(3) AIA mandates transparency for deployers of emotion recognition and biometric categorization systems, requiring individual notification of exposure (except for authorized law enforcement). This provision explicitly requires compliance with both the GDPR and, where applicable, the Law Enforcement Directive. This reflects CJEU jurisprudence on biometric data processing (e.g., *Ministerstvo*).[249] Deploying these systems necessitates robust data protection, impacting entities like banks (Article 16 AIA). Despite GDPR compliance claims when using biometric-based tools like Sentinel,[250] HyperVerge,[251] and Oz

[235] Commission, 'Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 884 final (4 February 2025) para 73, 90.
[236] Ibid., para 72 n 63.
[237] Case C-646/22 *Compass Banca SpA v Autorità Garante della Concorrenza e del Mercato (AGCM)* [2024] EU:C:2024:957 [43], [53], [57], [59].
[238] Commission, 'Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 884 final (4 February 2025) para 72 n 63.
[239] Ibid., para 71.
[240] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [58], [59], [60], [61].
[241] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [49], [58], [60], [69], [70], [72], [74].
[242] Commission, 'Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 884 final (4 February 2025) para 145.
[243] European Parliament, Criminal procedural laws across the European Union – a comparative analysis of selected main differences and the impact they have over the development of EU legislation. https://www.europarl.europa.eu/RegData/etudes/STUD/2018/604977/IPOL_STU(2018)604977_EN.pdf, 2018 (accessed 6 June 2025).
[244] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957; Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117; Case C-434/16 *Peter Nowak v Data Protection Commissioner* [2017] ECLI:EU:C:2017:994.
[245] Deepware, Privacy Policy. https://deepware.ai/privacy-policy/ (accessed 6 June 2025). Deepware's privacy policy indicates potential GDPR breaches, including: data minimisation (GDPR, art 5(1)(c)) through excessive data collection (e.g. marital status, social security numbers) without demonstrated necessity; purpose limitation (GDPR, art 5(1)(b)) with vague data use specifications; security of processing (GDPR, art 32) via weak assurances; rules on international transfers (GDPR, ch V) concerning data transfers to Turkey without specified adequate safeguards; and unclear mechanisms for data subject rights (e.g. objection (GDPR, art 21), restriction (GDPR, art 18)). The policy poses significant privacy risks and needs overhaul.
[246] Synthesia. https://www.synthesia.io/, 2017 (accessed 6 June 2025); The Guardian, 'It's not me, it's just my face': the models who found their likenesses had been used in AI propaganda. https://www.theguardian.com/technology/2024/oct/16/its-not-me-its-just-my-face-the-models-who-found-their-likenesses-had-been-used-in-ai-propaganda, 2024 (accessed 6 June 2025).
[247] Reality Defender, Visual deepfake detection explainability. https://www.realitydefender.com/blog/visual-deepfake-detection-explainability, 2024 (accessed 6 June 2025).
[248] Commission, 'Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 884 final (4 February 2025) para 56.
[249] Case C-205/21 *Ministerstvo na vatreshnite raboti (Enregistrement de données biométriques et génétiques par la police)* [2023] ECLI:EU:C:2023:49.
[250] Sentinel AI, Privacy Policy. https://thesentinel.ai/privacy-policy.html, 2024 (accessed 6 June 2025). Sentinel displays a lack of transparency (GDPR, arts 12-14) concerning: specific data categories collected; data retention periods (see art 5(1)(e)); details on security measures (see art 32); third-party sharing practices (see art 28); and clear procedures for exercising user rights (arts 15-22).
[251] HyperVerge, Privacy Policy. https://cdn.hyperverge.co/wp-content/uploads/2025/01/Privacy-Notice.pdf, 2025 (accessed 6 June 2025). HyperVerge, while showing improvements, still needs to address several GDPR requirements, including: justification of legitimate interests (GDPR, art 6); transparency (arts 12-14) regarding data categories and third-party processing (see also art 28); data minimization (art 5(1)(c)); data transfer safeguards (ch V); data retention periods (art 5(1)(e)); procedures for exercising user rights (arts 15-22); and specification of security measures (art 32).

Forensics[252] many deployers may not be fully compliant, exhibiting common gaps: missing Data Protection Impact Assessments (DPIAs) (Article 35); invalid lawful bases for processing biometric data (Articles 6 and 9); insufficient security measures (Article 32); unclear data retention policies (Article 5(1)(e)); and inadequate information provided regarding third-party processing (Article 28). Privacy-Enhancing Technologies (PETs), such as Keyless' Zero-Knowledge Biometrics™, offer a more privacy-preserving alternative.[253]

Finally, Article 50(4) requires disclosure by deployers of AI systems creating artificially generated or manipulated content, indirectly encouraging the use of deepfake detection technologies for verification (similar to the C2PA initiative).[254]

Despite acknowledging specific exemptions (e.g., law enforcement—Article 50(1)–(4)) and legitimate applications (e.g., satire, art—Recital 134), the AIA's existing framework is insufficient to effectively counter the potential for deepfake abuse. Critical next steps include: precise definitions and a clear articulation of its relationship with the DSA; unequivocal classification of fraudulent deepfakes (encompassing extortion) as "unacceptable-risk"; the establishment of robust enforcement mechanisms; and the bolstering of data protection through DPIAs and PETs. The internal tensions within the AIA, coupled with the significant variations in approach across the EU, US, UK, and China, starkly illustrate the complex and evolving landscape of deepfake regulation.

3.2. The EU General Data Protection Regulation and deepfake detection: A balancing act

3.2.1. Biometric data processing

Developing robust deepfake detection mechanisms to counter threats like financial fraud, political manipulation, and non-consensual content presents a complex landscape within the GDPR. While the GDPR aims to safeguard fundamental rights related to personal data protection, it simultaneously enables and constrains these detection efforts, demanding a delicate balance.

A core tension arises from the GDPR's definition of biometric data (Article 4(14)),[255] encompassing facial images/expressions and voice patterns—key characteristics deeply involved in both deepfake creation and detection. Consequently, deepfake detection activities fall squarely within the GDPR's ambit, demanding careful consideration of lawful bases and strict safeguards.[256] Anti-fraud tools like iProov (performing advanced liveness detection),[257] Pindrop (collecting voice biometrics),[258] and BioCatch (analyzing behavioural biometrics including keystroke dynamics)[259] all process data defined by GDPR.

Lawful processing is governed by Articles 6 and 9. Obtaining explicit consent (Articles 6(1)(a) and 9(2)(a)), particularly for training datasets used to identify deepfakes across various applications, presents a significant hurdle.[260] The practicalities of gaining freely given, specific, informed, and unambiguous consent for detecting circulating political deepfakes or non-consensual content are insurmountable, highlighting the tension between comprehensive training data needs and individual rights.[261]

Alternatives like processing for legal obligations (Article 6(1)(c)), such as the DSA-mandated removal of illegal deepfakes used in financial scams or for political interference, and processing necessary for reasons of substantial public interest (Article 9(2)(g)), such as combating disinformation affecting democratic processes, offer pathways necessitating a clear legal basis in EU or Member State law and adherence to necessity and proportionality.

The legitimate interests basis (Article 6(1)(f)) permits processing unless overridden by data subject rights, requiring a Legitimate Interests Assessment (LIA) carefully weighing detection interests (e.g., preventing fraudulent deepfakes) against individual rights (e.g., privacy or freedom of expression).[262] However, CJEU jurisprudence emphasizes strict

[252] Oz Forensics, Privacy Policy Oz Liveness Demo Application. https://ozforensics.com/legal/privacy_policy_liveness_demo_application, 2025 (accessed 6 June 2025). Oz Forensics' practices indicate numerous GDPR concerns, including: inadequate transparency (GDPR, arts 12-14) concerning the distinction between on-device and server-based processing, details of data processors (see also art 28), and policy change notifications; data minimization (art 5(1)(c)) by requiring personal data for demos; issues with lawfulness and fairness (art 5(1)(a)) by attempting to place responsibility for third-party data consent on users; lack of a clear legal basis (art 6) for collecting telemetry data and other processing activities; unaddressed data transfers to Singapore (ch V); reliance on vague consent (arts 4(11), 7); ill-defined or insufficient data retention periods (art 5(1)(e)); and overly general information on exercising user rights (arts 15-22). Additionally, its "do not track" statement appears ill-suited to a mobile app context.

[253] Keyless, Zero-Knowledge Biometrics™ the future of authentication. https://26689385.fs1.hubspotusercontent-eu1.net/hubfs/26689385/%5B2023%5D%20Downloadable%20Content/Keyless_Zero_Knowledge_Biometrics.pdf, 2023 (accessed 6 June 2025).

[254] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending certain regulations ('the AI Act') [2024] OJ L 178/1, rec 133 also emphasizes tagging and identification tools (e.g. watermarks, metadata, fingerprints) to trace content origin and prove authenticity, implicitly referencing C2PA.

[255] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1 art 4(14) defines biometric data as 'personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data'.

[256] Information Commissioner's Office, Biometric data guidance: biometric recognition. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/biometric-data-guidance-biometric-recognition/, 2024 (accessed 6 June 2025).

[257] iProov, Protect your business from deepfakes with dynamic liveness detection. https://www.iproov.com/deepfake-protection-liveness, 2025 (accessed 6 June 2025).

[258] Pindrop, Testing voice biometric security against AI deepfakes. https://www.pindrop.com/article/testing-voice-biometric-security-against-ai-deepfakes/, 2024 (accessed 6 June 2025).

[259] BioCatch, Why BioCatch. https://www.biocatch.com/why-biocatch?hsCtaTracking=a73571b6-f408-4a5f-981f-c9976932010a%7C22182ece-a206-47b4-bfd0-513b8cc46cea, 2024 (accessed 6 June 2025).

[260] F. Romero-Moreno, Generative AI and deepfakes: a human rights approach to tackling harmful content, International Review of Law, Computers & Technology 38(3) (2024) 297-326. https://doi.org/10.1080/13600869.2024.2324540.

[261] For further guidance on obtaining informed, freely given, specific, and unambiguous consent, see Case C-673/17 *Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] EU:C:2019:246 [72]; Case C-61/19 *Orange Romania SA v Autoritatea Națională de Supraveghere a Prelucrării Datelor cu Caracter Personal (ANSPDCP)* [2020] ECLI:EU:C:2020:901 [36].

[262] Information Commissioner's Office, How do we apply legitimate interests in practice? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/, 2024 (accessed 6 June 2025).

necessity and proportionality, demanding proof of essential processing and unavailability of less intrusive means,[263] potentially hindering broad deployments of deepfake detection technologies. Article 9(2) derogations, such as those for legal claims, may also apply. In contrast, the different regulatory approaches in jurisdictions like the US and China create alternative pathways, albeit with their own distinct challenges, for developing such detection tools.

The EU AIA's allowance for sensitive data processing (biometrics, ethnic origin) to combat AI discrimination (Article 10(5)) clashes with GDPR Article 9's stricter limitations, creating legal uncertainty and jeopardizing data subject rights, notably the right against automated decisions and profiling (Article 22 GDPR).[264] This tension, exemplified by CJEU rulings in *SCHUFA*[265] and *Dun & Bradstreet*,[266] is particularly concerning in high-risk law enforcement or finance systems. For instance, bias in deepfake detection can lead to inaccurate, legally significant automated decisions (e.g., false accusations), demanding transparency.[267] To reconcile data protection with bias correction, essential tools like XAI and C2PA are needed to overcome "black box" opacity.[268]

3.2.2. *Data protection: Design, security, and subject rights*

The GDPR's data protection by design and by default principle (Article 25) mandates upfront integration of safeguards in deepfake detection systems, ensuring minimal biometric data processing.[269] Given the sensitivity of data used in deepfake detection methods, this principle is critical. Consequently, a Data Protection Impact Assessment (DPIA) (Article 35) is almost invariably required due to the high risks of biometric processing for detection, including potential inaccuracies leading to false accusations.[270]

The principles of necessity and proportionality, as reinforced by the CJEU in *KNLTB*[271] and *HTB Neunte Immobilien Portfolio,*[272] demand that deepfake detection methods are strictly necessary and proportionate to the risks posed by both undetected deepfakes and the detection process itself. To achieve this balance, Privacy-Enhancing Technologies (PETs) offer valuable solutions by minimizing data collection and maximizing privacy, aligning with the GDPR's principles of data minimization, purpose limitation, and data security.[273] These PETs, including federated learning (decentralized training),[274] differential privacy (adding noise to protect individual data),[275] homomorphic encryption (computations on encrypted data),[276] and secure multi-party computation (collaborative detection without data sharing)[277] provide concrete mechanisms. For instance, projects like SecDFDNet illustrate how secure protocols, such as collaborative deepfake detection and secret sharing, can enable detection without directly accessing raw facial data.[278] By minimizing data collection and limiting its exposure, this approach safeguards individual privacy, further aligning with GDPR principles of data minimization, purpose limitation, and data security.

Considering the delicate nature of biometric data processed in deepfake analysis, robust data security and clear breach notification procedures are paramount. Article 32 mandates appropriate technical and organizational security measures—encryption, access controls, and regular audits—proportional to the risk.[279] Reflecting this concern, Article 33 sets out stringent data breach notification requirements, demanding notification to the competent supervisory authority without undue delay, and where feasible, within 72 h of becoming aware of a personal data breach involving biometric data.[280] Recent events, including the GenNomis service data breach that exposed thousands of non-consensual deepfake images, exemplify this issue.[281]

GDPR Articles 15–22 grant data subjects rights including access, rectification, erasure, restriction, portability, and objection. The right to object to legitimate interest processing (Article 21) is particularly pertinent to deepfake detection. However, exercising these rights against widely disseminated deepfakes, such as politically motivated manipulations or non-consensual content, presents inherent complexities due to the clash with competing rights like freedom of expression. For instance, rectifying a deepfake involves not only correcting the

[263] Case C-621/22 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECLI:EU:C:2024:857 [42], [51], [57], [58]; Joined Cases C-17/22 and C-18/22 *HTB Neunte Immobilien Portfolio geschlossene Investment UG & Co. KG and Ökorenta Neue Energien Ökostabil IV geschlossene Investment GmbH & Co. KG v Müller Rechtsanwaltsgesellschaft mbH and Others* [2024] EU:C:2024:738 [51], [59], [73], [74], [76], [78]; Case C-205/21 *Ministerstvo na vatreshnite raboti (Enregistrement de données biométriques et génétiques par la police)* [2023] ECLI:EU:C:2023:49 [126]-[128], [133].

[264] Think Tank European Parliament, Algorithmic discrimination under the AI Act and the GDPR. https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769509, 2025 (accessed 6 June 2025).

[265] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [3], [11], [42]-[50],[57], [59], [64].

[266] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [3], [8], [38], [46], [55]-[57].

[267] European Data Protection Supervisor, Deepfake detection. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en, 2024 (accessed 6 June 2025).

[268] Ibid., Complex machine learning models in many detection algorithms often operate as "black boxes," meaning their internal decision-making processes are opaque and difficult for even experts to interpret, thus limiting transparency.

[269] Information Commissioner's Office, Biometric data guidance: biometric recognition. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/biometric-data-guidance-biometric-recognition/, 2024 (accessed 6 June 2025).

[270] Ibid.

[271] Case C-621/22 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECLI:EU:C:2024:857 [42], [51], [57], [58].

[272] Joined Cases C-17/22 and C-18/22 *HTB Neunte Immobilien Portfolio geschlossene Investment UG & Co. KG and Ökorenta Neue Energien Ökostabil IV geschlossene Investment GmbH & Co. KG v Müller Rechtsanwaltsgesellschaft mbH and Others* [2024] EU:C:2024:738 [51], [59], [73], [74], [76], [78].

[273] Information Commissioner's Office, Privacy-enhancing technologies (PETs). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/, 2023 (accessed 6 June 2025).

[274] P. Kairouz, H. B. McMahan, et al., 2021. Advances and open problems in federated learning, Foundations and Trends in Machine Learning. arXiv. arXiv:1912.04977v3. https://arxiv.org/abs/1912.04977.

[275] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Foundations and Trends in Theoretical Computer Science 9 (3-4) (2014) 211-407. https://doi.org/10.1561/0400000042.

[276] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, A survey on homomorphic encryption schemes: theory and implementation, ACM Computing Surveys (CSUR) 51(4) 79 (2018) 1-35. https://dl.acm.org/doi/10.1145/3214303.

[277] Y. Lindell, B. Pinkas, Secure multiparty computation for privacy-preserving data mining, Journal of Privacy and Confidentiality 1(1) (2009) 59-98. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/566.

[278] B. Chen, X. Liu, Z. Xia, G. Zhao, 2023. Privacy-preserving deepfake face image detection, Digital Signal Processing. 143, 104233. https://www.sciencedirect.com/science/article/pii/S1051200423003287?via%3Dihub.

[279] Information Commissioner's Office, A guide to data security. https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/security/a-guide-to-data-security-0-0.pdf, 2023 (accessed 6 June 2025); Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [59], [66].

[280] European Data Protection Board, Guidelines 9/2022 on personal data breach notification under GDPR.https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-92022-personal-data-breach-notification-under_en, 2022 (accessed 6 June 2025)

[281] VPN Mentor, Thousands of AI & deepfake images exposed on Nudify service data breach. https://www.vpnmentor.com/news/report-gennomis-breach/, 2025 (accessed 6 June 2025).

manipulated data[282] but also tackling its potentially widespread dissemination,[283] highlighting this fundamental conflict.

Companies deploying deepfake detection tools at scale likely necessitate a Data Protection Officer (Article 37) due to the sensitive biometric data processed.[284] The CJEU *Bindl*[285] case's emphasis on robust international transfer safeguards and the allowance of damage claims for unlawful transfers have significant implications for cross-border deepfake analysis. Furthermore, the CJEU's *Dun & Bradstreet*[286] ruling (GDPR Recital 63) establishes a framework for Data Protection Authorities and courts to balance the trade secret protection of deepfake detection copyrighted algorithms against data subject access rights (AI transparency), a crucial aspect for accountability.

The GDPR necessitates a continuous balance between effective deepfake detection and individual freedoms, demanding meticulous consideration of lawful bases, proportionality, transparency, and data subject rights. This dual role—facilitating detection to counter harmful manipulation while constraining it to respect fundamental rights—offers a key framework for navigating AI's complex landscape, presenting both opportunities and hurdles compared to US and Chinese regulatory approaches. Though such requirements may temper rapid deployment, they incentivize responsible innovation, driving the development of XAI, C2PA, and PETs for explainable, transparent, privacy-preserving methods.

3.3. The EU Digital Services Act: Reconciling deepfake detection, online safety and fundamental rights

3.3.1. Navigating complexities and competing interests

The EU's Digital Services Act (DSA) aims to create a safer online environment in response to the significant threat that deepfakes pose to online discourse and democratic processes.[287] By imposing regulatory obligations on online intermediary platforms, particularly Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), regarding risk assessment and mitigation (Articles 34, 35), the DSA profoundly impacts deepfake detection.[288] The DSA incentivizes this detection, yet its current form risks over-removal and infringing fundamental rights. Such risks stem from vague definitions, the pressure for expeditious action, and technological limitations in distinguishing genuine content from synthetic media (e.g., reliance on AI-powered deepfake detection algorithms or biometric liveness checks), thereby creating an inherent conflict between online safety and freedoms.

A key challenge is the DSA's approach to defining "deepfakes." Unlike the EU AIA's specific definition (Article 3(60)), the DSA opts for a broader, contextual approach to "inauthentic use" contributing to disinformation (Recital 84).[289] Similarly, Recital (55), addressing automated content moderation, indirectly incentivizes platforms to invest in detection while mandating safeguards like transparency and user notification, thereby emphasizing accountability.[290] While this avoids rapid obsolescence, it creates legal uncertainty for developers regarding the tool's scope, liability for political misinformation, and content prioritization for non-consensual deepfakes. The lack of a clear definition necessitates examining how existing legal concepts, such as "illegal content" (Article 3(h)), are applied to deepfakes used in financial fraud or malicious impersonation.[291] This definitional ambiguity contrasts with potentially more precise definitions in jurisdictions such as the US[292] and China,[293] leading to jurisdictional tensions in cross-border enforcement.

2024 incidents highlight the real-world impact of deepfakes. For instance, a Marine Le Pen montage on X sparked debate in France,[294] while Italian PM Meloni testified against her pornographic deepfake creators.[295] Additionally, Australians lost $43.4 million to celebrity deepfake scams, prompting Meta to remove 9000 fraudulent Facebook pages and 1.2 billion fake accounts globally.[296]

Moreover, further provisions integrate deepfake detection into platform moderation. Article 39 incentivizes detection in advertising via mandated transparency.[297] Articles 16, 34, and 35, concerning notice-and-action, risk assessment, and mitigation, directly require platforms to address deepfake harms.[298] Article 34′s risk assessment mandate indirectly incentivizes detection technology investment, while Article 35 directly obligates VLOPs and VLOSEs to implement robust risk

[282] C. Novelli, F. Casolari, P. Hacker, G. Spedicato, L. Floridi, 2024. Generative AI in EU law: liability, privacy, intellectual property, and cybersecurity. Computer Law & Security Review. 55, 106066. https://doi.org/10.1016/j.clsr.2024.106066.

[283] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, F. Tramèr, What does it mean for a language model to preserve privacy? FAccT 22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (2022) 2280-2292. https://doi.org/10.1145/3531146.3534642.

[284] European Data Protection Supervisor, Coordinated enforcement action designation and position of Data Protection Officers. https://www.edpb.europa.eu/our-work-tools/our-documents/other/coordinated-enforcement-action-designation-and-position-data_en. 2024 (accessed 6 June 2025).

[285] Case T-354/22 *Thomas Bindl v European Commission* [2025] ECLI:EU:T:2025:4 [189]-[200]; see also Case C-362/14 *Maximillian Schrems v Data Protection Commissioner* [2015] ECLI:EU:C:2015:650 [67]-[106]; and Case C-311/18 *Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems* [2020] ECLI:EU:C:2020:559 [90]-[202].

[286] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [69], [70], [72], [74], [75].

[287] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1.

[288] Ibid., DSA arts 34(1)-(2) (requiring assessment of systemic risks, including from "intentional manipulation" of services) and art 35(1)(k) (requiring mitigation measures for manipulated image, audio, or video content).

[289] Ibid., DSA rec 84 (highlighting risks relevant to systemic risk assessment, such as "inauthentic use" including deepfakes, "misleading or deceptive content", "algorithmic amplification", and their connection to disinformation campaigns).

[290] Ibid., DSA rec 55 (discussing automated content moderation, including for "inauthentic use" of services relevant to manipulated or deceptive content, and the importance of transparency such as providing information on automation in statements of reasons and ensuring access to redress).

[291] Ibid., DSA art 3(h) (defining "illegal content" broadly as any information not in compliance with Union law or the law of a Member State).

[292] See, e.g., the varied state-level approaches targeting specific harms: California SB 926, 2023-2024 Reg Sess, ch 289 (2024); Texas SB 751, 89th Reg Sess (2025); Florida Laws 2022, ch 2022-212; Louisiana Acts 2023, No 175; California AB 2655, 2023-2024 Reg Sess, ch 261 (2024); California SB 942, 2023-2024 Reg Sess, ch 291 (2024).

[293] Art 23 of the Provisions on the Administration of Deep Synthesis Internet Information Services (Order No 12 of the Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security, 25 November 2022). <http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm> accessed 6 June 2025.

[294] Politico, Le gouvernement embarrassé par un "deepfake" visant Marine Le Pen. https://www.politico.eu/article/france-gouvernement-deepfake-marine-le-pen/, 2024 (accessed 6 June 2025).

[295] Politico, Italy's Giorgia Meloni called to testify in deepfake porn case. https://www.politico.eu/article/italian-pm-giorgia-meloni-called-to-testify-in-deepfake-porn-case/, 2024, (accessed 6 June 2025).

[296] The Guardian, More than 9,000 scam Facebook pages deleted after Australians lose $43.4m to celebrity deepfakes. https://www.theguardian.com/technology/2024/oct/02/more-than-9000-scam-facebook-pages-deleted-after-australians-lose-millions-to-celebrity-deepfakes, 2024 (accessed 6 June 2025).

[297] Ibid., DSA art 39(1)-(2) (mandating VLOPs to provide a public repository of advertisements detailing, inter alia, ad content, sponsor, payer, presentation period, and targeting parameters as specified in art 39(2)(a)-(e)).

[298] Ibid., DSA arts 16, 34, 35 (collectively requiring VLOPs and VLOSEs) to address harms from, e.g., illegal deepfakes, through user reporting mechanisms (art 16), systemic risk assessment including for intentional manipulation (art 34 (1)-(2)), and mitigation measures including the detection of deepfakes (art 35 (1)(k))).

management systems, potentially including these technologies.[299] Moreover, Articles 44 and 45, regarding standards and codes of conduct, reinforce detection's importance and encourage algorithm development collaboration, particularly for VLOPs and VLOSEs.[300]

3.3.2. Challenges and risks: Over-removal, chilling effects, and fundamental rights

However, these incentives are balanced by significant challenges, notably its emphasis on "expeditious action" (Recital 22) and specific timeframes for illegal content, such as the 24 h benchmark for illegal hate speech (Recital 87).[301] This pressures platforms to rapidly assess and remove potentially illegal deepfakes, creating a substantial over-removal risk due to false positives,[302] especially concerning nuanced content like misclassified political satire or artistic expression[303] where accurate and rapid detection is technically challenging.[304] This emphasis on speed over accuracy creates a balancing challenge within the DSA, potentially chilling freedom of expression,[305] a concern that is also present in the US framework, which has a strong emphasis on free speech.

The CJEU Advocate General in *Poland v Council and Parliament*[306] warned against filtering with high false positive rates, a direct concern for deepfake detection. False positives, as established in CJEU case law like *Sabam v Scarlet*,[307] and *Netlog*,[308] risk erroneously removing legitimate content such as satire or artistic expression. The *Glawischnig-Piesczek*[309] ruling, allowing removal of "equivalent" defamatory information, compounds this risk by potentially capturing protected expression, particularly relevant for deepfakes where distinguishing harmful falsehoods from satire is inherently difficult.[310] The DSA's emphasis on speed, coupled with liability for inaction, further incentivizes over-removal, arguably chilling freedom of expression,[311] especially given the subjective nature of judging deepfakes, often depending on context, cultural norms, and individual interpretation.[312]

CJEU decisions, including *Sabam v Scarlet*,[313] and *Netlog*,[314] have established limitations on automated filtering, highlighting its impact on copyright exceptions and user rights under Articles 8 (data protection) and 11 (freedom of expression) of the EU Charter of Fundamental Rights. The further CJEU *SCHUFA*[315] and *Dun & Bradstreet*[316] rulings, referencing GDPR Recital 71, mandate safeguards for automated decision-making, including appropriate procedures, data security, and the right to human intervention. These safeguards are crucial for deepfake detection in content moderation due to the potential impact on fundamental rights. The DSA's "expeditious action" requirement, however, risks undermining these safeguards by prioritizing speed over meaningful human review, possibly infringing protected rights.[317] The DSA's empowerment of private companies to decide on removing deepfakes used in online harassment or disinformation raises concerns about private censorship and democratic oversight,[318] a regulatory challenge regarding accountability and due process.

Furthermore, DSA Recital 69 highlights the risk of manipulative advertising, including potentially discriminatory deepfakes. This is particularly relevant considering Meta's US testing of facial recognition to detect deepfakes featuring celebrities in advertisements.[319] Despite its purported immediate data deletion,[320] this practice raises concerns under stricter EU data protection law, a potential jurisdictional tension stemming from GDPR as interpreted by the CJEU. The *Meta v Bundeskartellamt*[321] case addressed the legality of Meta's data processing practices, finding that the bundling of data from various services and the conditions of consent did not comply with GDPR, particularly regarding the legal basis under Article 6 (specifically legitimate interests) and the requirement for freely given consent. Additionally, *Schrems v Meta*[322] established that Meta's extensive data collection for advertising purposes, including sensitive data, breached the GDPR's data minimization principle (Article 5(1)(c)). This could push EU-based platforms towards alternative, GDPR-compliant detection techniques like anonymized data analysis, XAI, and C2PA solutions.

The DSA presents a complex challenge: reconciling the need to

[299] Ibid., DSA arts 34(1)-(2), 35(1)(k) (requiring VLOPs and VLOSEs) to assess systemic risks (art 34(1)), including from 'intentional manipulation' relevant to deepfakes (art 34(2)), and to implement mitigation measures, specifically including deepfake detection (art 35(1)(k))).

[300] Ibid., DSA art 44 (encouraging standards for platform operations relevant to deepfake detection, cf. art 44(1)(e)-(i)); art 45 (fostering codes of conduct for collaborative deepfake mitigation, including objectives and key performance indicators (KPIs), cf. art 45(1)-(2)).

[301] Ibid., DSA rec 22 (emphasizing the general need for 'expeditious action' against illegal content); rec 87 (highlighting the 24 h benchmark for processing illegal hate speech removal notifications under the 2016 Code of Conduct on countering illegal hate speech online as an important example of operational measures).

[302] A. Turillazzi, M. Taddeo, L. Floridi, F. Casolari, The Digital Services Act: an analysis of its ethical, legal, and social implications, Law, Innovation and Technology 15(1) (2023) 83-106. https://doi.org/10.1080/17579961.2023.2184136.

[303] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [52]; Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [50].

[304] B. Cavia, E. Horwitz, T. Reiss, Y. Hoshen, 2024. Real-Time Deepfake Detection in the Real World. arXiv. arXiv:2406.09398. https://arxiv.org/abs/2406.09398.

[305] M. Husovec, The Digital Services Act's red line: what the Commission can and cannot do about disinformation, Journal of Media Law 16(1) (2024) 47-56. https://doi.org/10.1080/17577632.2024.2362483.

[306] AG opinion in Case C-401/19 *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [214].

[307] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [52].

[308] Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [50].

[309] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821 [38], [39], [45], [46], [53], [55].

[310] Co-creation Studio, JUST JOKING! 2023 Action plan from questions to action. https://cocreationstudio.mit.edu/just-joking-action-plan/, 2022 (accessed 6 June 2025); M. Labuz, Deep fakes and the Artificial Intelligence Act—an important signal or a missed opportunity?, Policy & Internet 16(4) (2024) 1-18. https://onlinelibrary.wiley.com/doi/10.1002/poi3.406.

[311] M. Husovec, The Digital Services Act's red line: what the Commission can and cannot do about disinformation, Journal of Media Law 16(1) (2024) 47-56. https://doi.org/10.1080/17577632.2024.2362483.

[312] UN, 'Intensification of efforts to eliminate all forms of violence against women and girls: technology-facilitated violence against women and girls: Report of the Secretary-General' (2024) UN Doc A/79/500.

[313] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [51]-[53].

[314] Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [48]-[51].

[315] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [3], [45], [53], [54], [56], [57], [59], [66].

[316] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [3], [55], [56]-[58].

[317] G. Frosio, C. Geiger, Taking fundamental rights seriously in the Digital Services Act's platform liability regime, European Law Journal 29(1-2) (2023) 31-77. https://doi.org/10.1111/eulj.12475.

[318] M. Husovec, The Digital Services Act's red line: what the Commission can and cannot do about disinformation, Journal of Media Law 16(1) (2024) 47-56. https://doi.org/10.1080/17577632.2024.2362483.

[319] Meta, Testing new ways to combat scams and help restore access to compromised accounts. https://about.fb.com/news/2024/10/testing-combat-scams-restore-compromised-accounts/, 2024 (accessed 6 June 2025).

[320] Ibid.

[321] Case C-252/21 *Meta Platforms Inc and Others v Bundeskartellamt* [2023] ECLI:EU:C:2023:537 [27], [30], [62], [103], [104], [117], [148], [150], [151].

[322] Case C-446/21 *Maximilian Schrems v Meta Platforms Ireland Limited, anciennement Facebook Ireland Limited* [2024] ECLI:EU:C:2024:834 [45]-[65].

combat harmful deepfakes with the protection of fundamental rights. Its current structure, with broad definitions and an emphasis on speed, risks over-removal and chilling effects. Future implementation must prioritize transparency, accountability, and robust safeguards for an optimal balance between online safety and essential freedoms, a goal approached differently and with varying degrees of emphasis across the EU, US, UK, and China.

3.4. Deepfake detection in the US: A regulatory roadblock

3.4.1. The uncertainty of federal inaction

Despite proposed AI regulation efforts,[323] the US lacks a federal privacy law like the EU's GDPR or the UK Data Protection Act. This creates uncertainty, especially concerning biometric data crucial for deepfake detection, compared to the EU and UK's defined legal frameworks and causes jurisdictional tension regarding the permissible use of detection data. The absence of GDPR principles (data minimization, purpose limitation, security)[324] in US law further exacerbates this jurisdictional tension, potentially hindering clearer guidelines for deepfake technology.

The lack of specific federal deepfake legislation severely cripples innovation, leaving individuals vulnerable to deepfake harms. US case law such as *Young v Neocortext*,[325] focused on traditional copyright, is ill-equipped to address the unique challenges posed by deepfakes, including their rapid spread in political smear campaigns or financial scams, the erosion of trust, and the difficulty in identifying them. This contrasts with the EU's AIA, which, despite its ambiguities, attempts a more direct regulatory approach to defining and addressing deepfakes.[326]

Federal progress has been slow due to the failure of legislative efforts like the Malicious Deep Fake Prohibition Act of 2018.[327] The 2020 National Defense Authorization Act[328] initiated some action with disinformation reports and a detection technology prize.[329] However, subsequent legislative inaction, including the failed Deepfake Report Act (2019),[330] DEEPFAKES Accountability Act (DAA) (2019),[331] and Identifying Outputs of Generative Adversarial Networks Act (IOGAN) (2020),[332] leaves critical gaps, including defining actionable deepfakes and incentivizing detection technology for identifying manipulated media in non-consensual scenarios.

The DAA's proposed mandated disclosures and provenance standards,[333] similar to the EU's push for C2PA,[334] remain unenforceable without legislative backing, highlighting a tension between recognizing the need for such measures and the federal inability to enact them. Consequently, the absence of clear guidance on addressing specific deepfake types, like political disinformation versus satire, continues to hinder progress and creates a jurisdictional divergence compared to regions with more specific regulations such as the EU, UK or China.

Furthermore, the 2021 NO FAKES Act,[335] while addressing unauthorized digital replicas in intellectual property (IP) relevant to financial fraud involving celebrity endorsements, inadvertently disincentivizes detection through its "willful avoidance" provision,[336] potentially creating liability for platforms that invest in such technologies. This creates a tension between protecting copyright and fostering anti-deepfake tools. This, coupled with the recent passage of the TAKE IT DOWN Act (S. 146)[337] in 2025, defining deepfakes as "digital forgeries" and introducing a "notice-and-removal" requirement for non-consensual instances, risks hindering deepfake detection innovation. Its focus on removal within 48 h, without guidelines for detection technology, may inadvertently sideline crucial identification tools.

Finally, the lack of consistent federal funding and coordination,[338] unlike the more directed initiatives seen in the UK[339] and EU,[340] has hindered deepfake detection research, leaving the US vulnerable to sophisticated manipulations.

3.4.2. State-level fragmentation: Regulatory disparity

Federal inaction has resulted in a patchwork of state laws addressing deepfake harms, including non-consensual pornography, political disinformation (e.g., the deepfake Biden robocall),[341] and fraud. However, recent 2024 incidents, notably the Taylor Swift deepfakes with

[323] See, e.g., Exec Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed Reg 75191 (30 Oct 2023); National Institute of Standards and Technology, AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework, 2023 (accessed 6 June 2025); Federal Trade Commission, Artificial Intelligence Compliance Plan. https://www.ftc.gov/ai, (accessed 6 June 2025); The White House Office of Science and Technology Policy, *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People* (October 2022) https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[324] See Regulation (EU) 2016/679 (GDPR) [2016] OJ L 119/1, art 5(1)(b) (purpose limitation), art 5(1)(c) (data minimization), and art 5(1)(f) (security). These principles are mirrored in the UK GDPR, which is part of UK law under the Data Protection Act 2018.

[325] *Kyland Young v NeoCortext Inc*, 2:23-cv-02496 (CD Cal Apr 2023).

[326] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013 and (EU) 2018/858, and Directives (EU) 2015/2366 and (EU) 2020/1828 [2024] OJ L 2024/1689, art 3(60), art 50.

[327] S 3805, 115th Cong (2018).

[328] Pub L No 116-92, 133 Stat 1296 (2019).

[329] National Defense Authorization Act for Fiscal Year 2020, Pub L No 116-92, ss 1731-1733, 133 Stat 1296 (2019).

[330] S Rep 116-93, 116th Cong (2019).

[331] HR 3230, 116th Cong (2019).

[332] HR 6821, 116th Cong (2020).

[333] HR 3230, 116th Cong (2019), DEEPFAKES Accountability Act.

[334] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013 and (EU) 2018/858, and Directives (EU) 2015/2366 and (EU) 2020/1828 [2024] OJ L 2024/1689, rec 133.

[335] (2021) HR 3953, 117th Cong (2021).

[336] S 3875, 117th Cong (2021), NO FAKES Act.

[337] Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, Pub L No 119-12, 139 Stat 321 (2025).

[338] HR 6821, 116th Cong (2020), Identifying Outputs of Generative Adversarial Networks Act (IOGAN). This bill, which proposed funding for deepfake detection research at the National Science Foundation and the National Institute of Standards and Technology, was not implemented, hindering US research efforts and increasing vulnerability.

[339] GOV.UK Accelerated Capability Environment, Innovating to detect deepfakes and protect the public. https://www.gov.uk/government/case-studies/innovating-to-detect-deepfakes-and-protect-the-public, 2025 (accessed 6 June 2025).

[340] European Digital Media Observatory, "AI Against Disinformation" Cluster. https://edmo.eu/resources/connected-horizon-europe-projects/, started in 2021 (accessed 6 June 2025).

[341] Federal Communications Commission, FCC proposes $6 million fine for illegal robocalls that used Biden deepfake generative AI voice message. https://docs.fcc.gov/public/attachments/DOC-402762A1.pdf, 2024 (accessed 6 June 2025).

explicit content viewed by 47 million on platform X before takedown,[342] and Oprah Winfrey's misleading endorsements of controversial self-help courses,[343] highlight the inadequacy of this fragmented approach.

While some states are acting, their limited scope and varying definitions create a regulatory disparity, hindering uniform standards for deepfake detection and mitigation, unlike the EU's unified approach. For example, California's SB 926,[344] informed by *People v Sol Ecom*,[345] addresses sexually explicit deepfakes and Texas SB 751[346] targets deepfakes harming political candidates. Similarly, Florida's SB 1798[347] and Louisiana Acts 2023, No 175[348] prohibit deepfakes depicting child sexual abuse. These well-intentioned state efforts often have limited scope, causing inconsistencies and compliance burdens for national technology companies – an internal US tension. A comprehensive federal framework is crucial for uniform standards against deepfake threats.

California's AB 2655[349] (political deepfakes) and SB 942[350] (AI transparency) illustrate state-level regulation complexities. AB 2655's vaguely defined "state-of-the-art techniques" mandate (Sections 20513 (a) and 20514(a)), targeting political deepfake dissemination, raises over-censorship and developer uncertainty concerns, potentially chilling detection innovation. The extensive data analysis implied by user reporting (20515(a)) creates significant data protection risks. These risks—profiling,[351] re-identification,[352] biased reporting,[353] mass surveillance,[354] data breaches,[355] and mission creep[356] (repurposing data without user control)—are heightened without GDPR-level protection, creating a challenge in balancing effective detection and user privacy. The bill's rapid response requirements (Sections 20513(b), 20514(b), and 20515(a)) also pose practical challenges like malicious actors flooding the reporting system.[357]

SB 942, operative in 2026, mandates AI-generated content transparency via free detection tools revealing provenance data but faces enforcement, technical feasibility, and Application Programming Interface (API) misuse hurdles. Metadata is easily removed,[358] watermarking degrades quality,[359] and robust techniques like cryptographic signatures and blockchain have scalability, interoperability, and cost challenges.[360] Specifically, Section 22757.3(b)'s requirement for embedded provenance data and a detection API (as supported by Section 22757.2) necessitates data collection potentially linked to sensitive user information, raising concerns about tracking, deanonymization, and vulnerable centralized data repositories.[361] While Section 22757.3(b) attempts to limit data collection, its vague language offers weak safeguards.

Furthermore, legal challenges like *Kohls v Bonta*[362] against California's AB 2655[363] and AB 2839,[364] highlight the difficulty of regulating political deepfakes and protecting free speech, especially satire and parody. While *Hustler v Falwell*[365] protected outrageous speech, deepfakes' deceptive nature questioning "actual malice" creates new legal issues,[366] with potential pre-emption and Commerce Clause challenges.[367] Additionally, data privacy issues, exemplified by *Clarkson v OpenAI*[368] revealing companies' ability to amass detailed user profiles (based on contact details, IP addresses, and browsing history) and train AI on sensitive information without sufficient transparency or consent, further hinder detection relying on training data understanding. This underscores the necessity of stronger federal privacy protections.

The current US regulatory landscape—characterized by federal inaction and fragmented state laws—creates a chilling effect on deepfake detection innovation while struggling to adequately protect individual rights. A strong federal framework, encompassing a robust privacy law and XAI/C2PA standards, is essential. This would establish clear guidelines, foster responsible innovation, and effectively protect against harmful deepfakes, aligning the US with proactive regulations in

[342] J. Sturges, 2024. Taylor Swift, deepfakes, and the First Amendment: changing the legal landscape for victims of non-consensual artificial pornography, Georgetown Journal of Gender and the Law 25(2) (2024) 1-11. https://www.law.georgetown.edu/gender-journal/online/volume-xxv-online/taylor-swift-deepfakes-and-the-first-amendment-changing-the-legal-landscape-for-victims-of-non-consensual-artificial-pornography/.

[343] BBC News, Piers Morgan and Oprah Winfrey 'deepfaked' for US influencer's ads. https://www.bbc.co.uk/news/technology-67703018, 2024 (accessed 6 June 2025).

[344] California Senate Bill 926, 2023-2024 Reg Sess, ch 289 (2024).

[345] *People of the State of California (David Chiu) v Sol Ecom Inc et al*, CGC-24-617237 (Cal Super Ct SF 2024); San Francisco City Attorney's Office, City Attorney sues most-visited websites that create nonconsensual deepfake pornography. https://www.sfcityattorney.org/2024/08/15/city-attorney-sues-most-visited-websites-that-create-nonconsensual-deepfake-pornography/, 2024 (accessed 6 June 2025); National Association of Women Judges, AI and the Courts: digital evidence and deepfakes in the age of AI. https://www.nawj.org/uploads/files/annual_conference/2024-annual-conference/fri845a-sgipsonrankinnawjhandout.pdf, 2024 (accessed 6 June 2025).

[346] Texas Senate Bill 751, 89th Reg Sess (2025).

[347] Florida Laws 2022, ch 2022-212.

[348] Louisiana Acts 2023, No 175.

[349] California Assembly Bill 2655, 2023-2024 Reg Sess, ch 261 (2024).

[350] California Senate Bill 942, 2023-2024 Reg Sess, ch 291 (2024).

[351] Gil, C., Parra-Arnau, J., Forné, J., 2025. Privacy protection against user profiling through optimal data generalization. Computers & Security. 148, 104178. https://doi.org/10.1016/j.cose.2024.104178.

[352] A. Farzanehfar, F. Houssiau, Y-A. de Montjoye, 2021. The risk of re-identification remains high even in country-scale location datasets. Patterns. 2(3), 100204. https://doi.org/10.1016/j.patter.2021.100204.

[353] J. T. Van der Steen, G. ter Riet, C. A. van den Bogert, L. M. Bouter, 2019. Causes of reporting bias: a theoretical framework. F1000Research. 8:280. https://doi.org/10.12688/f1000research.18310.2.

[354] The Royal Academy of Engineering, Dilemmas of privacy and surveillance challenges of technological change. https://raeng.org.uk/media/2hwab54p/dilemmas_of_privacy_and_surveillance_report.pdf, 2007 (accessed 6 June 2025).

[355] R. Ong, Mandatory data breach notification: its role in protecting personal data, Journal of International and Comparative Law 10(1) (2023) 87-112. https://www.jicl.org.uk/storage/journals/June2023/J1fjQwq0kHkl3Q6Yb2Fy.pdf.

[356] B-J. Koops, The concept of function creep, Law, Innovation and Technology 13(1) (2021) 29-56. https://doi.org/10.1080/17579961.2021.1898299.

[357] S. Stockwell, M. Hughes, P. Swatton, A. Zhang, J. Hall KC, Kieran, AI-enabled influence operations: safeguarding future elections, The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-safeguarding-future-elections, 2024 (accessed 6 June 2025).

[358] Reuters Institute & University of Oxford, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 (accessed 6 June 2025).

[359] Brookings, Detecting AI fingerprints: a guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/, 2024 (accessed 6 June 2025).

[360] G. Tripathi, M. A. Ahad, G. Casalino, 2023. A comprehensive review of blockchain technology: underlying principles and historical background with future challenges. Decision Analytics Journal. 9, 100344. https://doi.org/10.1016/j.dajour.2023.100344.

[361] A. Vilesov, Y. Tian, N. Sehatbakhsh, A. Kadambi, Solutions to deepfakes: can camera hardware, cryptography, and deep learning verify real images?. arXiv. arXiv:2407.04169. https://arxiv.org/abs/2407.04169.

[362] *Kohls v Bonta*, 2:24-cv-02527 (ED Cal 2024).

[363] California Assembly Bill 2655, 2023-2024 Reg Sess, ch 261 (2024).

[364] California Assembly Bill 2839, 2023-2024 Reg Sess, ch 262 (2024).

[365] *Hustler Magazine v Falwell*, 485 US 46 (1988) at 51, 52, 56.

[366] J. Lindsey, The place for illusions: deepfake technology and the challenges of regulating unreality, University of Florida Journal of Law & Public Policy 33 (2) Article 7 (2023) 309-332. https://scholarship.law.ufl.edu/jlpp/vol33/iss2/7/.

[367] Brookings, Constitutional constraints on regulating Artificial Intelligence. https://www.brookings.edu/articles/constitutional-constraints-on-regulating-artificial-intelligence/, 2024 (accessed 6 June 2025).

[368] *PM et al v OpenAI LP*, 3:23-cv-03199 (ND Cal 2023) [151], [152], [163], [248], [249], [269], [298].

the EU, UK, and China.

3.5. The UK Online Safety Act: Deepfake detection's indirect impact and key challenges

3.5.1. The indirect influence of the OSA

The UK Online Safety Act 2023 (OSA), while not explicitly addressing "deepfakes," significantly impacts the balance between fostering deepfake detection innovation and protecting individual rights. This legislation obligates a wide range of online platforms with UK links.[369] The OSA applies to platforms with a significant UK user base, those targeting the UK market, or those posing a risk of significant harm to UK users, regardless of location.[370]

The OSA (Section 1(3)) imposes a duty of care on providers to protect users, especially children, from harmful online content, including that facilitated by deepfakes. This duty covers "illegal content" (Section 1(3)(a)), such as deepfake-generated child sexual abuse material (CSAM), and "harmful content" (Section 1(3)(b)), defined by Schedule 1 as content likely to cause material physical or psychological harm to children (paragraph 2(a)) or a serious adverse effect on mental health (paragraph 2(b)). For example, platforms would need to detect and remove deepfake endorsements by public figures of dangerous online challenges to avoid being considered harmful under paragraph 2.[371]

Beyond the OSA's general duty, deepfakes facilitate non-consensual pornography, as evidenced by a 2021 UK conviction.[372] The Sexual Offences Act 2003, as amended by the OSA 2023 (Section 66B), criminalizes sharing non-consensual intimate deepfakes.[373] The UK government further plans to criminalize the creation of such content where there is intent to cause distress or sexual gratification, and no reasonable belief of consent.[374]

The OSA's requirement for platforms to assess and mitigate harmful content risks (Sections 9, 28–33) directly incentivizes deepfake detection tool development and deployment. Deepfakes, such as in Martin Lewis scam ads, fuel financial fraud.[375] In UK online apps, 75 % of users encounter deepfakes, with 19 % having been victimized and 22 % knowing someone affected. UK romance scams, often using deepfakes, have cost victims £410 million since 2020.[376] This indirect incentivization contrasts with the more direct mandates of the EU's DSA.

However, the OSA's fragmented approach risks hindering deepfake detection innovation and unbalancing technological advancement with individual rights. While mandating platform action against deepfakes (enforced by the Competition and Markets Authority - CMA),[377] it lacks defined detection standards, relying on non-binding Ofcom guidance.[378] This expands its remit to model developers and hosting services, creating legal uncertainty.[379] Conversely, the Digital Regulation Cooperation Forum's (DRCF), aligned with Information Commissioner's Office (ICO) practices, advocates data sharing for better deepfake detection and scam prevention, a strategy impeded by the OSA's preference for policy over legislation. This approach aims to improve deepfake detection and support the Financial Conduct Authority (FCA) in combating scams.[380] This inconsistency—DRCF promoting data sharing, OSA lacking mandates—obstructs collaboration. Furthermore, the OSA's silence on the UK Data Protection Act 2018 (DPA) further conflicts with ICO/DRCF data sharing guidance for model training,[381] a gap only partly covered by potential Ofcom Codes.

3.5.2. Regulatory and technical challenges in OSA deepfake mitigation

The OSA's broad definitions of "illegal" (Section 1(3)(a)) and "harmful" content (Section 1(3)(b)), coupled with the lack of a specific legal definition for "deepfake," generate a significant regulatory challenge that contributes to numerous technical problems. This ambiguity obstructs focused deepfake detection tool development and testing, creating uncertainty regarding which deepfakes are detectable and technical standards for applications ranging from political misinformation to non-consensual image manipulation. Ofcom's non-binding guidance, issued following the 2024 *Telegraph* investigation into AI chatbots mimicking deceased British teenagers Brianna Ghey and Molly Russell, highlights these definitional problems.[382] Ofcom's clarification

[369] GOV.UK Guidance Online Safety Act: explainer. https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer, 2024 (accessed 6 June 2025); specifically, this legislation covers search, content posting, and user interaction platforms with UK links, including social networks, cloud storage, video-sharing platforms, forums, dating apps, and messaging services.

[370] Ibid.

[371] For example, to avoid being considered harmful under paragraph 2(a) (if targeted at children) or 2(b) (if causing widespread anxiety), platforms, under their duty of care, would need to detect and remove deepfake endorsements. This could include instances where a public figure appears to endorse a dangerous online challenge or promotes misleading health advice.

[372] BBC News, 'I was deepfaked by my best fried'. https://www.bbc.co.uk/news/uk-68673390, 2024 (accessed 6 June 2025).

[373] Sexual Offences Act 2003, s 66B (inserted by Online Safety Act 2023, s 188).

[374] GOV.UK, Press release, better protection for victims thanks to new law on sexually explicit deepfakes.https://www.gov.uk/government/news/better-protection-for-victims-thanks-to-new-law-on-sexually-explicit-deepfakes#:~:text=The%20Government%20has%20tabled%20an,without%20reasonable%20belief%20in%20consent, 2025 (accessed 6 June 2025).

[375] MoneySavingExpert, Have the 'Martin Lewis' scammers finally been uncovered?. https://www.moneysavingexpert.com/news/2023/april/martin-lewis-bbc-scam-warning, 2023 (accessed 6 June 2025); BBC News 'I was scammed out of £75k by Martin Lewis deepfake advert'. https://www.bbc.co.uk/news/articles/clyvj754d9lo, 2024 (accessed 6 June 2025); MoneySavingExpert, Martin Lewis scam adverts, he doesn't do ads – so any you see are fake. https://www.moneysavingexpert.com/shopping/fake-martin-lewis-ads/, 2025 (accessed 6 June 2025).

[376] Sumsub, One in five single Brits have already been duped by deepfakes on dating apps. https://sumsub.com/newsroom/one-in-five-single-brits-have-already-been-duped-by-deepfakes-on-dating-apps/, 2025 (accessed 6 June 2025).

[377] Digital Regulation Cooperation Forum, The future of synthetic media. https://www.drcf.org.uk/siteassets/drcf/pdf-files/the-future-of-synthetic-media.pdf?v=385978, 2024 (accessed 6 June 2025).

[378] Ofcom, Deepfake defenses mitigating the harms of deceptive deepfakes. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754, 2024 (accessed 6 June 2025).

[379] Ibid.

[380] Digital Regulation Cooperation Forum, The future of synthetic media. https://www.drcf.org.uk/siteassets/drcf/pdf-files/the-future-of-synthetic-media.pdf?v=385978, 2024 (accessed 6 June 2025).

[381] Ibid; see also ICO, Data sharing: a code of practice. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/data-sharing-a-code-of-practice/, 2021 (accessed 6 June 2025); ICO, Guidance on AI and data protection. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/, 2023 (accessed 6 June 2025); ICO, How to use AI and personal data appropriately and lawfully. https://ico.org.uk/media/for-organisations/documents/4022261/how-to-use-ai-and-personal-data.pdf, 2022 (accessed 6 June 2025); ICO, Privacy-enhancing technologies (PETs). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/, 2023 (accessed 6 June 2025).

[382] Telegraph, Digital clones of Brianna Ghey and Molly Russell created by 'manipulative and dangerous' AI. https://www.telegraph.co.uk/business/2024/10/30/digital-clones-brianna-ghey-molly-russell-created-ai/, 2024 (accessed 6 June 2025).

that the OSA takes a broad approach treating all AI-generated content as user-generated[383] further complicates the development of clear technical standards for deepfake detection.

These regulatory gaps and ambiguities significantly challenge the balance between data protection, legitimate uses (art, satire, parody), freedom of expression, and business operations – an internal tension within the Act. Critics argue the OSA's focus on reactive takedowns, rather than preventative "staydown" measures (like preventing non-consensual deepfakes upon upload), severely limits its effectiveness.[384] "Staydown" deepfake prevention is more effective than post-dissemination "takedown"; however, it arguably raises human rights concerns.[385] While the OSA does not explicitly mandate proactive monitoring, its framework necessitates it for many platforms, especially Category 1 services handling illegal content. Sections 9–11 impose duties to prevent users from encountering illegal content (Schedule 4 "priority offences" like terrorism and child sexual abuse). The requirement for "proportionate steps" to prevent exposure, minimize content duration (10(4)(c)(i), 11(2)(c)(i)), and ensure swift removal (10(4)(c)(ii), 11(2)(c)(ii)) implies proactive measures.

A key technical difficulty arises from using hashing for content identification.[386] While effective for detecting exact copies of known deepfakes,[387] it presents significant privacy risks, including potential user re-identification,[388] increased data breach risks,[389] centralized surveillance,[390] and function creep[391] threatening free expression.[392] Furthermore, false positives (due to hash collisions)[393] and over-blocking (as evidenced in UK Supreme Court *Cartier v BT*)[394] can wrongly flag content and stifle artistic expression, particularly satirical deepfakes, creating a conflict with freedom of expression. The risk of over-blocking, exacerbated by the OSA's broad powers and ambiguous guidance regarding context and intent (as in *Chambers v DPP*),[395] threatens to stifle the development of these tools, ironically undermining the Act's intent.

Critically, unlike the EU AIA or China's Deep Synthesis Provisions, the OSA fails to address crucial technical aspects of deepfake detection, such as AI training data, model accuracy, and bias mitigation. This lack of technical guidance, particularly regarding the composition and potential biases within training data, undermines the effectiveness of detection tools and creates ambiguity. Studies have demonstrated algorithmic bias resulting in higher error rates for certain demographic groups, including individuals with darker skin tones and often favoring male detection.[396] As exemplified by the UK's *R (Bridges)*[397] ruling, deploying biased technology without regulatory oversight poses significant dangers. This bias undermines the OSA's duty to safeguard individuals.

Despite the OSA's aim to regulate online harms, its indirect nature necessitates explicit deepfake detection provisions in Ofcom's Codes. These Codes must address definitions, data protection (per UK DPA 2018), free expression, business freedom, and robust technical standards (bias mitigation, XAI, C2PA). This is essential to effectively mitigate deepfake harms and protect UK users, potentially drawing lessons from the EU and China's direct regulatory and technical standard engagements.

[383] Ofcom, Open letter to UK online service providers regarding Generative AI and chatbots. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/open-letter-to-uk-online-service-providers-regarding-generative-ai-and-chatbots/, 2024 (accessed 6 June 2025).

[384] C McGlynn and L Woods, Written evidence submitted to Ofcom Consultation on 'protecting people from illegal harms' (2024) [IIA003] pp 1-6 https://committees.parliament.uk/writtenevidence/130477/pdf/, 2024 (accessed 6 June 2025); L. Woods, W. Perrin, Online harm reduction – a statutory duty of care and regulator, Collective Wellbeing Carnegie UK (2019) 1-71. https://carnegieuk.org/publication/online-harm-reduction-a-statutory-duty-of-care-and-regulator/.

[385] J. M. Urban, J. Karaganis, B. Schofield, 2017. Notice and takedown in everyday practice. UC Berkeley Public Law Research. 2755628. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628; F. Romero-Moreno, 'Notice and staydown' and social media: amending Article 13 of the Proposed Directive on Copyright, International Review of Law, Computers & Technology 33(2) (2019) 187-210. https://www.tandfonline.com/doi/full/10.1080/13600869.2018.1475906; F. Romero-Moreno, 'Upload filters' and human rights: implementing Article 17 of the Directive on Copyright in the Digital Single Market, International Review of Law, Computers & Technology 34(2) (2020) 153-182. https://www.tandfonline.com/doi/full/10.1080/13600869.2020.1733760.

[386] Ofcom, Deepfake defenses mitigating the harms of deceptive deepfakes. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754, 2024 (accessed 6 June 2025); hashing generates unique digital 'fingerprints' (hashes) for online content, stored in shared databases. Platforms leverage these to rapidly identify and share information on harmful material, including child sexual abuse material (CSAM), terrorist propaganda, non-consensual images, and potentially known deepfakes.

[387] Software Engineering Institute, Comparing the performance of hashing techniques for similar function detection. https://insights.sei.cmu.edu/blog/comparing-the-performance-of-hashing-techniques-for-similar-function-detection/, 2024 (accessed 6 June 2025).

[388] A. Sadeghi-Nasab, V. Rafe, A comprehensive review of the security flaws of hashing algorithms, Journal of Computer Virology and Hacking Techniques 19 (2023) 287-302. https://link.springer.com/article/10.1007/s11416-022-00447-w.

[389] M. A. Almaiah, L. M. Saqr, L. A. Al-Rawwash, L. A. Altellawi, R. Al-Ali, O. Almomani, Classification of cybersecurity threats, vulnerabilities and countermeasures in database systems, Computers, Materials and Continua 81(2) (2024) 3189-3220. https://www.sciencedirect.com/org/science/article/pii/S1546221824008154.

[390] Ofcom, Overview of perceptual hashing technology. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/other/perceptual-hashing-technology.pdf?v=328806, 2022 (accessed 6 June 2025).

[391] D. Leblanc-Albarel, B. Preneel, 2024. Black-box collision attacks on widely deployed perceptual hash functions, Cryptology ePrint Archive. 1869. https://eprint.iacr.org/2024/1869.pdf.

[392] B. Heller, Combating terrorist-related content through AI and information sharing, Transatlantic Working Group, The Carr Center for Human Rights Policy (2019) 1-8. https://www.ivir.nl/publicaties/download/Hash_sharing_Heller_April_2019.pdf.

[393] Ofcom, Deepfake defenses mitigating the harms of deceptive deepfakes. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754, 2024 (accessed 6 June 2025).

[394] *Cartier International AG v British Telecommunications Plc* [2018] UKSC 28 [5].

[395] *Chambers v DPP* [2012] EWHC 2157 (Admin) [28]-[32], [38].

[396] J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, Proceedings of Machine Learning Research 81 (2018) 1-15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf; S. Perkowitz, The bias in the machine: facial recognition technology and racial disparities, MIT Schwarzman College of Computing (2021). https://mit-serc.pubpub.org/pub/bias-in-machine/release/1; P. Grother, M. Ngan, K. Hanaoka, 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR. 8280. https://doi.org/10.6028/NIST.IR.8280.

[397] *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police* [2020] EWCA Civ 1058 [188], [192], [199].

3.6. China's Administrative Provisions on Deep Synthesis: Balancing innovation and control in deepfake detection

3.6.1. Broad definitions, compliance burdens, and the need for standardization

China's 2023 Provisions on the Administration of Deep Synthesis of Internet-Based Information Services (the Provisions) mark a significant step in regulating deepfakes.[398] By assigning responsibilities to deep synthesis service (DSS) providers, technical supporters, and users,[399] these Provisions create tension between centralized information control and deepfake detection innovation. This friction highlights the necessity of a balanced approach including robust countermeasures against deepfakes used in financial fraud, political misinformation, and non-consensual content, while safeguarding freedoms and promoting innovation. This balance is crucial, despite the Provisions' aim to mitigate harm through their broad scope and control.

A key aspect of China's regulatory approach is the strong emphasis on aligning AI technologies and their regulation with the Chinese Communist Party's (CCP) ideology and maintaining social and political stability.[400] This represents a significant divergence from the more decentralized media environments and regulatory philosophies of the US and EU.[401]

Article 23 of the Provisions broadly defines "deep synthesis technology" as generative or synthetic algorithms producing text, images, audio, video, and virtual scenes,[402] potentially extending beyond typical deepfakes. This broad definition, combined with Article 4′s mandate for "correct political orientation"[403] – aligning content with the CCP ideology – creates a significant tension that challenges the free and objective operation of deepfake detection technologies. Developers might prioritize censorship over addressing manipulated media,[404] contrasting sharply with the emphasis on freedom of expression in the US and EU. This chilling effect[405] could stifle the development and use of diverse media synthesis tools, hindering creativity and beneficial applications, including deepfake detection for satire or art.

Despite Articles 16–18 mandating labelling of synthesized content, the lack of specific technical standards (e.g., C2PA)[406] creates an inconsistency. Moreover, simple labelling may not be effective governance, especially for those with low moral sensitivity; negatively framed ethical messages about deepfake harms could improve perceived deception and reduce harmful engagement.[407] Thus, even with mandated labelling for AI-generated political propaganda or non-consensual deepfakes, the absence of standardized protocols hinders interoperability and allows manipulation like removal of watermarks and alteration of metadata, undermining transparency.[408] Controlling the spread of synthetic content, including deepfakes in financial scams, is inherently difficult.[409] While real-name verification and labelling aid authorities in tracking sources (reflecting China's state control focus), DSS providers struggle to stop the spread of decoupled mis/disinformation, a global challenge.[410] Ultimately, without standardized protocols and XAI adoption, the Provisions risk insufficient transparency and explainability in identifying synthetic content, potentially leading to widespread unidentified manipulated media.

Concerns about misuse are evident in real-world cases. The 2019 ZAO deepfakes app controversy, highlighting privacy issues, foreshadowed these risks.[411] Since then, deepfakes have enabled scams, financial fraud, and privacy violations, exemplified by the 2023 Caro-Lailai deepfake pornography case causing severe harm and the Fuzhou AI face-swap fraud costing 4.3 million yuan.[412] The use of AI news anchors for state propaganda further highlights manipulation potential.[413] These instances underscore the necessity of a balanced approach to deepfake regulation.

3.6.2. Integrating compliance, security, and privacy into detection development

The Provisions impose a heavy compliance burden on DSS providers, demanding substantial investments in content moderation, real-name registration, security assessments, and mandated labelling (Articles 6–15). This burden, coupled with broad government oversight (Articles

[398] Provisions on the Administration of Deep Synthesis Internet Information Services (Order No 12 of the Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security, 25 November 2022). <http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm> accessed 6 June 2025.

[399] Allen & Gledhill, China seeks to regulate deep synthesis services and technology. https://www.allenandgledhill.com/sg/publication/articles/22947/seeks-to-regulate-deep-synthesis-services-and-technology, 2023 (accessed 6 June 2025).

[400] J. Xu, Opening the 'black box' of algorithms: regulation of algorithms in China, Communication Research and Practice 10(3) (2024) 288–296. https://doi.org/10.1080/22041451.2024.2346415.

[401] E. Hine, L. Floridi, New deepfake regulations in China are a tool for social stability, but at what cost? Nature Machine Intelligence 4 (2022) 608-610. https://www.nature.com/articles/s42256-022-00513-4.

[402] Art 23 of the Provisions on the Administration of Deep Synthesis Internet Information Services (Order No 12 of the Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security, 25 November 2022). <http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm> accessed 6 June 2025 defines 'deep synthesis technology' as the use of advanced algorithms (e.g., deep learning, virtual reality) to generate or modify digital information, including text, images, audio, video, and virtual environments, encompassing realistic creation or significant alteration of digital content.

[403] S-F. Lee, Deepfakes with Chinese characteristics: PRC influence operations in 2024, China Brief 24(7) (2024). https://jamestown.org/program/deepfakes-with-chinese-characteristics-prc-influence-operations-in-2024/.

[404] Ibid.

[405] F. Schauer, Fear, risk and the First Amendment: unravelling the chilling effect, William & Mary Law School Scholarship Repository (1978) 685-732. https://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=2010&context=facpubs.

[406] C2PA. https://c2pa.org/, founded in 2021 (accessed 6 June 2025).

[407] M. Li, Y. Wan, L. Zhou, H. Rao, 2024. An enhanced governance measure for deep synthesis applications: Addressing the moderating effect of moral sensitivity through message framing. Information & Management. 61, 103982. https://doi.org/10.1016/j.im.2024.103982.

[408] TechCrunch, DEEPFAKES Accountability Act would impose unenforceable rules — but it's a start. https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/, 2019 (accessed 6 June 2025).

[409] Software Engineering Institute, A framework for detection in an era of rising deepfakes. https://insights.sei.cmu.edu/blog/a-framework-for-detection-in-an-era-of-rising-deepfakes/#:~:text=It%20is%20difficult%2C%20but%20not,generation%20techniques%20become%20increasingly%20sophisticated, 2024 (accessed 6 June 2025).

[410] E. Hine, L. Floridi, New deepfake regulations in China are a tool for social stability, but at what cost? Nature Machine Intelligence 4 (2022) 608-610. https://www.nature.com/articles/s42256-022-00513-4.

[411] A. Antoniou, Zao's deepfake face-swapping app shows uploading your photos is riskier than ever, The Conversation. https://theconversation.com/zaos-deepfake-face-swapping-app-shows-uploading-your-photos-is-riskier-than-ever-122334, 2019 (accessed 6 June 2025).

[412] Herbert Smith Freehills, AI-Deep Synthesis Regulations and legal challenges: recent face swap fraud cases in China. https://www.herbertsmithfreehills.com/notes/data/2023-08/ai-deep-synthesis-regulations-and-legal-challenges-recent-face-swap-fraud-cases-in-china, 2023 (accessed 6 June 2025).

[413] H. Levy-Landesberg, X. Cao, Anchoring voices: the news anchor's voice in China from television to AI, Media, Culture & Society 47(2) (2024) 229-251. https://doi.org/10.1177/01634437241270937.

3, 13, 19–22), raises serious concerns about abuse and unfair treatment.[414] Consequently, state control, particularly through the Ministry of Public Security's (MPS) involvement (Article 3) and its history of disinformation campaigns,[415] tensions the fostering of a thriving deepfake detection industry. This dual oversight by the Cyberspace Administration of China and the MPS amplifies the risk of state-sponsored deepfake creation and dissemination.[416] Therefore, unclear assessment criteria and heavy penalties chill innovation, potentially hindering research, development, and investment in the deep synthesis sector.

Commendably, Articles 7 and 14–15 of the Provisions address crucial data security and ethical AI training data for reliable detection methods, while Article 14 specifically mandates that DSS providers instruct users to obtain consent for editing biometric data (faces/voices). However, the real-name user registration requirement (Article 9) introduces a significant tension. Its potential integration with social credit scoring creates serious data protection and privacy concerns, a conflict with fundamental rights addressed differently in the EU (GDPR), UK (DPA), and US (fragmented privacy laws). This social credit system, assessing individual trustworthiness based on financial creditworthiness, social behavior, and adherence to laws,[417] consequently risks government profiling, predictive policing, and behavioral monitoring via big data analytics[418] and facial recognition.[419] Furthermore, while these broad data management requirements heavily burden smaller developers,[420] they will likely not deter sophisticated actors who can easily circumvent authentication systems.[421]

Recent Chinese court rulings significantly impact deepfake detection and AI governance. In 2023, *Li v Liu*[422] extended personality rights to AI voice cloning, reinforcing the need for synthetic voice detection and consent. The 2024 *Ultraman*[423] case held AI platforms liable for user-generated copyright infringement, incentivizing detection tool development. However, the 2025 *Feng v Dongshan Company*[424] critically denied copyright to AI-generated pictures, asserting that less verifiable and controllable human input in AI-generated content weakens its legal protection, complicating deepfake analysis and legal challenges. Collectively, these rulings largely strengthen the legal framework around synthetic content while highlighting the crucial role of provable human contribution for legal protection and its impact on deepfake detection and recourse.

Critics highlight a fundamental divergence in deepfake regulation: China's Deep Synthesis Provisions, prioritizing regime stability, starkly contrast with the more liberal, ex-post remedies in the US and the rights-based approach in the EU.[425] This fragmented global landscape demands domestic initiatives and international collaboration for consistent deepfake detection and moderation standards, favoring solutions like XAI and C2PA which the Chinese Provisions currently overlook.

While China's Deep Synthesis Provisions offer a framework for addressing deepfakes, their broad definitions, compliance burdens, and emphasis on control threaten to hinder innovation, chilling creativity, and raising privacy concerns. Effectively mitigating deepfake harms requires a strategy that integrates national approaches with international cooperation, fosters responsible innovation in deepfake detection, and carefully balances security with individual freedoms. Such a strategy is essential to address harmful deepfakes across diverse political and social systems.

4. Bridging the divide: National strategies and global cooperation in a fragmented world

The escalating threat of malicious deepfakes risks undermining public trust and democratic processes globally. However, as shown in Table 3, the fragmented international regulatory landscape necessitates a pragmatic shift towards robust national-level action, given the potential for unchecked proliferation with limited global consensus. The preceding analysis of diverse regulatory approaches in the EU, US, UK, and China underscores this challenge: the EU's comprehensive yet internally strained approach, the US's federal inaction and state-level fragmentation due to free speech considerations, the UK's indirect online safety focus lacking specific deepfake standards, and China's control-oriented system raising concerns about innovation and freedoms all highlight the difficulty of achieving unified global regulation. This lack of cohesive international action, therefore, strengthens the urgency for strong domestic strategies.

The current geopolitical climate, marked by trade tensions and increasing nationalistic agendas,[426] further undermines the feasibility of widespread international cooperation on deepfakes. This prioritization of national interests often translates to a reluctance to adopt universal standards or cede regulatory authority, making reliance solely on

[414] China Briefing, China to regulate deep synthesis (deepfake) technology starting 2023. https://www.china-briefing.com/news/china-to-regulate-deep-synthesis-deep-fake-technology-starting-january-2023/, 2022 (accessed 6 June 2025).

[415] S-F. Lee, Deepfakes with Chinese characteristics: PRC influence operations in 2024, China Brief 24(7) (2024). https://jamestown.org/program/deepfakes-with-chinese-characteristics-prc-influence-operations-in-2024/.

[416] Ibid.

[417] State Council, 'Shehui xinyong tixi jianshe guihua gangyao (Planning Outline for the Construction of the Social Credit System, 2014-2020)' (14 June 2014); Financial Times, China: when big data meets big brother. https://next.ft.com/content/b5b13a5e-b847-11e5-b151-8e15c9a029fb, 2016 (accessed 6 June 2025).

[418] Y. Chen, A. S. Cheung, The transparent self under big data profiling: privacy and Chinese legislation on the social credit system, The Journal of Comparative Law 12(2) (2017) 356-378. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2992537.

[419] T. Liu, B. Yang, Y. Geng, S. Du, 2021. Research on face recognition and privacy in China—based on social cognition and cultural psychology. Frontiers in Psychology. 12, 809736. https://doi.org/10.3389/fpsyg.2021.809736.

[420] E. Hine, L. Floridi, New deepfake regulations in China are a tool for social stability, but at what cost? Nature Machine Intelligence 4 (2022) 608-610. https://www.nature.com/articles/s42256-022-00513-4.

[421] Biometric Update, Deepfakes and synthetic IDs are already a problem; just wait for the next upgrade. https://www.biometricupdate.com/202410/deepfakes-and-synthetic-ids-are-already-a-problem-just-wait-for-the-next-upgrade, 2024 (accessed 6 June 2025).

[422] *Li v Liu* [2023] Jing 0491 Min Chu No 11279 (Beijing Internet Court) https://english.bjinternetcourt.gov.cn/pdf/BeijingInternetCourtCivilJudgment112792023.pdf accessed 6 June 2025.

[423] *Tsuburaya Productions Co Ltd v Guangzhou Blue Arc Culture Communication Co Ltd* [2024] Yue 0192 Min Chu 113 (Guangzhou Internet Court) https://archive.org/details/scla-v-ai-company-guangzhou-internet-court-02082024-with-english-translation accessed 6 June 2025.

[424] *Feng v Zhangjiagang Dongshan Cultural Commc'n Co Ltd* [2025] (2024) Su 0582 Min Chu No 9015 (People's Court of Zhangjiagang City, Jiangsu Province) https://www.kwm.com/cn/en/insights/latest-thinking/chinese-court-found-ai-generated-pictures-not-copyrightable-convergence-with-the-us-standard.html accessed 6 June 2025.

[425] E. Hine, L. Floridi, New deepfake regulations in China are a tool for social stability, but at what cost? Nature Machine Intelligence 4 (2022) 608-610. https://www.nature.com/articles/s42256-022-00513-4.

[426] P. Fajgelbaum, A. Khandelwal, The economic impacts of the US-China trade war, NBER Working Paper 29315 (2021) 1-28. http://www.nber.org/papers/w29315; P. Fajgelbaum, P. Goldberg, P. Kennedy, A. Khandelwal, D. Taglioni, The US-China trade war and global reallocations, American Economic Review: Insights 6(2) (2024) 295-312. https://www.aeaweb.org/articles?id=10.1257/aeri.20230094; T. Yang, W.-Y. Lau, E. N. A. Bahri, The impact of US-China trade war on China's exports: evidence from difference-in-differences model, SAGE Open (2025) 1-15. https://doi.org/10.1177/21582440251328482.

Table 3
Key differences in global regulatory approaches to deepfake detection.

Approach	Focus	Strengths	Weaknesses
EU	Comprehensive legal framework (AIA, GDPR, DSA)	Risk-based approach, strong data protection, emphasis on fundamental rights	Complexity, potential over-regulation, risk of stifling innovation
US	Limited federal action, fragmented state laws	State-level initiatives address specific harms	Lack of federal privacy law, regulatory uncertainty, inconsistent standards, potential for First Amendment conflicts
UK	Duty of care on online platforms (Online Safety Act)	Addresses online harms, encourages platform responsibility	Broad definitions, lack of technical standards, reliance on policy, potential for over-removal, human rights concerns
China	Centralized control and content regulation (Deep Synthesis Provisions)	Proactive measures, emphasis on data security and ethical training	Broad definitions, compliance burdens, potential for censorship and abuse, privacy concerns, risk of regulatory overreach

comprehensive international solutions unrealistic in the near to medium term.[427] This geopolitical reality reinforces the concerning trajectory of escalating deepfake threats in the absence of collective global action.[428]

In this complex environment, decisive national-level action is paramount for bolstering deepfake defenses and potentially paving the way for future, albeit currently limited, international collaboration.[429] Nations should enact clear and comprehensive legislation precisely defining deepfakes and outlining prohibited uses including financial fraud, non-consensual content, and political manipulation, drawing inspiration from the EU's AIA (recognizing its imperfections)[430] while respecting national constitutional principles. Significant government investment in national research and development of deepfake detection and mitigation technologies, including explainable AI (XAI)[431] and content provenance (C2PA)[432] implementation tailored to national contexts, is crucial to reduce reliance on potentially inaccessible international solutions, as exemplified by the UK's Accelerated Capability Environment initiatives.[433]

Establishing national technical standards for labelling, watermarking, and the interoperability of detection tools within domestic digital ecosystems is vital, creating a baseline for national platforms, as demonstrated by Spain's AI bill mandating clear labelling and imposing substantial fines.[434] Public awareness and media literacy campaigns, such as the TRUE Project 2024, are essential to educate citizens on identifying deepfakes and promoting critical media consumption, empowering individuals as a crucial first line of defence.[435]

Strengthening national data protection frameworks, akin to GDPR, is necessary to govern data collection and use for both deepfake creation and detection, safeguarding individual rights while enabling responsible innovation.[436] Fostering national cross-sector collaboration between research institutions, technology companies, media, and civil society, as seen in the UK's Deepfake Detection Challenge, is key to developing and implementing effective national strategies.[437] Finally, clear national legal frameworks are needed to address the admissibility of deepfake evidence and establish liability for harmful deepfakes.[438]

While ideal international cooperation faces geopolitical hurdles,[439] prioritizing these comprehensive national measures can significantly enhance resilience. Successful national strategies and the development of effective, rights-respecting technologies at the national level could, over time, serve as models and potentially facilitate more targeted and achievable international collaboration in the future.

5. Deepfake detection analysis under UN Resolution 78/265

Following the analysis of the fragmented regulatory landscape and the imperative of national strategies to bridge the divide in addressing deepfakes, this section examines deepfake detection and broader regulatory pathways, focusing on the potential for international cooperation through the framework of UN Resolution 78/265

[427] S. MacIsaac, B.C. Duclos, Trade and conflict: trends in economic nationalism, unilateralism and protectionism, Canadian Foreign Policy Journal 26(1) (2020) 1–7. https://doi.org/10.1080/11926422.2020.1714682.

[428] GOV.UK Department for Science, Innovation & Technology, AI Safety Institute, International AI Safety Report 2025. https://www.gov.uk/government/publications/international-ai-safety-report-2025, 2025 (accessed 6 June 2025).

[429] Ibid.

[430] See the EU AIA analysis in Section 3.1.

[431] European Data Protection Supervisor, TechDispatch #2/2023 - Explainable Artificial Intelligence. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Explainable Artificial Intelligence needs human intelligence. https://www.edps.europa.eu/press-publications/press-news/blog/explainable-artificial-intelligence-needs-human-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Deepfake detection. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en, 2024 (accessed 6 June 2025).

[432] C2PA. https://c2pa.org/, founded in 2021 (accessed 6 June 2025).

[433] GOV.UK Accelerated Capability Environment, Innovating to detect deepfakes and protect the public. https://www.gov.uk/government/case-studies/innovating-to-detect-deepfakes-and-protect-the-public, 2025 (accessed 6 June 2025).

[434] Reuters, Spain to impose massive fines for not labelling AI-generated content. https://www.reuters.com/technology/artificial-intelligence/spain-impose-massive-fines-not-labelling-ai-generated-content-2025-03-11/, 2025 (accessed 6 June 2025).

[435] GOV.UK Government Office for Science, Deepfakes and media literacy. https://www.gov.uk/government/publications/deepfakes-and-media-literacy/deepfakes-and-media-literacy, 2025 (accessed 6 June 2025).

[436] See the EU GDPR analysis in Section 3.2.

[437] GOV.UK Accelerated Capability Environment, Innovating to detect deepfakes and protect the public. https://www.gov.uk/government/case-studies/innovating-to-detect-deepfakes-and-protect-the-public, 2025 (accessed 6 June 2025).

[438] H.B. Dixon Jr., The "Deepfake Defense": an evidentiary conundrum, The Judges' Journal 63(2) (2024) 38-40. https://www.americanbar.org/content/dam/aba/publications/judges_journal/vol63no2-jj2024-tech.pdf; R.A. Delfino, The Deepfake Defense—exploring the limits of the law and ethical norms in protecting legal proceedings from lying lawyers, Ohio State Law Journal 84(5) (2024) 1068-1124. https://ssrn.com/abstract=4355140; C. Kellner, The end of reality? How to combat deepfakes in our legal system, ABA Journal (2025). https://www.abajournal.com/columns/article/the-end-of-reality-how-to-combat-deepfakes-in-our-legal-system.

[439] Peterson Institute for International Economics, Trump's trade war timeline 2.0: An up-to-date guide. https://www.piie.com/blogs/realtime-economics/2025/trumps-trade-war-timeline-20-date-guide, 2025 (accessed 6 June 2025).

Table 4
Key recommendations from UN Resolution 78/265 on safe, secure, and trustworthy AI.

Section	Key Themes	Main Points
Global AI Content ID Framework	International cooperation, robust C2PA tools, transparency	Promote robustness, accessibility, adaptability, and international interoperability (through labeling and watermarking) to support media verification; guided by the UN and EU with emphasis on multi-layered security and information sharing
Global Collaboration, Deepfake Detection Training and Testing	Standards, data sharing, UN database	Establish internationally interoperable standards for AI training and testing ensuring fairness, accuracy, and data protection; cross-border enforcement aligned with GDPR; UN centralized database to support global coordination
Intellectual Property	Copyright, AI training data	Deepfake datasets raise copyright concerns; standardized data usage protocols including consent, compensation, and attribution are essential; emphasize transparency and explore solutions like CR and NFTs
Data Privacy, Transparency and Accountability	Data protection, GDPR, biometric data, transparency, XAI, accuracy	GDPR compliance is critical, particularly for biometric data; implement DPIAs and PETs; ensure XAI decisions, robust human oversight for bias mitigation, and verification of vendor claims; balance transparency with trade secrets
Safeguards and Impact Assessments	Impact assessments, XAI, bias	Develop agile legislation; conduct comprehensive lifecycle risk assessments to uphold human rights and mitigate harm; evaluate explainability, algorithmic bias, and ensure diverse datasets to avoid discriminatory outcomes

concerning safe, secure, and trustworthy AI systems.[440] Although non-binding,[441] as shown in Table 4, this Resolution offers crucial, internationally recognized principles that extend beyond mere technical interoperability, encompassing ethical considerations, data governance, and frameworks for international collaboration relevant to combating deepfakes and aiming to facilitate future international cooperation. The Resolution advocates for robust, accessible, adaptable, and internationally interoperable tools to combat harmful content, including deepfakes, emphasizing transparency, accountability, and global cooperation in addressing crucial areas such as data protection, copyright, and bias.[442]

[440] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.
[441] United Nations, How decisions are made at the UN. https://www.un.org/en/model-united-nations/how-decisions-are-made-un, (accessed 6 June 2025).
[442] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

5.1. Global AI content ID framework: Upholding UN robustness, accessibility, adaptability and interoperability

The escalating global challenge of AI-generated content, including deceptive deepfakes, necessitates a coordinated international response. This involves developing robust, accessible, adaptable, and interoperable deepfake detection tools. UN Resolution 78/265 provides a crucial foundation, advocating for labelling and watermarking tools to empower individuals in verifying the authenticity and origin of digital media.[443] This aligns with the EU AIA (Recital 133), which mandates the ongoing refinement of techniques such as watermarking, labelling, metadata identification, and cryptography to effectively combat harmful deepfakes.

C2PA offers a valuable framework by embedding provenance information within media using metadata, fingerprinting, and watermarking.[444] However, the potential for manipulation of even C2PA metadata[445] underscores the need for continuous improvement, diligent monitoring, and layered security approaches for deepfake detection tools, in line with GDPR Article 32. Malicious AI-generated content poses broad societal threats: political misinformation via manipulated candidate videos, financial fraud via synthesized executive audio, and non-consensual deepfake pornography.

Robust detection requires a multi-layered approach. As highlighted by CJEU case law (e.g., *Scarlet Extended*,[446] *Netlog*,[447] *UPC Telekabel*),[448] online content authentication measures must be adaptable, necessary, and proportionate to balance effective detection with the protection of fundamental rights. This approach includes employing cryptographic techniques to ensure data integrity and prevent tampering, as well as tamper detection to identify alteration attempts.[449] Furthermore, blockchain-based provenance tracking offers an immutable record of content origin and modifications,[450] aligning with GDPR's purpose limitation and data minimization (Articles 5(1)(b, c)), and its security (Article 32) is crucial for preventing tampering and ensuring integrity.

Transparency and accessibility of information regarding the origin and authenticity of digital content—key principles of the GDPR (Article 5(1)(a) – transparency) and the DSA (Chapter III, focusing on transparency obligations of online platforms)—are paramount in combating deepfakes. The C2PA's emphasis on clear communication and adherence to Web Content Accessibility Guidelines ensures this information is readily understandable for all users, including those with disabilities or

[443] Ibid.
[444] C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).
[445] Hacker Factor, C2PA from the attacker's perspective. https://www.hackerfactor.com/blog/index.php?/archives/1031-C2PA-from-the-Attackers-Perspective.html, 2024 (accessed 6 June 2025).
[446] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [47]-[54].
[447] Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [45]-[52].
[448] Case C-314/12 *UPC Telekabel Wien GmbH v Constantin FilmVerleih GmbH and Wega Filmproduktionsgesellschaft GmbH* [2013] EU:C:2014:192 [62]-[64].
[449] S. Longpre, R. Mahari, N. Obeng-Marnu, W. Brannon, T. South, K. Gero, S. Pentland, J. Kabbara, Data authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?, Proceedings of Machine Learning Research 235 (2024) 32711–32725. https://proceedings.mlr.press/v235/longpre24b.html.
[450] J. Collomosse, A. Parsons, To authenticity, and beyond! Building safe and fair generative AI upon the three pillars of provenance, IEEE Computer Graphics and Applications 44(3) (2024) 82–90. https://dl.acm.org/doi/10.1109/MCG.2024.3380168; A. Vilesov, Y. Tian, N. Sehatbakhsh, A. Kadambi, 2024. Solutions to deepfakes: can camera hardware, cryptography, and deep learning verify real images?. arXiv. arXiv 2407.04169. https://doi.org/10.48550/arXiv.2407.04169.

limited internet access.[451] This also aligns with the CJEU ruling in *Dun & Bradstreet*.[452]

Rapidly advancing AI used by malicious actors for synthetic identities (e.g., fake selfies, documents for identity theft and financial scams)[453] necessitates a dynamic, decentralized ecosystem. This system, critical for long-term adaptability, must integrate information sharing (threat intelligence, best practices), open-source intelligence, adaptive AI detection models, and ongoing research and collaboration among researchers, developers, and policymakers to anticipate and counter emerging deepfake techniques.[454]

An effective global deepfake response demands international interoperability, requiring compatible detection tools. The UN's International Telecommunication Union (ITU) is fostering this by developing industry-wide standards and protocols for content provenance and authenticity (e.g., labelling, watermarking) to ensure seamless functionality across platforms and jurisdictions.[455]

However, achieving this vital global interoperability is challenged by varying legal frameworks, geopolitical tensions, a deficit of international trust,[456] and subjectivity in judging deepfakes based on context, cultural norms, and individual interpretation.[457] Overcoming these obstacles necessitates harmonized legal approaches, diplomatic collaboration, and clear data-sharing protocols for accountability.[458] The threat of malicious deepfakes requires sustained research, robust international cooperation, and continuous adaptation.

5.2. UN standards and global collaboration: Responsible training and testing of deepfake detectors

UN Resolution 78/265 highlights the importance of internationally interoperable frameworks and standards for training and testing AI systems, including deepfake detection tools.[459] This aligns with the AIA's focus on high-quality data and state-of-the-art solutions that ensure fairness, transparency, and responsible development (Recitals 66, 67, 121).

Establishing global standards offers significant benefits: improved accuracy, facilitated cross-border cooperation, and user empowerment in identifying manipulated political videos or fraudulent financial statements. However, diverging legal definitions of defamation, harassment, and election interference, along with variations in freedom of expression and privacy norms, create hurdles.[460] These differences complicate universal guidelines for handling deepfakes in online smear campaigns or non-consensual image sharing. Furthermore, overly rigid standards risk stifling innovation and disadvantaging smaller actors. The CJEU ruling in *Public.Resource.Org*,[461] focused on freely available harmonized standards, highlights open access to foster innovation and prevent barriers to accessing legal standards.

Regarding data protection, cross-border enforcement necessitates robust law enforcement cooperation and common protocols for GDPR-compliant data sharing when investigating international financial crimes involving deepfakes. This includes adhering to GDPR principles: lawful processing, data minimization, and purpose limitation (Articles 5 and 6); ensuring equivalent protection for extra-EU data transfers (CJEU: *Bindl*,[462] *Schrems I*,[463] *Schrems II*);[464] and respecting GDPR requirements on consent, data security (Article 32), and data subject rights (Articles 15–22), particularly during deepfake detection model training using biometric data.

A centralized, UN-managed database of validated deepfake samples and detection methodologies, adhering to ethical guidelines and accessibility standards, could significantly enhance international cooperation in identifying and countering novel deepfake techniques.[465] Data collection and storage must comply with data protection principles, particularly the GDPR's requirements for data minimization, purpose limitation, and data security regarding biometric data. For VLOPs (DSA Chapter III, Section 3), DSA compliance is vital, mandating illegal content identification/removal, risk assessments, mitigation, and transparency reporting with safeguards against widespread misinformation or harmful deepfakes. Consistent with the CJEU ruling in *Google Spain*,[466] individuals must have mechanisms to request the removal of their personal data if misused in stored deepfakes. A central international body, such as a dedicated UN agency (e.g., ITU), could coordinate investigations and information sharing, streamlining cross-border efforts against organized deepfake campaigns.

Deepfake detection standards impact various stakeholders. Technology companies and social media platforms, especially VLOPs (DSA

[451] C2PA, Guiding principles. https://c2pa.org/principles/, 2024 (accessed 6 June 2025).

[452] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [49], [50], [58], [65], [66], [77]; AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 57].

[453] Biometric Update, Deepfake raises concerns head of 2024 US elections. https://www.biometricupdate.com/202411/deepfake-raises-concerns-ahead-of-2024-us-elections, 2024 (accessed 6 June 2025).

[454] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, 2019. Combating fake news: a survey on identification and mitigation techniques. ACM Transactions on Intelligent Systems and Technology. 10, 21. https://doi.org/10.1145/3305260; S. Zannettou, M. Sirivianos, J. Blackburn, N. Kourtellis, 2019. The web of false information: rumours, fake news, hoaxes, clickbait, and various other shenanigans. Journal of Data and Information Quality. 11, 10. https://doi.org/10.1145/3309699; R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: a survey of face manipulation and fake detection, Information Fusion 64 (2020) 131–148. https://doi.org/10.1016/j.inffus.2020.06.014.

[455] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025).

[456] United Nations, Secretary-General urges statesmanship to end geopolitical deadlock, warning humanity 'ever closer to a great fracture', at opening of Annual General Assembly Session. https://press.un.org/en/2024/ga12579.doc.htm, 2024 (accessed 6 June 2025).

[457] UN, 'Intensification of efforts to eliminate all forms of violence against women and girls: technology-facilitated violence against women and girls: Report of the Secretary-General' (2024) UN Doc A/79/500.

[458] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025).

[459] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[460] Ofcom, Use of AI in online content moderation. https://www.ofcom.org.uk/online-safety/safety-technology/online-content-moderation/, 2023 (accessed 6 June 2025).

[461] Case C-588/21 *Public.Resource.Org, Inc. and Right to Know CLG v European Commission* [2024] ECLI:EU:C:2024:201 [89].

[462] Case T-354/22 *Thomas Bindl v European Commission* [2025] ECLI:EU:T:2025:4 [189]-[193].

[463] Case C-362/14 *Maximillian Schrems v Data Protection Commissioner* [2015] ECLI:EU:C:2015:650 [67]-[106].

[464] Case C-311/18 *Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems* [2020] ECLI:EU:C:2020:559 [90]-[202].

[465] While a centralized, UN-managed database of deepfakes has not been formally proposed, initiatives like the Deepfake Detection Challenge and the Coalition for Content Provenance and Authenticity demonstrate the potential for collaborative efforts to combat deepfakes through shared resources and technical standards.

[466] Case C-131/12 *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González* [2024] ECLI:EU:C:2014:317 [89], [99].

Chapter III), must adapt content moderation practices. CJEU case law such as *Glawischnig-Piesczek*[467] clarifies platform responsibility for illegal deepfakes. Standardization also benefits journalists verifying political videos, citizens assessing financial advice authenticity, and those at risk of likeness misuse. Furthermore, CJEU case law on automated decision-making, notably *SCHUFA*[468] and *Dun & Bradstreet*,[469] is crucial for ensuring transparency, explainability, and protecting individual rights.

UN Report A/73/348 on AI and human rights underscores the necessity of independent auditing, certification, and robust redress and accountability mechanisms for victims of AI misuse, including those harmed by malicious deepfakes.[470] Consequently, mitigating deepfake-related harm requires an integrated approach encompassing standardized detection techniques and international cooperation, ensuring victim redress and perpetrator accountability.[471] Ongoing collaboration on adaptable standards, ethical reflection, public awareness, and continuous research remains essential for effective deepfake mitigation.

5.3. UN framework: Copyright and deepfake detection tensions

UN Resolution 78/265 recognizes the importance of protecting intellectual property rights, particularly copyright, for developing effective AI tools.[472] Deepfake detection tools, vital for combating manipulated media (e.g., political satire), often rely on copyrighted training data, creating tension between copyright protection and mitigating harmful deepfakes.

Deepfake dataset sourcing highlights this tension. Datasets like FaceForensics++ (sourced from YouTube videos potentially containing copyrighted material)[473] and WildDeepfake (internet-sourced deepfakes which may include copyrighted content)[474] raise copyright and GDPR concerns due to unauthorized use clashing with data protection principles (Articles 5(1)(a)-(c)). In contrast, the Deepfake Detection Challenge Dataset[475] demonstrates compliance with explicitly consented deepfake synthetic data from paid actors (GDPR Articles 4(11), 7), showcasing a path towards ethically sourced training data for detecting deepfakes.

CJEU case law (e.g., *SABAM v Scarlet*[476] and *Netlog*)[477] highlights the risk of technical measures inadvertently infringing on legitimate user activities (e.g., artistic deepfakes). Overly broad tools could misclassify content as infringement or illegal deepfakes, creating DSA liability for VLOPs (Chapter III, Section 3) and infringing free expression in political commentary or social satire.

The ITU's advocacy for standardized AI training data usage—including enhanced opt-out mechanisms, efficient control over copyrighted works, and clear data licensing and attribution guidelines—offers twofold benefits: protecting copyright holders[478] and ensuring GDPR's right to object (Article 21). These efforts aim to simplify copyright management and ensure compliance with international standards and data protection.

Transparency is paramount in AI development and content moderation. Aligning with GDPR consent (Articles 4(11), 7), the AIA mandates training data disclosure (Article 53(1)(d), Recital 107), and the DSA emphasizes content moderation transparency and accountability, as exemplified by Synthesia's explicit AI avatar consent policy.[479] Global standardization of training data disclosure would enhance accountability and streamline compliance,[480] particularly for VLOPs' DSA risk mitigation (Chapter III, Section 3) regarding malicious deepfakes.

Adobe's proposed Content Credentials (CR) offer a promising solution by embedding metadata into digital content to identify training data. For deepfake detection datasets, CR ensures creator attribution and potential compensation. It also enables verification of ethical sourcing by tracing the data origins of deepfakes. Combining CR with Non-

[467] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:820.

[468] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957.

[469] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117.

[470] David Kaye (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), 'Promotion and protection of the right to freedom of opinion and expression' (29 August 2018) UN Doc A/73/348.

[471] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025); Interpol, Beyond Illusions: unmasking the threat of synthetic media for law enforcement. https://www.interpol.int/content/download/21179/file/BEYOND%20ILLUSIONS_Report_2024.pdf, 2024 (accessed 6 June 2025).

[472] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[473] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, Face Forensics++: learning to detect manipulated facial images, Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 1-11. https://ieeexplore.ieee.org/document/9010912. Sourcing datasets like FaceForensics++ from YouTube raises copyright concerns (unauthorized use of copyrighted videos). It also presents GDPR concerns due to the unlikely availability of a valid legal basis (Article 6). While consent (Articles 4(11), 7) and legitimate interest are general legal grounds, obtaining valid consent for large, pre-existing video datasets is virtually impossible, and legitimate interest would face a stringent, often unmeetable, balancing test. Even with specific criteria for scientific research (Article 89), such unauthorized use risks violating the core principles of lawfulness, fairness, and transparency (Article 5(1)(a)).

[474] B. Zi, M. Chang, J. Chen, X. Ma, Y.-G. Jiang, 2024. WildDeepfake: a challenging real-world dataset for deepfake detection. arXiv. arXiv:2101.01456v2. https://arxiv.org/abs/2101.01456. Sourcing deepfakes from the internet for datasets like WildDeepfake raises copyright concerns (unauthorized use). It also implicates GDPR non-compliance, primarily due to the unlikely availability of a valid legal basis (Article 6). This practice risks violating core principles of lawfulness, fairness, transparency (Article 5(1)(a)), purpose limitation (Article 5(1)(b)), and data minimization (Article 5(1)(c)). While scientific research provisions (Article 89) exist, the proportionality of collecting all online content remains questionable.

[475] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. Canton Ferrer, 2020. The Deepfake Detection Challenge (DFDC) Dataset. arXiv. arXiv:2006.07397. https://doi.org/10.48550/arXiv.2006.07397. The DFDC Dataset exemplifies ethical data sourcing. It comprises synthetic deepfakes, like face swaps and lip-syncing, created from paid actors' performances with their explicit consent (GDPR Article 7).

[476] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [52].

[477] Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [50].

[478] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025).

[479] MIT Technology Review, An AI startup made a hyper realistic deepfake of me that's so good it is scary. https://www.technologyreview.com/2024/04/25/1091772/new-generative-ai-avatar-deepfake-synthesia/, 2024 (accessed 6 June 2025).

[480] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025).

Fungible Tokens (NFTs) could enhance transparency and ethical copyright use by recording ownership on a secure ledger.[481] This combination potentially addresses GDPR data security (Article 32) through auditable training data usage for detecting deepfake copyright infringement.

While CR offers a promising approach, AI copyright presents a fundamental dilemma: balancing creators' rights with innovation. *Getty Images (US) v Stability AI*[482] highlights the complexity of fair use in AI training on vast datasets potentially including copyrighted material used to generate or detect deepfakes of copyrighted characters or artworks. Requiring licenses for every image in massive datasets could stifle innovation due to prohibitive costs.[483] Conversely, unchecked use undermines creators and restricts data essential for effective deepfake detection.[484] Navigating AI copyright issues requires clear legal frameworks and responsible data use. Policymakers should balance creator rights with deepfake detection innovation. Developers should prioritize ethical practices, including seeking consent, using licensed datasets, exploring synthetic data, and promoting solutions like CR.

5.4. UN trustworthiness and security: Deepfake detection, data privacy, transparency and accountability

UN Resolution 78/265 emphasizes robust frameworks encompassing stringent data protection and privacy throughout the AI lifecycle. This Resolution requires transparent, accountable data usage AI practices,[485] principles central to navigating the ethical complexities of deepfake detection.

The GDPR demands adherence to data minimization, purpose limitation, and storage limitation (Article 5(1)(b, c, e)), and crucially, data security (Article 32), especially when processing sensitive biometric data (faces identifying non-consensual deepfakes, voice patterns detecting fraudulent audio). Consistent with CJEU rulings (e.g., *Digital Rights Ireland*,[486] *Ministerstvo*),[487] this necessitates that deepfake detection tools like Sentinel,[488] Oz Forensics,[489] and HyperVerge[490] embed privacy by design and default.

Protecting biometric data (e.g., political deepfakes, financial scams) requires robust security throughout its lifecycle—encryption, access controls (GDPR Article 32, as recognized in *SCHUFA*)[491]—and effective breach response plans (Article 33). The CJEU's emphasis on strict necessity in *KNLTB*[492] and *HTB Neunte Immobilien Portfolio*[493] necessitates privacy-enhancing technologies (PETs),[494] especially when analyzing data from diverse populations. Crucially, addressing inherent biases in datasets is paramount, as demonstrated by the UK *R (Bridges)*.[495] Diverse training data, coupled with rigorous Data Protection Impact Assessments for high-risk processing (GDPR Article 35), are essential to prevent discriminatory outcomes[496] in detecting deepfakes.

Beyond data protection, transparency is critical to build trust and ensuring accountability in deepfake detection. Effective reporting on these tools requires XAI, providing accessible explanations of why AI decisions are made (e.g., textual/visual highlighting inconsistencies in non-consensual deepfake identification, reasons for flagging political deepfakes) and identifying how key detection features contribute.[497] This aligns with the AIA's emphasis on traceability and explainability (Recital 27), the DSA's transparency mandates for content moderation and removal (Articles 14, 17), and the CJEU's emphasis on accessible AI information (e.g., *Dun & Bradstreet*).[498]

The lack of standardized, independent validation for vendor-claimed accuracy rates (e.g., HyperVerge at 98.5 %)[499] undermines the DSA's effectiveness in mitigating deepfake disinformation (Articles 34 and 35). Therefore, prioritizing open validation standards, similar to the US *Daubert*[500] standard, and wider access to independently verified tools are vital for fostering public scrutiny. Companies like DuckDuckGoose,[501] and Reality Defender,[502] are leading the way in transparency

[481] J. Collomosse, A. Parsons, To authenticity, and beyond! Building safe and fair generative AI upon the three pillars of provenance, IEEE Computer Graphics and Applications 44(3) (2024) 82–90. https://dl.acm.org/doi/10.1109/MCG.2024.3380168.

[482] *Getty Images (US), Inc. v. Stability AI, Inc.*, 1:23-cv-00135 (US District Court. Del. 2023); *Getty Images (US) Inc & Ors v Stability AI Ltd* [2023] EWHC 3090 (Ch).

[483] M. A. Lemley, B. Casey, Fair learning, Texas Law Review 99(4) (2021) 743-785. https://texaslawreview.org/fair-learning/.

[484] World Intellectual Property Organization, Generative AI navigating intellectual property. https://www.wipo.int/export/sites/www/about-ip/en/frontier_technologies/pdf/generative-ai-factsheet.pdf, 2024 (accessed 6 June 2025).

[485] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[486] Joined Cases C-293/12 and C-594/12 *Digital Rights Ireland and Seitlinger and Others* [2014] ECLI:EU:C:2014:238 [29], [36], [37], [38], [52], [66],[67].

[487] Case C-205/21 *Ministerstvo na vatreshnite raboti (Enregistrement de données biométriques et génétiques par la police)* [2023] ECLI:EU:C:2023:49 [63], [116], [126], [127], [128].

[488] Sentinel AI, Defending against deepfakes and information warfare. https://thesentinel.ai/, (accessed 6 June 2025).

[489] Oz Forensics, Liveness detection and biometric software. https://ozforensics.com/#main_window, founded in 2017 (accessed 6 June 2025).

[490] HyperVerge, Deepfake Detection. https://hyperverge.co/use-cases/deepfake-detection/, founded in 2013 (accessed 6 June 2025).

[491] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [65]-[66].

[492] Case C-621/22 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECLI:EU:C:2024:857 [42], [51], [57], [58].

[493] Joined Cases C-17/22 and C-18/22 *HTB Neunte Immobilien Portfolio geschlossene Investment UG & Co. KG and Ökorenta Neue Energien Ökostabil IV geschlossene Investment GmbH & Co. KG v Müller Rechtsanwaltsgesellschaft mbH and Others* [2024] EU:C:2024:738 [51], [59], [73], [74], [76], [78].

[494] Information Commissioner's Office, Privacy-enhancing technologies (PETs). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/, 2023 (accessed 6 June 2025).

[495] *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police* [2020] EWCA Civ 1058 [188], [192], [199].

[496] The EU AI Act reinforces the need for diverse training data and robust risk assessments for AI systems, particularly those deemed 'high-risk' (Articles 10 and 35).

[497] I. Ul Haq, K.M. Malik, K. Muhammad, 2024. Multimodal neurosymbolic approach for explainable deepfake detection. ACM Transactions on Multimedia Computing Communications and Applications 20, 341. https://doi.org/10.1145/3624748; European Data Protection Supervisor, TechDispatch #2/2023 - Explainable Artificial Intelligence. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Explainable Artificial Intelligence needs human intelligence. https://www.edps.europa.eu/press-publications/press-news/blog/explainable-artificial-intelligence-needs-human-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Deepfake detection. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en, 2024 (accessed 6 June 2025).

[498] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [49], [50], [58], [65], [66], [77]; see also AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 57].

[499] HyperVerge, Deepfake Detection. https://hyperverge.co/use-cases/deepfake-detection/, founded in 2013 (accessed 6 June 2025).

[500] *Daubert v Merrell Dow Pharmaceuticals Inc.,* 509 US 579, 589–595 (1993).

[501] DuckDuckGoose AI, https://www.duckduckgoose.ai/, founded in 2020 (accessed 6 June 2025).

[502] Reality Defender, Visual deepfake detection explainability. https://www.realitydefender.com/blog/visual-deepfake-detection-explainability, 2024 (accessed 6 June 2025).

by providing detailed performance metrics and leveraging XAI in detecting harmful deepfakes.

CJEU rulings *Dun & Bradstreet*[503] and *SCHUFA*[504] enable Data Protection Authorities and courts to balance trade secret protection (Recital 63 GDPR) with AI transparency. This allows for achieving meaningful information about decision-making logic (GDPR Article 15(1)(h)) without full algorithm disclosure (GDPR Article 12(1), Recitals 58, 63).[505] This empowers individuals to understand data usage in detecting deepfakes used against them. Furthermore, it enables them to challenge automated decisions (GDPR Article 22(1)).[506] The DSA further reinforces accountability through complaint mechanisms (Article 20), conforming to the GDPR human review of automated decisions requirement (Article 22(2)(b)). VLOPs must publish transparency reports (Article 42) detailing content moderation practices related to widespread deepfake campaigns, aligning with the CJEU's focus on data security and integrity (e.g., *Schrems II*).[507] Responsible deepfake detection demands integrated legal frameworks, continuous technical advancements, and unwavering adherence to industry best practices prioritizing data privacy, transparency, and accountability.

5.5. UN Resolution on ethical AI: Responsible deepfake detection, safeguards and impact assessments

UN Resolution 78/265 emphasizes imperative safeguards and impact assessments for all AI systems, directly applicable to deepfake detection.[508] The reliance of these tools on substantial datasets raises concerns about privacy, bias, and potential misuse. To mitigate risks, the World Economic Forum advocates proactive scenario planning and stakeholder engagement, requiring adaptable regulations and regulatory sandboxes.[509]

Existing legal frameworks offer a foundation for responsible development and deployment. Specifically, UN Report A/73/348 advocates for lifecycle impact assessments to ensure transparency throughout an AI system's development.[510] The EU AIA (Article 27) mandates Fundamental Rights Impact Assessments for high-risk AI, including deepfake detection tools used in sensitive areas like political content moderation or law enforcement investigations, to safeguard individual rights.[511] GDPR (Article 35) requires Data Protection Impact Assessments for biometric data processing, which is crucial for deepfake detection.[512] Furthermore, the DSA (Articles 34, 35) obligates VLOPs and VLOSEs to assess and mitigate systemic risks associated with deepfakes, implicitly promoting the use of XAI.[513]

To ensure compliance with GDPR (Article 15(1)(h)) and CJEU rulings, including *SCHUFA*[514] and *Dun & Bradstreet*,[515] Algorithmic Impact Assessments (AIAs) for deepfake detection must prioritize explainability alongside accuracy. AIAs must incorporate dedicated explainability assessments that provide clear and accessible explanations of the technology's logic.[516] This requires a comprehensive media analysis, clearly detailing the analysis procedure, including facial tracking, audio analysis, and temporal inconsistency checks.[517] Detection principles must be transparently outlined, explicitly stating the core detection principles, such as lip-sync anomalies, blink rate, and lighting discrepancies.[518] Granular data point disclosure is essential, specifying data points, how they are weighted, how algorithms are trained, and how thresholds are set, including error margins.[519] XAI implementations must adopt explicit methodologies, providing concrete outputs like confidence scores and heatmaps,[520] and validating performance through adversarial XAI benchmarking.[521] Accessible explanations are crucial, ensuring they are understandable for diverse users, including law enforcement,

[503] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [69], [70], [72], [74], [75].

[504] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [59], [60]; AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 54], [AG 57], [AG 58].

[505] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [57], [59]-[61]; Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [53], [66].

[506] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [55], [56], [58]; Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [66].

[507] Case C-311/18 *Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems* [2020] ECLI:EU:C:2020:559 [94], [96], [179]-[187], [191], [192].

[508] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[509] World Economic Forum, Governance in the age of generative ai: a 360º approach for resilient policy and regulation White Paper October 2024. https://www3.weforum.org/docs/WEF_Governance_in_the_Age_of_Generative_AI_2024.pdf, 2024 (accessed 6 June 2025).

[510] David Kaye (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), 'Promotion and protection of the right to freedom of opinion and expression' (29 August 2018) UN Doc A/73/348.

[511] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013 and (EU) 2018/858, and Directives (EU) 2015/2366 and (EU) 2020/1828 [2024] OJ L, 2024/1689, art 27.

[512] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1, art 35.

[513] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1, arts 34, 35.

[514] Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [58], [59], [61].

[515] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [49], [58], [60].

[516] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [49], [50], [58], [65], [66], [77]; Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [56], [57]; AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 57], [AG 58].

[517] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [38], [42], [43], [50], [58], [61], [65], [66], [77].

[518] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [38], [41]-[46], [50], [58], [60], [61], [65], [66], [77]; Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [56], [57].

[519] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [24], [42]-[43], [62]; Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [56], [57]; AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 58].

[520] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artificial Intelligence 267 (2019) 1-38. https://doi.org/10.1016/j.artint.2018.07.007.

[521] H. Baniecki, P. Biecek, 2024. Adversarial attacks and defenses in explainable artificial intelligence: a survey. Information Fusion. 107, 102303. https://doi.org/10.1016/j.inffus.2024.102303; N. Agrawal, I. Pendharkar, J. Shroff, J. Raghuvanshi, A. Neogi, S. Patil, R. Walambe, K. Kotecha, A-XAI: adversarial machine learning for trustable explainability, AI and Ethics 4 (2024) 1143-1174. https://link.springer.com/article/10.1007/s43681-023-00368-4.

journalists, and the public, to foster accountability.[522]

AIAs, as mandated by EU AIA Article 10(5), must include rigorous bias assessments, that evaluate demographic disparities in deepfake detection. This requires testing on diverse datasets and transparently reporting performance metrics. AIAs must provide concrete statistics on accuracy rates, false positive rates, and false negative rates across ethnicity, gender, and age.[523] For example, UK *R (Bridges)* showed significant gender bias, with a 34 % accuracy rate for men compared to 18 % for women, and an 82 % false positive rate for women.[524] Deepfake detection using Xception on FaceForensic++ also revealed racial bias, misclassifying Black men as fake 39.1 % of the time, compared to 15.6 % for white women.[525] Recognizing that AI systems perpetuate existing societal biases, particularly in facial recognition, and that training data is often the cause,[526] these statistical breakdowns are essential for identifying, mitigating, and rectifying bias, ensuring fairness and preventing harmful outcomes in areas like law enforcement or content moderation. Examples of societal bias in AI must be acknowledged and addressed, including higher error rates in facial recognition for darker-skinned women,[527] gender bias in NLP models,[528] age-related variations in deepfake detection accuracy,[529] and targeted deepfake campaigns against specific demographic groups.[530]

5.6. Reconciling UN Resolution with deepfake detection

The UN Resolution's focus on robust, accessible, adaptable, and internationally interoperable tools, such as labelling and watermarking, calls for detection methods that function across diverse platforms and jurisdictions.[531] This directly addresses current regulatory fragmentation demanding globally applicable solutions.

Building on the necessity of interoperability, the Resolution promotes internationally interoperable frameworks and standards for AI training and testing. This highlights the need for diverse and unbiased datasets to train effective detection models. The Resolution also stresses standardized evaluation metrics—e.g., precision, recall, and F1-score—to ensure consistent global performance assessment, which is vital for building trust and facilitating collaboration.[532]

Furthermore, the Resolution's prioritization of privacy and intellectual property (specifically copyright) necessitates careful consideration of how deepfake detection technologies process sensitive biometric data and the ethical implications of using copyrighted content for AI training datasets.[533] The emphasis on transparency aligns with the increasing importance of XAI in ensuring trustworthiness[534] and the legal admissibility of detection results, particularly for holding malicious actors accountable.[535] Similarly, the stress on accountability and the need for robust safeguards and thorough lifecycle risk impact assessments are directly relevant to deploying deepfake detection tools, especially in sensitive areas, ensuring these tools are human rights-compliant.[536]

The principles of the UN Resolution demonstrate strong alignment with the EU's comprehensive AI governance framework (AIA, GDPR, DSA), providing a robust legal foundation for ethical and rights-respecting deepfake detection. For instance, the AIA's focus on detecting and disclosing manipulated content[537] mirrors the Resolution's call

[522] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [48]-[50], [61]; Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [56], [57]; AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 57].

[523] Organization for Economic Co-operation and Development (OECD), 'Recommendation of the Council on Artificial Intelligence' (22 May 2019) OECD/LEGAL/0449 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; NIST, 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI. 100-1. https://doi.org/10.6028/NIST.AI.100-1; P. Grother, M. Ngan, K. Hanaoka, 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR. 8280. https://doi.org/10.6028/NIST.IR.8280; J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, Proceedings of Machine Learning Research 81 (2018) 1-15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf).

[524] *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police* [2020] EWCA Civ 1058 [188].

[525] Y. Ju, S. Hu, S. Jia, G. H. Chen, S. Lyu, 2023. Improving fairness in deepfake detection. arXiv. arXiv:2306.16635v3. https://doi.org/10.48550/arXiv.2306.16635.

[526] *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police* [2020] EWCA Civ 1058 [188], [192], [199].

[527] S. Perkowitz, The bias in the machine: facial recognition technology and racial disparities, MIT Schwarzman College of Computing (2021). https://mit-serc.pubpub.org/pub/bias-in-machine/release/1; P. Grother, M. Ngan, K. Hanaoka, 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR. 8280. https://doi.org/10.6028/NIST.IR.8280.

[528] K. Stanczak, I. Augenstein, 2021. A survey on gender Bias in Natural Language Processing. arXiv. arXiv:2112.14168. https://doi.org/10.48550/arXiv.2112.14168.

[529] J. Lovato, J. St-Onge, R. Harp, G.S. Lopez, S.P. Rogers, I. Ul Haq, L. Hébert-Dufresne, J. Onaolapo, 2024. Diverse misinformation: impacts of human biases on detection of deepfakes on networks. NPJ Complexity. 1, 5. https://doi.org/10.1038/s44260-024-00006-y.

[530] I. Kaate, J. Salminen, R.A. Al Tamime, S. Jung, B.J. Jansen, 2025. Is deepfake diversity real? Analyzing the diversity of deepfake avatars. Expert Systems With Applications. 269, 126382. https://doi.org/10.1016/j.eswa.2025.126382.

[531] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[532] Ibid.

[533] Ibid.

[534] European Data Protection Supervisor, TechDispatch #2/2023 - Explainable Artificial Intelligence. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Explainable Artificial Intelligence needs human intelligence. https://www.edps.europa.eu/press-publications/press-news/blog/explainable-artificial-intelligence-needs-human-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Deepfake detection. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en, 2024 (accessed 6 June 2025).

[535] H.B. Dixon Jr., The "Deepfake Defense": an evidentiary conundrum, The Judges' Journal 63(2) (2024) 38-40. https://www.americanbar.org/content/dam/aba/publications/judges_journal/vol63no2-jj2024-tech.pdf; R.A. Delfino, The Deepfake Defense—exploring the limits of the law and ethical norms in protecting legal proceedings from lying lawyers, Ohio State Law Journal 84(5) (2024) 1068-1124. https://ssrn.com/abstract=4355140; C. Kellner, The end of reality? How to combat deepfakes in our legal system, ABA Journal (2025). https://www.abajournal.com/columns/article/the-end-of-reality-how-to-combat-deepfakes-in-our-legal-system.

[536] David Kaye (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), 'Promotion and protection of the right to freedom of opinion and expression' (29 August 2018) UN Doc A/73/348.

[537] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013 and (EU) 2018/858, and Directives (EU) 2015/2366 and (EU) 2020/1828 [2024] OJ L, 2024/1689, art 50.

for media verification tools,[538] while the GDPR's DPIA mandate[539] and the DSA's VLOP/VLOSE risk mitigation obligations[540] align with the Resolution's emphasis on risk impact assessments.[541] This existing synergy suggests the EU framework could serve as a practical model for implementing the UN's broader principles regionally and potentially globally.

In this context, various deepfake detection methods—including artifact-based, behavioral/physiological (often incorporating liveness detection), deep learning-based (enhanced through adversarial training), and hybrid multimodal approaches—can be evaluated against the UN Resolution's principles. For example, accessibility implies developing user-friendly and widely deployable tools, potentially integrated into existing online platforms.[542] Accuracy and fairness necessitate rigorously addressing biases in training data to prevent misidentification based on demographic factors.[543] Privacy prioritization calls for adopting PETs in detection processes to minimize data exposure.[544] Transparency underscores the importance of XAI to ensure the reliability and accountability of detection outcomes, building crucial user trust.[545]

Therefore, the UN Resolution offers a valuable ethical and normative framework for guiding deepfake detection technology development and deployment. Its internationally recognized principles can help navigate complex technical and societal challenges, even absent fully harmonized legal frameworks. Serving as a crucial benchmark for national and regional efforts and a potential foundation for future international agreements, it ultimately supports this paper's comprehensive and rights-respecting approach.

6. Towards a trustworthy deepfake ecosystem: Integrating detection, ethics, adaptive governance, and accountability

The emergence of sophisticated deepfakes presents a profound dilemma: while offering potential benefits in artistic expression,[546] accessibility,[547] training,[548] medical treatment,[549] education,[550] and even creating AI avatars of victims for remembrance,[551] their capacity for malicious deployment—perpetrating financial fraud, disseminating political misinformation, and inflicting non-consensual harms—fundamentally threatens the integrity of our digital society and the safeguarding of individual rights. Addressing this complex threat necessitates a holistic and adaptive strategy, one that intricately weaves together advancements in technical detection, robust ethical considerations, agile governance frameworks, and clearly defined accountability measures across all stakeholders.

As this study has explored, the technological landscape of deepfake detection is characterized by a relentless cycle of innovation and circumvention. While methods ranging from artifact analysis to advanced AI and techniques like blockchain and quantum computing offer valuable tools, their effectiveness is continuously tested by increasingly sophisticated generation techniques.

Furthermore, analysis of the legal and regulatory landscape across the EU, US, UK, and China reveals a fragmented and often reactive patchwork of approaches, marked by jurisdictional divergences and internal tensions in balancing innovation with fundamental rights.[552] This lack of international harmonization exacerbates the challenge of effectively combating a technology that inherently rapidly transcends borders.

Moving forward, while robust national strategies are essential, a cohesive and globally coordinated response is paramount. For lawmakers and regulators, the imperative lies in establishing harmonized

[538] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[539] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1, art 35.

[540] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1, arts 34, 35.

[541] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[542] C2PA, Guiding Principles. https://c2pa.org/principles/, 2024 (accessed 6 June 2025).

[543] Organization for Economic Co-operation and Development (OECD), 'Recommendation of the Council on Artificial Intelligence' (22 May 2019) OECD/LEGAL/0449 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; NIST, 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI. 100-1. https://doi.org/10.6028/NIST.AI.100-1; P. Grother, M. Ngan, K. Hanaoka, 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR. 8280. https://doi.org/10.6028/NIST.IR.8280; J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, Proceedings of Machine Learning Research 81 (2018) 1-15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf).

[544] Information Commissioner's Office, Privacy-enhancing technologies (PETs). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/, 2023 (accessed 6 June 2025); Case C-621/22 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECLI:EU:C:2024:857 [42], [51], [57], [58]; Joined Cases C-17/22 and C-18/22 *HTB Neunte Immobilien Portfolio geschlossene Investment UG & Co. KG and Ökorenta Neue Energien Ökostabil IV geschlossene Investment GmbH & Co. KG v Müller Rechtsanwaltsgesellschaft mbH and Others* [2024] EU:C:2024:738 [51], [59], [73], [74], [76], [78].

[545] For an analysis of XIA and C2PA solutions, refer to Section 2.6.

[546] World Intellectual Property Organization, Artificial intelligence: deepfakes in the entertainment industry. https://www.wipo.int/web/wipo-magazine/articles/artificial-intelligence-deepfakes-in-the-entertainment-industry-42620, 2022 (accessed 6 June 2025).

[547] ABC News, David Beckham 'speaks' 9 languages for new campaign to end malaria. https://abcnews.go.com/International/david-beckham-speaks-languages-campaign-end-malaria/story?id=62270227, 2019 (accessed 6 June 2025).

[548] Breacher AI, Deepfake simulations: Preparing your team for deepfake threats. https://breacher.ai/deepfake/deepfake-simulations/, 2024 (accessed 6 June 2025).

[549] NorhJersey, ALS silenced him. But AI and voice banking gave this NJ man a new way to communicate. https://eu.northjersey.com/story/news/health/2024/02/15/ai-voice-banking-nj-educator-als-new-voice/71799482007/, 2024 (accessed 6 June 2025).

[550] CereProc, JFK Unsilenced. https://www.cereproc.com/en/jfkunsilenced, 2018 (accessed 6 June 2025).

[551] 404 Media, 'I loved that AI:' judge moved by AI-generated avatar of man killed in road rage Incident. https://www.404media.co/i-loved-that-ai-judge-moved-by-ai-generated-avatar-of-man-killed-in-road-rage-incident/, 2025 (accessed 6 June 2025).

[552] Deepfake regulation involves several key actors regionally within the EU (European Commission, AI Office, European Data Protection Board, Digital Services Coordinators, national authorities); nationally in the US (Congress, Federal Trade Commission, state legislatures, with support from Department of Homeland Security and National Institute of Standards and Technology); in the UK (Ofcom, Digital Regulation Cooperation Forum); and in China (Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security). Organizations like the World Economic Forum and the Organization for Economic Co-operation and Development advocate for harmonized international governance of generative AI, including responsible deepfake development and use. For an analysis of the implications of these regulatory approaches, see Section 3.

international standards[553] and adaptable national legislation.[554] These frameworks must not only define prohibited uses and establish clear liability for the creation and dissemination of malicious deepfakes but also foster an environment that encourages responsible innovation in detection technologies.[555] Drawing inspiration from the EU's rights-based approach and aligning with the ethical principles outlined in the UN Resolution, nations must strive for legal clarity and interoperability.[556] Furthermore, in line with the legal precedent set in *Cartier v BT*,[557] legal frameworks should adopt the "follow the money" principle, targeting advertising platforms and payment processors that financially benefit from the distribution of harmful deepfakes.

For technology actors, including detection technology providers and large corporations, the focus must be on ethical development and deployment. This entails prioritizing accuracy, reliability, and transparency through the integration of XAI[558] and content provenance standards like C2PA.[559] Collaboration across the industry, coupled with substantial investment in cutting-edge research and robust testing protocols, is essential.[560] Furthermore, a commitment to user education and the promotion of media literacy are crucial responsibilities in empowering individuals to navigate the deepfake landscape.[561] Educating users to identify and interpret AI content labels is vital for assessing credibility and mitigating misinformation.[562] Technology corporations should also actively work to disrupt the economic incentives for deepfake creation and dissemination by collaborating with advertising platforms and payment processors to identify and cut off revenue streams.[563] Social media platforms must implement comprehensive and transparent deepfake policies,[564] encompassing content labelling (C2PA) and watermarking,[565] targeted removal of demonstrably harmful content,[566] and the avoidance of indiscriminate monitoring,[567] while also being held accountable for facilitating the dissemination of illegal content, considering their level of control and involvement, as established in cases like *L'Oréal v eBay*,[568] and *Google v Louis Vuitton*.[569] Deepfake application providers must implement safeguards, including watermarking,[570] explicit consent requirements,[571] and the prohibition of applications facilitating non-consensual nudity (following the precedent set by Google and Apple).[572] Additionally, illegal deepfake sites, such as the now defunct MrDeepFakes (designed to primarily host non-consensual content),[573]

[553] International Telecommunication Union, Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop, the need for standards collaboration on AI and multimedia authenticity 2024 Report. https://acrobat.adobe.com/id/urn:aaid:sc:EU:764a0bb2-52cc-4617-b8c3-690cf6f2d022, 2024 (accessed 6 June 2025).

[554] World Economic Forum, Governance in the age of generative ai: a 360º approach for resilient policy and regulation White Paper October 2024. https://www3.weforum.org/docs/WEF_Governance_in_the_Age_of_Generative_AI_2024.pdf, 2024 (accessed 6 June 2025).

[555] C. Han, A. Li, D. Kumar, Z. Durumeric, 2024. Characterizing the MrDeepFakes sexual deepfake marketplace. arXiv. arXiv:2410.11100v1. https://arxiv.org/html/2410.11100v1.

[556] UNGA Res 78/265 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (21 March 2024) UN Doc A/RES/78/265.

[557] *Cartier International AG & Ors v British Sky Broadcasting Ltd & Ors* [2014] EWHC 3354 (Ch) [197]-[217].

[558] European Data Protection Supervisor, TechDispatch #2/2023 - Explainable Artificial Intelligence. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Explainable Artificial Intelligence needs human intelligence. https://www.edps.europa.eu/press-publications/press-news/blog/explainable-artificial-intelligence-needs-human-intelligence_en, 2023 (accessed 6 June 2025); European Data Protection Supervisor, Deepfake detection. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en, 2024 (accessed 6 June 2025).

[559] C2PA. https://c2pa.org/, founded in 2021 (accessed 6 June 2025).

[560] See, e.g., Accenture, Accenture invests in Reality Defender to help fight deepfake extortion, fraud and disinformation. https://newsroom.accenture.com/news/2024/accenture-invests-in-reality-defender-to-help-fight-deepfake-extortion-fraud-and-disinformation, 2024 (accessed 6 June 2025); Meta AI, Deepfake Detection Challenge Dataset. https://ai.meta.com/datasets/dfdc/, 2020 (accessed 6 June 2025); Google Cloud, MWISE Conference 2024: your front-row seat to the future of cybersecurity. https://cloud.google.com/blog/products/identity-security/mwise-conference-2024-your-front-row-seat-to-the-future-of-cybersecurity, 2024 (accessed 6 June 2025); Google Cloud, Powering the next generation of AI startups with Google Cloud. https://cloud.google.com/blog/topics/startups/ai-startups-at-next24, 2024 (accessed 6 June 2025).

[561] See, e.g., AI for Education, Uncovering deepfakes, classroom guide + discussion questions.https://www.aiforeducation.io/ai-resources/uncovering-deepfakes, 2024 (accessed 6 June 2025); A. Lewis, P. Vu, R. M. Duch, A. Chowdhury, 2023. Deepfake detection with and without content warnings. Royal Society Open Science. 10, 231214. https://doi.org/10.1098/rsos.231214.

[562] S.A. Fisher, Something AI should tell you – the case for labelling synthetic content, Journal of Applied Philosophy 42(1) (2025) 272-286. https://doi.org/10.1111/japp.12758.

[563] *Cartier International AG & Ors v British Sky Broadcasting Ltd & Ors* [2014] EWHC 3354 (Ch) [197]-[217]; 404 Media, Instagram advertises nonconsensual AI nude apps. https://www.404media.co/instagram-advertises-nonconsensual-ai-nude-apps/, 2024 (accessed 6 June 2025); Crikey, 'Nudify' economy: how Australians are paying for sexual deepfakes to exploit others. https://www.crikey.com.au/2025/03/13/nudify-apps-sexual-deepfakes-exploitation-australia-cryptocurrency/, 2025 (accessed 6 June 2025).

[564] Case C-203/22 *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH* [2025] ECLI:EU:C:2025:117 [49], [50], [58], [65], [66], [77]; AG opinion in Case C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 57].

[565] C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).

[566] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821 [34], [35], [37], [41], [45].

[567] Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [35]-[53]; Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [33]-[51]; see also Regulation (EU) 2022/2065 (Digital Services Act), Article 8, which explicitly prohibits imposing a general obligation on intermediary service providers to monitor transmitted or stored information or to actively seek out indications of illegal activity.

[568] Case C-324/09 *L'Oréal SA and others v eBay International AG and others* [2011] ECLI:EU:C:2011:474 [113], [116], [120].

[569] Joined Cases C-236/08 and C-238/08 *Google France SARL and Google Inc. v Louis Vuitton Malletier SA* [2010] ECLI:EU:C:2010:159 [120].

[570] C2PA, Technical Specification. https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html, 2024 (accessed 6 June 2025).

[571] The GDPR distinguishes standard consent (Art 4(11)) from explicit consent, a stricter requirement under Art 9(2)(a) for processing special category data (e.g., race, sexual orientation). Explicit consent requires a clear, specific statement of unambiguous agreement for specified purposes. General conditions for consent (Art 7), such as demonstrability and withdrawal rights, and potential explicit consent requirements for significant automated decisions (Art 22), underscore the GDPR's stringent approach to sensitive data.

[572] TechCrunch, Google Play cracks down on AI apps after circulation of apps for making deepfake nudes. https://techcrunch.com/2024/06/06/google-play-cracks-down-on-ai-apps-after-circulation-of-apps-for-making-deepfake-nudes/, 2024 (accessed 6 June 2025).

[573] CBC News, This Canadian pharmacist is key figure behind world's most notorious deepfake porn site. https://www.cbc.ca/news/canada/mrdeepfakes-porn-website-key-figure-1.7527626, 2025 (accessed 6 June 2025).

should be globally delisted and blocked.[574] The principles of intermediary liability[575] and the allocation of costs for detection and takedown[576] also offer avenues for ensuring platform accountability.

For researchers in AI, computer vision, and digital forensics, the challenge lies in pushing the boundaries of detection capabilities while addressing critical limitations such as accuracy,[577] bias,[578] "greener" tools,[579] and real-time processing efficiency.[580] Understanding the evolving societal impact of deepfakes and informing policy through rigorous, interdisciplinary research is equally vital.

Finally, building resilience against deepfake harms requires a critical and immediate shared societal endeavor. For the global community, fostering critical media literacy and promoting a culture of cautious information consumption are essential.[581] Collaboration across sectors, including media organizations (e.g., PRISA Media)[582] and civil society organizations (e.g., WITNESS),[583] is crucial in raising awareness and developing effective strategies for identifying and mitigating the impact of manipulated media. To protect vulnerable individuals and combat misinformation, public awareness campaigns must educate users about deepfake harms, detection limitations, and ethical responsibilities.[584]

Ultimately, navigating the evolving deepfake landscape demands a dynamic and adaptive ecosystem where technological advancements, ethical considerations, robust governance, and clear accountability mechanisms, including disrupting financial incentives, are seamlessly integrated. This requires ongoing dialogue, collaboration, and a sustained shared commitment from all stakeholders to ensure a future where technology empowers and informs, rather than deceives and endangers humanity.

Declaration of generative AI and AI-assisted technologies in the writing and table creation process

During the preparation of this work, the author utilized Google's Gemini to enhance the readability and language of the text. Additionally, Gemini was employed in the creation of the tables presented within this document. Following the use of these tools, the author meticulously reviewed and edited the content, ensuring accuracy and coherence. The author retains full responsibility for the content of the published article, including the text and all tables.

Declaration of competing interest

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author would like to express his sincere gratitude to the two anonymous peer reviewers for their insightful feedback on previous drafts of this manuscript. Their thoughtful comments and constructive suggestions were invaluable and significantly improved the quality of this paper.

Data availability

No data was used for the research described in the article.

[574] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821 [50], [52]. This ruling established that national courts could order host providers to remove or block access to unlawful content (including identical or equivalent material) worldwide, provided such injunctions comply with relevant international law. This precedent supports calls for the global delisting and blocking of sites, like MrDeepFakes, primarily hosting illegal content such as non-consensual deepfakes.

[575] Intermediary liability for online content is addressed in cases like Case C-324/09 *L'Oréal SA and others v eBay International AG and others* [2011] ECLI:EU:C:2011:474; Joined Cases C-236/08 and C-238/08 *Google France SARL and Google Inc. v Louis Vuitton Malletier SA* [2010] ECLI:EU:C:2010:159; and *Cartier International AG and others v British Telecommunications plc* [2018] UKSC 28, [2018] 1 WLR 3997; see also Regulation (EU) 2022/2065 (Digital Services Act), which establishes liability exemptions for intermediary service providers, detailed in Article 4 ('Mere conduit'), Article 5 ('Caching'), and Article 6 ('Hosting'). These exemptions are conditional on providers meeting specific criteria, such as not initiating or modifying transmissions, and for hosting, lacking actual knowledge of illegal content and acting expeditiously upon awareness.

[576] *Cartier International AG and others v British Telecommunications plc* [2018] UKSC 28, [2018] 1 WLR 3997 [27]-[38].

[577] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, F. Xia, 2024. Deepfake video detection: challenges and opportunities. Artificial Intelligence Review. 57, 159. https://doi.org/10.1007/s10462-024-10810-6.

[578] L. Trinh, Y. Liu, 2021. An examination of fairness of AI models for deepfake detection. arXiv. arXiv:2105.00558v1. https://doi.org/10.48550/arXiv.2105.00558; Y. Xu, P. Terhörst, K. Raja, M. Pedersen, 2024. Analyzing fairness in deepfake detection with massively annotated databases. arXiv. arXiv:2208.05845v4. https://doi.org/10.48550/arXiv.2208.05845; Y. Ju, S. Hu, S. Jia, G. H. Chen, S. Lyu, 2023. Improving fairness in deepfake detection. arXiv. arXiv:2306.16635v3. https://doi.org/10.48550/arXiv.2306.16635.

[579] H. Luo, W. Sun, 2024. Addition is all you need for energy-efficient Language Models. arXiv. arXiv:2410.00907v2. https://doi.org/10.48550/arXiv.2410.00907; Google, SynthID: tools for watermarking and detecting LLM-generated Text. https://g.co/kgs/BM1FSWw, 2024 (accessed 6 June 2025).

[580] N. Bansal, T. Aljrees, D.P. Yadav, K.U. Singh, A. Kumar, G.K. Verma, T. Singh, 2023. Real-time advanced computational intelligence for deep fake video detection. Applied Sciences. 13, 3095. https://doi.org/10.3390/app13053095; A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, F. Xia, 2024. Deepfake video detection: challenges and opportunities. Artificial Intelligence Review. 57, 159. https://doi.org/10.1007/s10462-024-10810-6.

[581] PA, Deepfakes: a human challenge, how can tooling be used to assist humans in deepfake detection?. https://www.weprotect.org/wp-content/uploads/Deepfakes_A-Human-Challenge_PA-Report_v3.pdf, 2024 (accessed 6 June 2025).

[582] Reuters Institute & University of Oxford, How a Spanish media group created an AI tool to detect audio deepfakes to help journalists in a big election year. https://reutersinstitute.politics.ox.ac.uk/news/how-spanish-media-group-created-ai-tool-detect-audio-deepfakes-help-journalists-big-election, 2024 (accessed 6 June 2025).

[583] WITNESS, Deepfakes, synthetic media and generative AI. https://www.gen-ai.witness.org/, launched 2023 (accessed 6 June 2025); WITNESS, Written evidence submitted to the House of Lords Communications and Digital Select Committee on Large language models (LLM0050). https://committees.parliament.uk/writtenevidence/124270/pdf, 2023 (accessed 6 June 2025).

[584] Brookings, Watch out for false claims of deepfakes, and actual deepfakes, this election year. https://www.brookings.edu/articles/watch-out-for-false-claims-of-deepfakes-and-actual-deepfakes-this-election-year/, 2024 (accessed 6 June 2025).