Computer Law & Security Review 62 (2026) 106376

Contents lists available at ScienceDirect

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

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

The enforced technical mandate: A multi-layered governance model for deepfake fraud and biometric integrity

Felipe Romero-Moreno

School of Law, Education and Society, University of Hertfordshire, UK

ARTICLE INFO

Keywords:
Deepfake fraud
Generative AI
Agentic AI
Biometric integrity
Privacy-Enhancing Technologies
Online Safety Act
General Data Protection Regulation
EU AI Act
Digital Services Act
Digital identity assurance

ABSTRACT

The rapid escalation of financial deepfake fraud—driven by the emergence of Fraud-as-a-Service—has outpaced existing regulatory frameworks, creating a critical vulnerability in global digital security. This paper argues that the current legal response remains fragmented, trapped between the European Union's rights-based architecture (General Data Protection Regulation, AI Act, Digital Services Act) and the United Kingdom's safety-oriented technology-forcing imperatives (Online Safety Act). Through a doctrinal and functional comparative analysis, this study constructs a three-layered governance paradigm for the digital economy: Layer 1 (Source Control) identifies a compliance black hole in biometric data erasure; Layer 2 (Distribution Control) contrasts systemic risk management with proactive technical detection; and Layer 3 (Accountability) evaluates the shift toward strict corporate criminal liability. Critically, the study evaluates the June 2026 Digital Omnibus updates, identifying a 12-month governance vacuum created by the disparity between the December 2026 functional bans and the delayed December 2027 application timelines for high-risk systems. The paper concludes by advancing six strategic policy recommendations to counter scalable injection attacks, including the enforcement of NIST IAL2 zero-retention biometric standards and obligatory digital provenance (C2PA). Most notably, it proposes a Transatlantic Regulatory and Financial Interoperability Framework, advocating for the integration of biometric integrity protocols directly into ISO 20022 messaging schemas to enforce a real-time financial blockade against non-compliant jurisdictions.

1. Introduction: governing the deepfake fraud crisis

Generative AI—fuelled by advanced models like xAI's Grok[1] and illicit variants like WormGPT and FraudGPT[2]—has precipitated a global digital trust emergency. This technological pivot has transformed deepfake technology (realistic synthetic media generated by AI, typically targeting identity and voice) from a theoretical risk to a precision financial weapon.[3] The economic fallout is immediate and severe: the average loss from criminal deepfake fraud now exceeds $500,000, with high-profile incidents driving single-event losses of up to $50 million.[4]

The acceleration of AI-driven attacks has catalysed a global "fraud tidal wave".[5] By 2030, projections indicate the financial ramifications of fraud will reach $58.3 billion—a 153% surge from the 2025 baseline.[6] "Frankenstein" synthetic identities—composite personas engineered from blended data—will cost the UK economy £4.2 billion by 2027, revealing a profound systemic asymmetry.[7] While these figures signal a

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

[1] xAI, Grok [software], https://grok.com/, 2024 [Accessed 5 July 2026].

[2] P.V. Falade, Decoding the Threat Landscape: ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks, Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol. 9 (2023) 185–198. https://doi.org/10.48550/arXiv.2310.05595.

[3] D. Elliott, '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 5 July 2026].

[4] Brightside Team, Deepfake CEO Fraud: $50M Voice Cloning Threat CFOs. https://www.brside.com/blog/deepfake-ceo-fraud-50m-voice-cloning-threat-cfos, 2025 [Accessed 5 July 2026].

[5] Juniper Research, Fraud to Cost Financial Institutions $58 Billion Globally by 2029; Growth in Account Takeover Attacks and Authorised Push Payment Fraud Creating Major Challenges. https://www.juniperresearch.com/press/fraud-to-cost-financial-institutions-58bn/, 2024 [Accessed 5 July 2026].

[6] Ibid.

[7] LexisNexis Risk Solutions, The Hallmarks of a Synthetic Identity. https://risk.lexisnexis.co.uk/insights-resources/article/seven-signs-for-spotting-synthetic-identities, 2025 [Accessed 5 July 2026].

https://doi.org/10.1016/j.clsr.2026.106376

Available online 17 July 2026

catastrophic breakdown of traditional verification, they uncover a more profound systemic asymmetry. Deepfake fraud is industrialized, yet the global deepfake detection market remains nascent. With only 59 catalogued providers averaging a meagre £25 million each in funding, private-sector innovation cannot scale to match the velocity of synthetic threats.[8] This market failure necessitates a definitive transition toward mandated architectural integrity.

While existing literature addresses specific dimensions of the deepfake threat (e.g., algorithmic detection, criminal law implications),[9] a pivotal oversight deficit remains: current research lacks a comprehensive, integrated analysis bridging technical fraud prevention and the multi-layered legal lifecycle. In contemporary legal discourse, scholarship has focused narrowly on the UK Online Safety Act (OSA)[10] and non-consensual intimate imagery (NCII),[11] overlooking the architectural depth of the EU's Digital Single Market regulations (General Data Protection Regulation (GDPR),[12] AI Act (AIA),[13] Digital Services Act (DSA)).[14] Significantly, these critiques fail to synthesize how these overarching EU frameworks contrast with and challenge the OSA's distinct technology-forcing directives regarding financial deepfake fraud and the critical conflicts these create with privacy law.

A profound legal tension magnifies this gap: data protection law inherently restricts the deployment of biometric authentication mechanisms to counter this rising threat.[15] Biometric data constitutes special category data under the GDPR, subjecting its processing to the strictest level of legal scrutiny.[16] This paper argues that the legal response to deepfake fraud has crystallized into a unified, three-layered governance model, where legal enforcement and technical defence are inextricably intertwined.

Employing a doctrinal and functional comparative approach, this paper analyses the interplay between the EU's rights-based construct (GDPR, AIA, DSA)[17] and the UK's safety-based composition (OSA)[18] to determine their efficacy against the deepfake lifecycle. This analytical methodology is necessary due to the jurisdictions' divergent, yet globally influential, regulatory strategies. The European Union's GDPR establishes the international standard for data protection,[19] while the post-Brexit OSA creates a unique, technology-forcing liability model that fundamentally shifts the burden of fraud mitigation onto private platforms.[20] Evaluating this divergence—EU's rights-based systemic regulation versus UK's safety-based statutory technological intervention—is the most effective way to pinpoint a future-proof governance model for the global digital economy.

This analysis scrutinizes primary legal texts and jurisprudence from the UK, the Court of Justice of the EU (CJEU), and European Court of Human Rights (ECtHR) are analysed to construct a normative argument: that effective governance requires an obligatory shift from *ex-post* content moderation to *ex-ante* identity verification.[21] This research resolves the apparent conflict between privacy and security by defining the specific architectural threshold—zero-retention biometrics—required to satisfy both regimes.

The analysis proceeds through three governance layers. Layer 1 (Source Control) examines how the GDPR and AIA enforce zero-retention biometrics. Layer 2 (Distribution Control) is split into two distinct regulatory modalities: the EU's systemic risk management under the DSA, and the UK's unique technology-forcing requirements under the OSA, which compel proactive fraud detection. Layer 3 (Accountability) evaluates *ex-post* enforcement, including criminal and corporate liability. Section 6 (Discussion) synthesizes these findings via a US-China comparative analysis and explicitly delineates the paper's primary contributions to the fields of computer science and legal theory. Finally, Section 7 (Conclusion and Policy Recommendations) proposes a Transatlantic Regulatory and Financial Interoperability Framework—advocating for the integration of high-assurance standards (NIST IAL2)[22] with ISO 20022 messaging schemas[23] to enforce a real-time financial blockade against non-compliant jurisdictions.

2. Background: the technical foundation of deepfake fraud

2.1. The move from GANs to diffusion architectures

Deepfakes traditionally relied on Generative Adversarial Networks (GANs), employing a competitive logic where a generator creates synthetic media and a discriminator attempts to identify the forgery, both learning until the output achieves photorealism While GANs remain prevalent for high-speed, real-time identity manipulation—such as live

[8] Department for Science, Innovation & Technology, Deepfake detection technology. https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].

[9] Z. Ke, S. Zhou, Y. Zhou, C.H. Chang, R. Zhang, Detection of AI Deepfake and Fraud in Online Payments Using GAN-Based Models, in: 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE), IEEE, 2025. https://doi.org/10.1109/ICAACE65325.2025.11020513; G. Vecchietti, G. Liyanaarachchi, G. Viglia, Managing deepfakes with artificial intelligence: Introducing the business privacy calculus, J. Bus. Res. 186 (2025) 115010. https://doi.org/10.1016/j.jbusres.2024.115010; A.P. Singh, Legal Implications of Deepfake Technology in Criminal Law, Int. J. Law Manag. Humanit. 8 (2025) 1645–1661. https://doij.org/10.10000/IJLMH.119051.

[10] Online Safety Act 2023 (UK), c. 50.

[11] C. McGlynn, L. Woods, A. Antoniou, Pornography, the Online Safety Act 2023 and the need for further reform, J. Media Law 16 (2024) 211–239. https://doi.org/10.1080/17577632.2024.2357421; B. Kira, When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act, Comput. Law Secur. Rev. 54 (2024) 106024. https://doi.org/10.1016/j.clsr.2024.106024; J. Hörnle, Deepfakes and the Law: Why Britain needs stronger protections against technology-facilitated abuse, Queen Mary University of London. https://www.qmul.ac.uk/law/news/2025/items/deepfakes-and-the-law-why-britain-needs-stronger-protections-against-technology-facilitated-abuse.html, 2025 [Accessed 5 July 2026].

[12] 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), OJ L 119 (2016) 1.

[13] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), OJ L 2024/1689 (2024).

[14] 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), OJ L 277 (2022) 1.

[15] F. Romero-Moreno, Deepfake detection in generative AI: A legal framework proposal to protect human rights, Comput. Law Secur. Rev. 58 (2025) 106162. https://doi.org/10.1016/j.clsr.2025.106162.

[16] Regulation (EU) 2016/679 (General Data Protection Regulation), OJ L 119 (2016) 1, Art 9(1).

[17] Regulation (EU) 2016/679 (General Data Protection Regulation), OJ L 119 (2016) 1; Regulation (EU) 2024/1689 (Artificial Intelligence Act), OJ L 1689 (2024); Regulation (EU) 2022/2065 (Digital Services Act), OJ L 277 (2022) 1.

[18] Online Safety Act 2023 (UK), c. 50.

[19] European Council, The General Data Protection Regulation. https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/, 2025 [Accessed 5 July 2026].

[20] Online Safety Act 2023 (UK), c. 50.

[21] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [101]–[102], [105]–[106]; *K.U. v. Finland*, App. no. 2872/02 (ECtHR, 2 December 2008), [46], [49]; Online Safety Act 2023 (UK), c. 50, ss. 10, 12, 81; *Ofcom v AVS Group* [2025] (Regulatory Enforcement).

[22] NIST, Digital Identity Guidelines, Special Publication 800-63-4, National Institute of Standards and Technology, Gaithersburg, MD, 2025. https://doi.org/10.6028/NIST.SP.800-63-4.

[23] ISO 20022, ISO 20022: The universal financial industry message scheme. https://www.iso20022.org/, 2025 [Accessed 5 July 2026].

face-swapping[24]—they often fail to replicate the complex motion cues and physical realism required to circumvent advanced liveness-detection systems. Traditionally, these technical shortcomings have allowed security frameworks to detect forgeries by analysing inconsistencies in facial biometrics and contextual cues.[25]

The innovation frontier has since shifted toward Diffusion Transformer (DiT) architectures, a paradigm reorientation exemplified by OpenAI's Sora 2.[26] Unlike the frame-by-frame processing of GANs, diffusion models synthesize high-fidelity video by reversing a noise-corruption process. By replacing the traditional U-Net backbone with a Transformer that processes data as spacetime patches, Sora 2 achieved unprecedented object permanence and physical realism. This evolution effectively neutralizes traditional detection methods that rely on frame-by-frame inconsistencies, creating a significantly more resilient form of synthetic deception.[27]

Notably, although OpenAI decommissioned the Sora application on March 24, 2026,[28] its technical legacy remains an irreversible blueprint for video fraud. This situation highlights the Sora paradox: the withdrawal of a centralized, moderated gatekeeper does not mitigate systemic risk when open-source ecosystems like Open Sora 2 systematize the underlying capability.[29]

Consequently, the threat has transitioned from a platform-centric risk to a decentralized "workflow risk".[30] Synthetic content now sidesteps centralized oversight by proliferating through unmoderated channels—including encrypted messaging and "replay attacks"—where threat actors frequently strip provenance metadata, rendering detection significantly harder to enforce.[31]

This development heralds a transition into an agentic epoch, characterized by reasoning models capable of autonomous, multi-step execution. Unlike previous iterations of deepfakes which acted as static masks, agentic AI functions as a reasoning brain. This class of AI can navigate complex social engineering scenarios and autonomously iterate on adversarial prompts in real-time, a capability increasingly validated by benchmarks such as AgentHarm.[32]

2.2. Biometric authentication and evolving threat vectors

Biometric authentication protocols, classified into physiological (face, fingerprint, iris) and behavioural (voice/speech, keystroke dynamics, gait) characteristics, are now primary, high-intensity targets for deepfake fraud. Malicious agents easily harvest and synthesize biometric vectors remotely—specifically facial liveness checks and voice recognition—which now dominate the contemporary threat landscape.[33]

The primary industry defence rests on a two-pronged technical countermeasure: Presentation Attack Detection (PAD), which verifies that the input originates from a live human rather than physical spoofs like masks, and Injection Attack Detection (IAD), which prevents synthetic biometric data from being digitally inserted directly into the system feed.[34]

Reflecting a 2026 industry-wide shift identified by the Biometrics Institute, the sector is moving from simple liveness checks toward a comprehensive "proof of personhood" model.[35] This is an essential evolution because as deepfakes move to the virtual camera level, PAD becomes a legacy control. IAD is now the critical frontier, designed to counter the 2665% surge in digital injection attacks that evade physical camera sensors entirely.[36] Significantly, in a 2024 benchmark case, liveness detection thwarted an operative attempting to impersonate a US resident and infiltrate a federal agency by intercepting the virtual camera feed, causing the synthetic overlay to visibly glitch and exposing the deception.[37]

Europol currently categorizes deepfake technology as a sophisticated form of PAD that utilises Generative Adversarial Networks (GANs) and deep learning to impersonate individuals in real-time.[38] The 2024 Arup incident—where an employee transferred $25 million during a deepfake-compromised video conference—underscores the need for integrated shielding.[39] An $18.5 million Hong Kong cryptocurrency heist further exposed this vulnerability in 2025, when threat actors used a deepfake voice clone of a finance manager to authorize a fraudulent transfer.[40]

Parallel to these substantial heists, the threat has also evolved beyond immediate financial theft toward institutional organizational infiltration. Between 2024 and 2025, successful penetrations by North Korean IT operatives escalated by 220%, driven by the weaponization of generative AI at every stage of the recruitment lifecycle.[41] This development exposes a profound security gap in virtual onboarding

[24] Biometric Update, Goode Intelligence, 2025 Deepfake Detection Market Report & Buyer's Guide. https://www.biometricupdate.com/wp-content/uploads/2025/10/BU-Deepfake-Detection-Market-Report-and-Buyers-Guide.pdf, 2025 [Accessed 5 July 2026].

[25] Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera, et al., Deepfake Media Forensics: Status and Future Challenges, J. Imaging 11 (2025) 73. https://doi.org/10.3390/jimaging11030073.

[26] OpenAI, Sora 2. https://openai.com/index/sora-2/, 2025 [Accessed 5 July 2026].

[27] X. Peng, Z. Zheng, C. Shen, T. Young, X. Guo, et al., Open-Sora 2.0: Training a Commercial-Level Video Generation Model in $200k, arXiv [preprint], 2025. https://doi.org/10.48550/arXiv.2503.09642.

[28] Sora (@soraofficialapp), We're saying goodbye to the Sora app. To everyone who created with Sora, shared it, and built community around it: thank you. https://x.com/soraofficialapp/status/2036546752535470382, 2026 [Accessed 5 July 2026].

[29] Open Sora 2: FREE AI Video Generator, Open Sora 2. https://opensora2.com/, 2025 [Accessed 5 July 2026].

[30] N. Vekiarides, Sora is gone. Deepfake video risk isn't. Here's why, Attestiv. https://attestiv.com/sora-is-gone-deepfake-video-risk-isnt-heres-why/, 2026 [Accessed 5 July 2026].

[31] Ibid.

[32] xAI, Grok 4.1 Model Card, 2025. https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf [Accessed 5 July 2026].

[33] ScienceDirect, Biometric Characteristic. https://www.sciencedirect.com/topics/computer-science/biometric-characteristic, 2025 [Accessed 5 July 2026].

[34] A. Terekhin, Presentation Attacks: What Liveness Detection Systems Protect From Every Day, Regula Forensics. https://regulaforensics.com/blog/presentation-attack-detection/, 2025 [Accessed 5 July 2026].

[35] Biometrics Institute, Biometrics: Keeping it Real – 2026 Silver Jubilee Concepts and Solutions Report. https://www.biometricsinstitute.org/concepts-and-solutions-keeping-biometrics-real/, 2026 [Accessed 5 July 2026].

[36] iProov, Threat Intelligence Report 2025: Remote Identity Under Attack. https://www.iproov.com/reports/threat-intelligence-report-2025-remote-identity-attack, 2025 [Accessed 5 July 2026].

[37] B. Hall, How AI-driven fraud challenges the global economy – and ways to combat it. https://www.weforum.org/stories/2025/01/how-ai-driven-fraud-challenges-the-global-economy-and-ways-to-combat-it/, 2025 [Accessed 5 July 2026].

[38] Europol, Biometric Vulnerabilities: Ensuring Future Law Enforcement Preparedness, Observatory Report. https://www.europol.europa.eu/cms/sites/default/files/documents/Biometric-vulnerabilities.pdf, 2025 [Accessed 5 July 2026].

[39] D. Elliott, '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 5 July 2026].

[40] B. Colman, $18.5 Million Lost in a Crypto Scam: AI Voice Fraud's Growing Threat to Financial Security, Reality Defender. https://www.realitydefender.com/insights/millions-lost-in-hong-kong-crypto-scam, 2025 [Accessed 5 July 2026].

[41] A. Gerut, North Korean IT worker infiltrations exploded 220% over the past 12 months, with gen AI weaponized at every stage of the hiring process, Fortune. https://fortune.com/2025/08/04/north-korean-it-worker-infiltrations-exploded/, 2025 [Accessed 5 July 2026].

workflows and lends significant weight to Gartner's forecast that by 2028, one in four job applicant profiles will be fraudulent.[42]

By 2026, production pipelines have reached massive scale, incorporating "AI face models"—primarily women hired to provide the source imagery for up to 100 fraudulent video calls per day to ensure realistic, human-verified victim engagement.[43] In March, Qingdao police dismantled a network that had generated over 50,000 AI videos to circumvent liveness checks.[44] This operation illustrates the low-cost, high-margin nature of modern fraud: stolen identity data purchased for as little as 20 yuan was transformed into verified accounts worth 300 yuan—a 1400% value increase that formalizes identity theft as a repeatable production system.[45]

Deepfakes have also weaponised traditional phishing into high-fidelity "vishing" (voice and video-based phishing).[46] Using as little as three seconds of audio, attackers can clone a target's voice with such precision that 77% of victims now experience financial loss.[47] These perceptually indistinguishable voice deepfakes enable sophisticated social engineering by exploiting emotional vulnerability, including fake kidnapping ransom calls that bypass a victim's logical scepticism through manufactured emotional proximity.[48]

Beyond biometric evasion, deepfakes now facilitate synthetic document fraud, where threat actors use high-quality, AI-generated IDs to open mule accounts[49] and enhance complex schemes like stock price manipulation.[50] The 2025 Swedish case, in which AI-generated promotional content resulted in the theft of €44.5 million from 5000 individuals, further evidences such coordinated financial deception.[51] This technical convergence—using synthesized media across multiple sensory dimensions—demands a unified, integrated defence against the industrialized and highly scalable nature of modern fraud rings.

3. Layer 1: source control and the architecture of data integrity

Building on the technical foundations delineated in the preceding section, Layer 1 examines the transition of Privacy-Enhancing Technologies (PETs) from discretionary practices to binding legal obligations.[52] It scrutinizes the regulatory frameworks governing the AI architecture and biometric inputs—the raw material of deepfake fraud—to formulate a legal counter-narrative regulating the technology at its foundational inception.

3.1. Controlling the raw material: the data (GDPR) and the compliance paradox

The CJEU Grand Chamber's landmark ruling in *X v Russmedia Digital SRL* ("*Russmedia*")[53] cemented the shift toward architectural imperatives. The 2026 Department for Science, Innovation and Technology (DSIT) report makes this requirement urgent, noting that deepfake detection precision drops by 10%–20% in real-world environments compared to laboratories.[54] Because technical fragility renders *ex-post* detection unreliable, architectures utilizing Zero-Knowledge Proofs (ZKP)[55] or encryption-in-use[56] represent the only viable pathway to satisfy the Court's *Russmedia* verify-then-publish directive[57]—without triggering a "systematic and generalized" processing of biometric data forbidden under *Mousse v CNIL*.[58]

The *Russmedia*[59] workflow is further constrained by the rigorous standards promulgated in *Comdribus*[60] where the CJEU recently clarified that systematic biometric collection remains prohibited even for individuals "reasonably suspected" of an offense. This ruling adds an indispensable layer of procedural protection: competent authorities must provide a "sufficient statement of reasons" detailing why collection is "strictly necessary" in each instance, effectively shifting the burden of proof to the state.[61] This jurisprudence precludes indiscriminate

[42] Gartner, Gartner Survey Shows Just 26% of Job Applicants Trust AI Will Fairly Evaluate Them. https://www.gartner.com/en/newsroom/press-releases/2025-07-31-gartner-survey-shows-just-26-percent-of-job-applicants-trust-ai-will-fairly-evaluate-them, 2025 [Accessed 5 July 2026].

[43] M. Burgess, '100 Video Calls Per Day': Models Are Applying to Be the Face of AI Scams, Wired. https://www.wired.com/story/models-are-applying-to-be-the-face-of-ai-scams/, 2026 [Accessed 5 July 2026].

[44] OECD.AI, AI-Generated Deepfake Videos Used for Large-Scale Identity Fraud in China. https://oecd.ai/en/incidents/2026-03-16-56f6, 2026 [Accessed 5 July 2026].

[45] O.K. Mathurin, Police seized 50,000 AI-generated face videos used to bypass KYC. 20 yuan in → 300 yuan out. This is what industrialized identity fraud looks like [LinkedIn post]. https://www.linkedin.com/posts/omathurin_ai-fraud-identityfraud-activity-7441750665523224576-7ck4/, 2026 [Accessed 5 July 2026].

[46] U.S. Department of the Treasury, 2026 National money laundering risk assessment. https://home.treasury.gov/system/files/246/2026-NMLRA.pdf, 2026 [Accessed 5 July 2026].

[47] McAfee, Beware the Artificial Impostor: A McAfee Cybersecurity Artificial Intelligence Report. https://www.mcafee.com/content/dam/consumer/en-us/resources/cybersecurity/artificial-intelligence/rp-beware-the-artificial-impostor-report.pdf, 2023 [Accessed 5 July 2026].

[48] E. Salam, US mother gets call from 'kidnapped daughter' – but it's really an AI scam, The Guardian. https://www.theguardian.com/us-news/2023/jun/14/ai-kidnapping-scam-senate-hearing-jennifer-destefano, 2023 [Accessed 5 July 2026].

[49] Entrust Cybersecurity Institute, 2025 Identity Fraud Report. https://www.entrust.com/sites/default/files/documentation/reports/2025-identity-fraud-report.pdf, 2025 [Accessed 5 July 2026].

[50] U.S. Department of Homeland Security, Increasing Threats of Deepfake Identities. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf, 2020 [Accessed 5 July 2026].

[51] AI Incident Database, Incident 1256: Purportedly AI-Generated Deepfake Investment Ads Defrauded 5,000 Swedish Investors of 500 Million SEK [dataset]. https://incidentdatabase.ai/cite/1256/, 2025 [Accessed 5 July 2026].

[52] Information Commissioner's Office, Using privacy-enhancing technologies (PETs) to unlock value from data responsibly. https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/11/using-privacy-enhancing-technologies-pets-to-unlock-value-from-data-responsibly/, 2024 [Accessed 5 July 2026]; European Data Protection Supervisor, Secure multi-party computation: powering privacy through collaboration. https://www.edps.europa.eu/press-publications/press-news/blog/secure-multi-party-computation-powering-privacy-through-collaboration_en, 2025 [Accessed 5 July 2026]; European Data Protection Supervisor, TechDispatch #1/2025 - Federated Learning. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-06-10-techdispatch-12025-federated-learning_en, 2025 [Accessed 5 July 2026].

[53] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935.

[54] Department for Science, Innovation & Technology, Deepfake detection technology. https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].

[55] M. Frigo, A. Shelat, Anonymous Credentials from ECDSA. https://eprint.iacr.org/2024/2010.pdf, 2024 [Accessed 5 July 2026].

[56] A. Vemury, Third party biometric homomorphic encryption matching for privacy protection. US Patent US20230403132A1, 2023. https://patents.google.com/patent/US20230403132A1 [Accessed 5 July 2026].

[57] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [104], [106].

[58] Case C-394/23, *Mousse v Commission nationale de l'informatique et des libertés (CNIL),* ECLI:EU:C:2025:2, [42].

[59] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [104], [106].

[60] Case C-371/24, *HW v Ministère public (Comdribus)* ECLI:EU:C:2026:234, [54].

[61] Ibid, [71], [75].

collection, ensuring a user's refusal to provide biometric data is punishable only if the state fulfils its rigorous burden of proving strict necessity.[62]

Deploying biometric authentication to combat deepfake fraud demands elevated oversight, as genuine biometrics constitute special category data under Article 9(1) of the GDPR. CJEU jurisprudence mandates a high threshold for such processing; *OT v Vyriausioji*[63] defines these actions as a "particularly serious interference," while *Mousse v CNIL*[64] corroborates that the acute risks to fundamental rights preclude relying on legitimate interests under Article 6(1)(f). Combating fraud, therefore, requires a more robust legal basis, such as the "substantial public interest" provision under Article 9(2)(g) GDPR. While the Digital Services Act (DSA) obligates the removal of deepfake scams (Articles 6, 35(1)(k)), this requirement grants no exemption from Article 9's strict necessity standards.[65]

Substantial legal friction persists, however, regarding the definition of data usage. As Kindt argues, the GDPR maintains an unstable omission by distinguishing between "having" and "using" biometric data.[66] Because raw facial images only become biometric data after undergoing "specific technical processing" under Article 4(14), organizations may stockpile vast reservoirs of source material without triggering Article 9 protections.[67] This creates latent exploitation risks and provides a ready data supply for unauthorized model training. It also precipitates a conformity deadlock between the GDPR's minimization principle (Article 5 (1)(c)) and the AIA's imperative for high-accuracy systems under Article 15. While the AIA requires expansive, diverse datasets to ensure detection models are robust,[68] the GDPR strictly curtails retaining the very data required to achieve that reliability.[69] To resolve this, the Court in *Ministerstvo na vatreshnite raboti*[70] ascertained that necessity fails where "less restrictive" alternatives exist, a principle upheld in *Mousse v CNIL*,[71] which rejected "systematic and generalized processing" in favour of targeted, proportional protection.

The proposed Digital Omnibus Regulation seeks to codify this resolution via a draft Article 9(2)(l), which introduces a targeted exemption for lawful identity verification provided it relies on data held under the "sole control of the data subject."[72] If enacted, this provision transforms zero-retention architectures into a statutory requirement, establishing a definitive barrier between permissible 1:1 verification and prohibited 1:many identification. To withstand legal challenges, anti-deepfake systems must embed these principles through data protection by design under Article 25.[73] As fraud increasingly targets remote biometrics—specifically facial liveness and voice[74]—default settings must enforce zero-retention by deleting templates immediately following authentication. This satisfies the purpose limitation principle (Article 5(1)(b)) and avoids the indefinite storage barred in *NG v Direktor na Glavna direktsia*.[75] Furthermore, the inherent high risk makes a Data Protection Impact Assessment (DPIA) obligatory (Article 35).[76] Security by design under Article 32 mandates encryption and pseudonymization, measures affirmed by *VB v National Revenue Agency*[77] as proportionate to biometric data hazards. Notably, *Meta Platforms v Bundeskartellamt*[78] expressly prohibits automatically reusing verification data to train detection models without a fresh legal basis, severing the illicit data aggregation loop at the inception of the generative AI lifecycle.

Despite these safeguards, a material enforcement gap persists regarding the right to erasure (Article 17) once a model's parameters ingest biometric data. While the right to be forgotten is fundamental, as settled in *Google Spain*,[79] applying it to generative models creates a procedural deadlock. Removing the influence of specific data points is virtually impossible without precipitating model collapse,[80] and retroactive "machine unlearning"—forgetting specific training samples without full retraining—remains experimentally immature and prone to verification failures.[81]

Unless regulators enforce a strict retraining-on-request command—a penalty the European Data Protection Board (EDPB) acknowledges is financially prohibitive for large-scale models[82]—illicitly trained deepfake models function as compliance black holes. Non-compliance regarding these raw materials triggers the highest tier of administrative fines under Article 83(5)—up to €20 million or 4% of total global annual turnover—providing a decisive financial deterrent against biometric data misuse within the generative AI lifecycle (*Deutsche Wohnen*[83]).

[62] Ibid, [36], [51], [85], [91].

[63] Case C-184/20, *OT v Vyriausioji tarnybinės etikos komisija*, ECLI:EU:C:2022:601, [126].

[64] Case C-394/23, *Mousse v Commission nationale de l'informatique et des libertés (CNIL)*, ECLI:EU:C:2025:2, [64].

[65] F. Romero-Moreno, Deepfake detection in generative AI: A legal framework proposal to protect human rights, Comput. Law Secur. Rev. 58 (2025) 106162. https://doi.org/10.1016/j.clsr.2025.106162

[66] E.J. Kindt, Having yes, using no? About the new legal regime for biometric data, Comput. Law Secur. Rev. 34 (2018) 523–538. https://doi.org/10.1016/j.clsr.2017.11.002.

[67] Ibid.

[68] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, et al., Deepfake Media Forensics: State of the Art and Challenges Ahead, arXiv [preprint], 2024. https://doi.org/10.48550/arXiv.2408.00388.

[69] Regulation (EU) 2016/679 (General Data Protection Regulation), OJ L 119 (2016) 1 Art 5(1)(e).

[70] Case C-205/21, *Ministerstvo na vatreshnite raboti (Recording of biometric and genetic data by the police)*, ECLI:EU:C:2023:59, [126].

[71] Case C-394/23, *Mousse v Commission nationale de l'informatique et des libertés (CNIL)*, ECLI:EU:C:2025:2, [42].

[72] European Commission, Proposal for a Regulation as regards the simplification of the digital legislative framework (Digital Omnibus), COM(2025) 837 final, art 3(3).

[73] Information Commissioner's Office, Data Protection by Design and Default. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/data-protection-by-design-and-default/, 2025 [Accessed 5 July 2026].

[74] T. Stankovic, Liveness Detection: A Complete Guide for Fraud Prevention and Compliance in 2025, Sumsub. https://sumsub.com/blog/face-liveness-detection/, 2025 [Accessed 5 July 2026].

[75] Case C-118/22, *NG v Direktor na Glavna direktsia 'Natsionalna politsia' pri Ministerstvo na vatreshnite raboti*, ECLI:EU:C:2024:97, [43], [59]–[60], [72].

[76] Information Commissioner's Office, Data Protection Impact Assessments (DPIAs). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/, 2025 [Accessed 5 July 2026].

[77] Case C-340/21, *VB v Natsionalna agentsia za prihodite*, ECLI:EU:C:2023:986, [3 (Rec 83)], [7], [26], [42], [47].

[78] Case C-252/21, *Meta Platforms Inc v Bundeskartellamt*, ECLI:EU:C:2023:537, [6], [98]-[99], [108], [142], [151].

[79] Case C-131/12, *Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González*, ECLI:EU:C:2014:317.

[80] C. Novelli, F. Casolari, P. Hacker, G. Spedicato, L. Floridi, Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity, Comput. Law Secur. Rev. 55 (2024) 106066. https://doi.org/10.1016/j.clsr.2024.106066.

[81] K. Shrishak, AI-Complex Algorithms and effective Data Protection Supervision: Effective implementation of data subjects' rights, European Data Protection Board. https://www.edpb.europa.eu/system/files/2025-01/d2-ai-effective-implementation-of-data-subjects-rights_en.pdf, 2025 [Accessed 5 July 2026].

[82] Ibid.

[83] Case C-807/21, *Deutsche Wohnen SE v Staatsanwaltschaft Berlin*, ECLI:EU:C:2023:950, [6(5)], [58], [60].

3.2. Controlling the technology's functionality (EU AI Act)

While the GDPR safeguards the raw material of the deepfake economy's raw material through strict biometric non-retention, the AI Act (AIA) shifts the regulatory focus to the functionality and deployment of the technology itself. This legal binarism is central, yet it introduces a substantive jurisdictional tension between *ex-post* data protection and *ex-ante* product safety. The AIA substantially captures deepfake fraud under the "unacceptable risk" prohibition of Article 5, establishing an uncompromising red line against systems designed to deploy purposefully deceptive techniques that materially distort human behaviour and inflict significant harm.[84]

However, the efficacy of an *ex-ante* prohibition against a criminal market characterized by negligible entry costs and rapid evolution remains a paramount challenge for European regulators. Addressing this gap, the recently adopted Digital Omnibus Regulation—approved by the European Parliament[85] and the Council[86] in June 2026—explicitly expands the Article 5 list of prohibited practices to include AI "nudifier" apps. Establishing a strict compliance deadline of 2 December 2026, this prohibition targets systems capable of generating non-consensual intimate imagery of identifiable individuals, fundamentally outlawing systems placed on the market without adequate technical safeguards.[87]

3.2.1. Unacceptable risk: analyzing systemic vulnerability and the scale of harm

Classifying deepfake fraud as an unacceptable risk rests primarily on Article 5(1)(a)—addressing harmful manipulation—and Article 5(1)(b), which concerns the exploitation of specific vulnerabilities. The unprecedented scale and psychological gravity of current deepfake attacks unequivocally satisfy the "significant harm" criterion. This threshold captures high-value financial crimes, exemplified by the fraudulent authorization of $25.5 million following the 2024 Arup deepfake video conference.[88] Macro-economically, the "fraud tidal wave" is projected to reach $58.3 billion globally by 2030—a 153% surge from 2025 levels.[89] Within this landscape, deepfakes already constitute approximately 40% of all recorded biometric fraud.[90] Beyond financial metrics, the criterion encompasses severe psychological trauma. Recent data indicates 35% of victims experience profound emotional distress,[91] while 16% of victims reported considering suicide in 2023—a figure that has doubled since 2021 and represents more than triple the US national average of approximately 5%.[92]

The technology's marginal cost and accelerated production cycles, in which a sixty-second non-consensual video generates in mere minutes,[93] fuel diverse, lucrative schemes. These range from "grandparent scams," involving voice cloning to socially engineer the elderly into wiring funds (costing victims over $21million),[94] to sophisticated sextortion rackets demanding immediate payment under the threat of distributing synthesized explicit imagery.[95] The gendered, predatory nature of this technology is evidenced by 98% of deepfake videos constituting non-consensual pornography,[96] with platforms like Snapchat logging approximately 20,000 sextortion cases targeting minors.[97]

Structural vulnerabilities in human cognition amplify this widespread prejudice; individual deepfake detection capabilities are severely limited, typically hovering within the 55%–60% accuracy range.[98] This substantive failure of human cognitive resilience, corroborated by the CJEU's rationale in *Compass Banca*[99] and Leiser's analysis of AI-driven "psychological patterns," suggests the Article 5 prohibition represents a structural necessity rather than a mere policy preference.[100] The technology does not merely inform; it subverts user autonomy and distorts decision-making processes beyond conscious awareness.[101] This poses a pervasive threat to democratic discourse and public trust, a risk that the European Commission pointedly links to facilitating criminal

[84] F. Romero-Moreno, Generative AI and deepfakes: a human rights approach to tackling harmful content, Int. Rev. Law Comput. Technol. 38 (2024) 297–326. https://doi.org/10.1080/13600869.2024.2324540.

[85] European Parliament, AI Act: EP approves simplification measures and "nudifier" app ban. https://www.europarl.europa.eu/news/en/press-room/20260611IPR45207/ai-act-ep-approves-simplification-measures-and-nudifier-app-ban, 2026 [Accessed 5 July 2026].

[86] Council of the European Union, Artificial Intelligence: Council gives final green light to simplify and streamline rules, Press Release. https://www.consilium.europa.eu/en/press/press-releases/2026/06/29/artificial-intelligence-council-gives-final-green-light-to-simplify-and-streamline-rules/, 29 June 2026 [Accessed 5 July 2026].

[87] European Parliament, AI Act: EP approves simplification measures and "nudifier" app ban. https://www.europarl.europa.eu/news/en/press-room/20260611IPR45207/ai-act-ep-approves-simplification-measures-and-nudifier-app-ban, 2026 [Accessed 5 July 2026].

[88] D. Elliott, '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 5 July 2026].

[89] Juniper Research, Fraud to cost financial institutions $58.3 billion by 2030 globally, as synthetic identities threaten fraud tidal wave. https://www.juniperresearch.com/press/fraud-to-cost-financial-institutions-58bn/, 2025 [Accessed 5 July 2026].

[90] Entrust Cybersecurity Institute, 2025 Identity Fraud Report. https://www.entrust.com/sites/default/files/documentation/reports/2025-identity-fraud-report.pdf, 2025 [Accessed 5 July 2026].

[91] J. Dhaliwal, State of the Scamiverse – How AI is Revolutionizing Online Fraud, McAfee Blog. https://www.mcafee.com/blogs/internet-security/state-of-the-scamiverse/, 2025 [Accessed 5 July 2026].

[92] Identity Theft Resource Center, 2023 Consumer Impact Report. https://www.idtheftcenter.org/wp-content/uploads/2023/08/ITRC_2023-Consumer-Impact-Report_Final-1.pdf, 2023 [Accessed 5 July 2026].

[93] G. Regan, Nonconsensual Deepfake Pornography: A Growing Concern in the Age of AI, Reality Defender. https://www.realitydefender.com/insights/what-is-deepfake-pornography, 2024 [Accessed 5 July 2026].

[94] Federal Communications Commission, 'Grandparent' Scams Get More Sophisticated. https://www.fcc.gov/consumers/scam-alert/grandparent-scams-get-more-sophisticated, 2025 [Accessed 5 July 2026].

[95] United Nations General Assembly, 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, A/79/500. https://undocs.org/A/79/500, 2024 [Accessed 5 July 2026]; UN Women, How AI is exacerbating technology-facilitated violence against women and girls. https://www.unwomen.org/sites/default/files/2026-01/how-ai-is-exacerbating-technology-facilitated-violence-against-women-and-girls-en.pdf, 2025 [Accessed 5 July 2026].

[96] Security Hero, 2023 State of Deepfakes: Realities, Threats, and Impact. https://www.securityhero.io/state-of-deepfakes/, 2023 [Accessed 5 July 2026].

[97] J. Halliday, FBI and NSPCC alarmed at 'shocking' rise in online sextortion of children, The Guardian. https://www.theguardian.com/society/2025/aug/09/fbi-nspcc-alarmed-shocking-rise-online-sextortion-children, 2025 [Accessed 5 July 2026].

[98] B. Colman, Why detecting dangerous AI is key to keeping trust alive in the deepfake era, World Economic Forum. https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/, 2025 [Accessed 5 July 2026].

[99] Case C-646/22, *Compass Banca SpA v Autorità Garante della Concorrenza e del Mercato*, ECLI:EU:C:2024:957, [53], [57], [59].

[100] M. Leiser, Psychological Patterns and Article 5 of the AI Act: AI-Powered Deceptive Design in the System Architecture and the User Interface, Artif. Intell. Regul. 1 (2024) 5–23. https://www.lexxion.eu/wp-content/uploads/2024/02/AIRe_1-24_Article_Leiser.pdf.

[101] Ibid.

offenses.[102]

The inherent fragility of current technical controls is further demonstrated by the challenges facing high-profile generative models. For instance, following the release of the now-discontinued OpenAI's Sora 2[103] saw detection firms[104] and third-party applications elude commercial safety guardrails within twenty-four hours.[105] INTERPOL reports from 2026 indicate that threat actors now methodically integrate sextortion into romance fraud and investment scams, utilizing deepfakes to scale victimization to an industrial level.[106] The early 2026 Grok scandal—which exposed a catastrophic absence of output filters—serves as a telling proxy for deepfake fraud vulnerabilities.[107] If a model lacks the architectural integrity to prevent nudify requests at scale, it is mathematically incapable of reliably blocking fraudulent impersonation. In truly open-source architectures, such as Open Sora 2,[108] the regulatory challenge magnifies; downstream actors' ability to modify and redistribute models—a practice explicitly permitted under Recital 102 AIA—ensures that safety filters can be stripped with minimal effort.[109] Thus, Layer 1 governance remains incomplete unless it mandates internationally interoperable, tamper-proof watermarking at the model's architectural core, ensuring even modified iterations carry their origin's forensic signature.[110]

3.2.2. *High-risk classification and resilience*

The threat posed by deepfakes aligns directly with several high-risk operational domains defined in Annex III of the AIA, necessitating a high level of technical resilience under Article 15. This classification structurally reinforces the GDPR's "strict necessity" test, compelling enhanced compliance across sectors by explicitly linking risk magnitude to mandatory mitigation strategies. Under Article 55(1)(a), the AIA requires providers of general-purpose AI (GPAI) models presenting systemic risk to proactively mitigate inherent vulnerabilities—specifically fraudulent deepfakes—through rigorous safeguards, including conducting and documenting adversarial testing.

The severity of these risks is evidenced across multiple high-stakes environments. Remote biometric identification systems (Annex III, 1a) now necessitate the simultaneous deployment of Presentation Attack Detection (liveness verification) and Injection Attack Detection (digital fraud checks). Such measures are critical given that deepfake-driven attempts to bypass biometric authentication surged 704% in 2023,[111] with virtual camera attacks emerging as a dominant threat, increasing 2665%.[112] The fragility of current safeguards was highlighted when Reality Defender's red team successfully breached OpenAI's Sora 2 identity verification protocols within twenty-four hours using real-time deepfakes.[113] Similar vulnerabilities extend to the financial sector (Annex III, 5b), which currently bears 40% of all synthetic identity attacks[114] and requires stringent risk management under Article 9. This is further quantified by the observation of deepfake attempts occurring at five-minute intervals—coinciding with a 244% increase in digital document forgeries[115]—leading to quarterly losses exceeding $200 million in early 2025.[116] Furthermore, AI systems utilized in law enforcement (Annex III, 6c, 6e) must demonstrate rigorous accuracy validation under Article 15 to preserve the integrity of evidence against the influx of synthetic criminal media.

Beyond these specific use cases, the AIA imposes a distinct governance regime upon the technology's foundational sources. As noted by Gstrein, Haleem, and Zwitter, the Act represents a paradigm shift from reactive regulation toward "proactive AI governance," structurally enforcing safety standards before models are permitted to enter the market.[117] This operationalises through a bifurcated framework for GPAI. While standard models face transparency duties under Article 53—such as the documentation of training datasets—models characterized by "high-impact capabilities" are designated as presenting "systemic risk" under Article 51. A computational threshold of 10^{25} floating

[102] European Commission, Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act), COM(2025) 884 final, para 145.

[103] OpenAI, Sora 2. https://openai.com/sora/, 2025 [Accessed 5 July 2026].

[104] B. Colman, How We Bypassed Sora 2's Identity Safeguards in Under 24 Hours', Reality Defender. https://www.realitydefender.com/insights/sora-2-identity-bypass, 2025 [Accessed 5 July 2026].

[105] M. Gault, Sora 2 Watermark Removers Flood the Web, 404 Media. https://www.404media.co/sora-2-watermark-removers-flood-the-web/, 2025 [Accessed 5 July 2026].

[106] INTERPOL, INTERPOL report warns of increasingly sophisticated global financial fraud threat. https://www.interpol.int/en/News-and-Events/News/2026/INTERPOL-report-warns-of-increasingly-sophisticated-global-financial-fraud-threat, 2026 [Accessed 5 July 2026].

[107] European Commission, Commission investigates Grok and X's recommender systems under the Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/ip_26_203, 2026 [Accessed 5 July 2026]; Ofcom, Ofcom launches investigation into X over Grok sexualised imagery. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/ofcom-launches-investigation-into-x-over-grok-sexualised-imagery, 2026 [Accessed 5 July 2026]; Information Commissioner's Office, ICO announces investigation into Grok. https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2026/02/ico-announces-investigation-into-grok/, 2026 [Accessed 5 July 2026].

[108] Open Sora 2: FREE AI Video Generator, Open Sora 2. https://opensora2.com/, 2025 [Accessed 5 July 2026].

[109] W. Hawkins, C. Russell, B. Mittelstadt, Deepfakes on Demand: the rise of accessible non-consensual deepfake image generators, arXiv:2505.03859 [cs.CY], 2025. https://doi.org/10.48550/arXiv.2505.03859.

[110] 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.

[111] iProov, New Threat Intelligence Report Exposes the Impact of Generative AI on Remote Identity Verification. https://www.iproov.com/press/new-threat-intelligence-report-exposes-impact-generative-ai-remote-identity-verification, 2024 [Accessed 5 July 2026].

[112] iProov, iProov Issues Annual Identity Verification Threat Intelligence Report. https://www.iproov.com/press/annual-identity-verification-threat-intelligence-report, 2025 [Accessed 5 July 2026].

[113] B. Colman, How We Bypassed Sora 2's Identity Safeguards in Under 24 Hours, Reality Defender. https://www.realitydefender.com/insights/how-we-bypassed-sora-2-identity-safeguards, 2025 [Accessed 5 July 2026].

[114] T. Guillen, Deepfake Technology is Evolving: How Can AI Combat Deepfake Fraud in Financial Systems?, Signicat. https://www.signicat.com/blog/deepfake-technology-evolving-in-financial-services, 2025 [Accessed 5 July 2026].

[115] Entrust Cybersecurity Institute, 2025 Identity Fraud Report. https://www.entrust.com/sites/default/files/documentation/reports/2025-identity-fraud-report.pdf, 2025 [Accessed 5 July 2026].

[116] B. Colman, Why detecting dangerous AI is key to keeping trust alive in the deepfake era, World Economic Forum. https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/, 2025 [Accessed 5 July 2026].

[117] O.J. Gstrein, N. Haleem, A. Zwitter, General-purpose AI regulation and the European Union AI Act, Internet Policy Rev. 13 (2024). https://doi.org/10.14763/2024.3.1790.

point operations (FLOPs) triggers this status.[118]

While Article 50 attempts to mitigate deepfake risks through mandatory content labelling,[119] the systemic risk designation activates the far more stringent obligations of Article 55,[120] requiring exhaustive adversarial testing and robust cybersecurity protections against sophisticated exploits like prompt injection.

However, relying on a static computational threshold (10^{25} FLOPs) establishes a capacity-first regulatory logic, predicated on assuming higher compute power correlates linearly with increased risk. Although the AIA empowers the Commission to update these benchmarks responding to algorithmic advancements,[121] this reactionary mechanism may lack the agility to capture highly efficient models achieving high-impact capabilities through optimized architecture rather than raw scale. The 2026 Grok incident[122] provides definitive justification for the systemic risk designation under Article 51(1)(b) of the AIA. Whereas Article 51(2) establishes a quantitative presumption based on computational thresholds (10^{25} FLOPs), the Commission's qualitative power to grounds itself in the Article 3(65) systemic risk definition. This encompasses models creating a "significant impact on the Union market due to their reach" or those with "actual or reasonably foreseeable negative effects" on fundamental rights that propagate "at scale across the value chain". By linking the non-consensual sexual imagery (NCII) failure to these criteria, the Commission can utilize Article 51(1)(b) as a vital regulatory fail-safe for models demonstrating a fundamental lack of security by design, regardless of their raw compute scale. As cautioned by the Confederation of European Data Protection Organisations (CEDPO), this regime's efficacy depends on shifting focus to functional threat profiles with sufficient velocity to outpace innovation.[123]

Regarding the legality of training inputs—specifically the enforceability of Text and Data Mining (TDM) opt-outs—the industry awaits clarification in the CJEU case *Like Company v Google Ireland*.[124] Simultaneously, output controls are already undergoing codification. The General-Purpose AI Code of Practice requires providers to implement safeguards against the generation of copyright-infringing content.[125] Significantly, this output control functions as an auxiliary anti-deepfake mechanism; by forcing models to inhibit the exact mimicry of proprietary source material, regulators indirectly restrict cloning celebrity voices or likenesses for fraudulent purposes.

3.2.3. Technical defense, accuracy, and fairness

Deploying Presentation Attack Detection (PAD)—the current state-of-the-art in deepfake prevention—increasingly satisfies the GDPR's strict necessity test when integrated with obligatory zero-retention and single-use verification protocols. PAD methodologies, encompassing both facial and vocal liveness checks,[126] represent a complex multimodal synthesis achieving correctness rates of 94%–96% in optimized settings.[127] International benchmarks, such as ISO/IEC 30107–3, validate this technical performance, which regulatory certifications like those issued by the BSI further verify.[128]

The convergence of PAD and Privacy-Enhancing Technologies (PETs) provides an aligned multi-layered protection blueprint. While PAD verifies liveness at the point of capture, decentralized architectures secure the subsequent processing. By either fragmenting templates into mathematically inert shards via Secure Multi-Party Computation (sMPC)[129] or utilizing Homomorphic Encryption (HE) to perform matching on third-party servers without decryption,[130] these architectures ensure biometric data is never exposed in plaintext form while in transit, at rest, or in use.[131] Solutions like Keyless operationalize this via a Zero-Knowledge Biometrics (ZKB) method, ensuring no complete biometric image ever exists to be intercepted. By eliminating the single point of failure inherent in centralized databases,[132] this approach satisfies the Article 32 GDPR security by design requirements emphasized in *Russmedia*,[133] stipulating the legal mandate for controllers to implement proactive measures against unauthorized data extraction.

Despite these technical milestones, deepfake evolution continues exploiting fundamental limitations in model generalization, specifically regarding overfitting and underfitting phenomena. Overfitting occurs when a detection model learns the idiosyncratic details of specific historical attacks too granularly, resulting in false negatives when confronting novel variations; conversely, underfitting occurs when a model

[118] Ibid.

[119] F. Romero-Moreno, Generative AI and deepfakes: a human rights approach to tackling harmful content, Int. Rev. Law Comput. Technol. 38 (2024) 297–326. https://doi.org/10.1080/13600869.2024.2324540

[120] O.J. Gstrein, N. Haleem, A. Zwitter, General-purpose AI regulation and the European Union AI Act, Internet Policy Rev. 13 (2024). https://doi.org/10.14763/2024.3.1790.

[121] M. Moloney, J. Browne, What's a FLOP? How the AI Act Regulates General Purpose AI Systems, Confederation of European Data Protection Organisations (CEDPO). https://cedpo.eu/wp-content/uploads/How-General-Purpose-AI-Models-are-regulated-under-the-AI.pdf, 2024 [Accessed 5 July 2026].

[122] European Commission, Commission investigates Grok and X's recommender systems under the Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/ip_26_203, 2026 [Accessed 5 July 2026]; Ofcom, Ofcom launches investigation into X over Grok sexualised imagery. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/ofcom-launches-investigation-into-x-over-grok-sexualised-imagery, 2026 [Accessed 5 July 2026]; Information Commissioner's Office, ICO announces investigation into Grok. https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2026/02/ico-announces-investigation-into-grok/, 2026 [Accessed 5 July 2026].

[123] M. Moloney, J. Browne, What's a FLOP? How the AI Act Regulates General Purpose AI Systems, Confederation of European Data Protection Organisations (CEDPO). https://cedpo.eu/wp-content/uploads/How-General-Purpose-AI-Models-are-regulated-under-the-AI.pdf, 2024 [Accessed 5 July 2026].

[124] Case C-250/25, *Like Company v Google Ireland Limited*, OJ C, C/2025/3039.

[125] European Commission, The General-Purpose AI Code of Practice, Copyright Chapter. Measure 1.4. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai, 2025 [Accessed 5 July 2026].

[126] A. Terekhin, Presentation Attacks: What Liveness Detection Systems Protect From Every Day, Regula Forensics. https://regulaforensics.com/blog/presentation-attack-detection/, 2025 [Accessed 5 July 2026]; G. Tahaoglu, D. Baracchi, D. Shullani, M. Iuliani, A. Piva, Deepfake audio detection with spectral features and ResNeXt-based architecture

[127] B. Colman, Why detecting dangerous AI is key to keeping trust alive in the deepfake era, World Economic Forum. https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/, 2025 [Accessed 5 July 2026]; K. Thakur, M.A. Sayed, S.A. Tisha, M.K. Alam, M.T. Hasan, et al., Multimodal Deepfake Detection Using Transformer-Based Large Language Models: A Path Toward Secure Media and Clinical Integrity, Am. J. Eng. Technol. 7 (2025) 169–177. https://doi.org/10.37547/tajet/Volume07Issue05-16.

[128] Busch, C. (2018, January 26). *Biometrics and Presentation Attack Detection* [Presentation]. Swedish Biometrics Forum, Halmstad University, Halmstad, Sweden; iProov, iProov is First Biometrics Vendor to Demonstrate Deepfake Resilience, Under New NIST Digital Identity Requirements. https://www.iproov.com/press/nist-digital-identity-requirements-first-biometrics-vendor-demonstrating-deepfake-resilience, 2025 [Accessed 5 July 2026].

[129] Keyless, Keyless Protocol. https://docs.keyless.io/why-keyless/keyless-protocol, 2025 [Accessed 5 July 2026].

[130] A. Vemury, Third party biometric homomorphic encryption matching for privacy protection. US Patent US20230403132A1, 2023. https://patents.google.com/patent/US20230403132A1 [Accessed 5 July 2026].

[131] Keyless, Keyless Protocol. https://docs.keyless.io/why-keyless/keyless-protocol, 2025 [Accessed 5 July 2026]; A. Vemury, Third party biometric homomorphic encryption matching for privacy protection. US Patent US20230403132A1, 2023. https://patents.google.com/patent/US20230403132A1 [Accessed 5 July 2026].

[132] Keyless, Keyless Protocol. https://docs.keyless.io/why-keyless/keyless-protocol, 2025 [Accessed 5 July 2026].

[133] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [118], [121]–[122], [126].

is too simplistic to distinguish genuine biometrics from sophisticated forgeries, leading to false positives.[134] This lack of domain-robustness frequently results in aggregate failures when moving from controlled laboratory data to high-compression, real-world environments.[135] Reality Defender's red teaming demonstrated such a vulnerability, successfully breaching authentication systems by deploying a cloned executive voice within a low-latency VoIP environment—illustrating the potent fusion of technical manipulation and human factors.[136]

Furthermore, implementing PAD technology triggers a profound compliance conflict between the GDPR's accuracy principle (Article 5(1)(d)) and the AIA's fairness mandate (Article 10), a duty underscored by the CJEU *SCHUFA*[137] and *Dun & Bradstreet*[138] precedents. Deepfake detectors frequently fail this dual standard due to quantifiable algorithmic bias. Performance divergences between racial subgroups, particularly African and Asian populations, reveal error rate gaps as high as 10.7%.[139] This failure manifests in a disproportionately high false positive rate for female subjects and individuals with darker skin tones, resulting in discriminatory inaccuracy systems incorrectly flag genuine biometric submissions as forgeries.[140] As the Luxembourg Court specified in *SCHUFA*,[141] the controller is under a strict obligation to use "appropriate mathematical or statistical procedures" for profiling, ensuring minimised error risks and precluding "discriminatory effects" based on racial or ethnic origin (Recital 71 GDPR). A model exhibiting a 10.7% racial disparity fails this appropriateness standard, rendering the processing inherently unlawful.

Significant lab-to-live performance variances compound this pervasive bias. While providers often claim high accuracy based on controlled datasets, empirical evidence from the 2026 DSIT report suggests that correctness rates typically drop 10%–20% during real-world redeployment. This lack of domain-robustness stems from limited access to high-quality training datasets that lack real-world synthetic media or updated creation technologies.[142] The CJEU's determination in *Dun & Bradstreet*,[143] directly addresses such technical uncertainty, articulating that the "meaningful information" required under Article 15(1)(h) GDPR covers all relevant information concerning the "procedure and principles actually applied" to obtain a specific result. This technological reality necessitates a robust structure for procedural transparency. Automated service denials based on failed liveness checks trigger the right to human review under Article 22 and transparency obligations under Article 13. Following the precedents of *SCHUFA*[144] and *Dun & Bradstreet*,[145] individuals flagged by deepfake detection systems must receive a "genuine right to an explanation" regarding the underlying logic—including specific automated forensic principles like micro-expression analysis—to effectively contest false positives. As Wachter observes, this remains essential because machine learning black boxes' opacity can inadvertently reinforce systemic biases, creating significant barriers to effective legal redress.[146]

To resolve the inherent deadlock between verification accuracy and data minimization, the AIA and the recently adopted Digital Omnibus Regulation establish a specialized legal architecture for processing sensitive data. Article 10(3) AIA mandates that training datasets be "sufficiently representative," while the 2026 DSIT report emphasizes standardized evaluation metrics are now prerequisites for commercial trust and regulatory compliance.[147] To operationalize this, the Digital Omnibus package introduces a new Article 4a into the AIA, which expands upon the original bias-mitigation framework of Article 10(5). This provision establishes a clear "substantial public interest" legal basis alongside Article 9 of the GDPR, authorizing the processing of special categories of personal data when necessary for bias detection and correction across all AI systems and models[148]—a deliberate lowering of the original "strictly necessary" threshold that has drawn sharp criticism from European data protection authorities.[149] This legislative architecture empowers developers to utilize synthetic biometric data to measure and correct false-positive disparities[150] without infringing upon the GDPR's general prohibition on biometric processing. Through the harmonization of accuracy metrics and the utilization of diverse, updated datasets, the industry can develop algorithms capable of detecting the evolving nuances of generative forgeries[151] while fulfilling the quality management requirements of the AIA.

This strategy constitutes more than a policy preference; the "least intrusive means" doctrine crystalized in *Mousse v CNIL*[152] legally necessitates it. The CJEU observed that processing fails the necessity

[134] F. Romero-Moreno, Deepfake detection in generative AI: A legal framework proposal to protect human rights, Comput. Law Secur. Rev. 58 (2025) 106162. https://doi.org/10.1016/j.clsr.2025.106162.

[135] S. Anlen, R. Vázquez Llorente, Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail, Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024 [Accessed 5 July 2026].

[136] D. Khambholia, Inside Reality Defender's AI Red Team, Reality Defender. https://www.realitydefender.com/insights/ai-red-teaming, 2025 [Accessed 5 July 2026].

[137] Case C-634/21, *OQ v Land Hessen (SCHUFA)*, ECLI:EU:C:2023:957, [3], [5], [59], [66].

[138] Case C-203/22, *CK v Magistrat der Stadt Wien (Dun & Bradstreet),* ECLI:EU:C:2025:117, [53], [63], [65], [68], [72], [74].

[139] L. Trinh, Y. Liu, An Examination of Fairness of AI Models for Deepfake Detection, in: Proc. 30th Int. Jt. Conf. Artif. Intell. (IJCAI-21), 2021, pp. 1–7. https://doi.org/10.48550/arXiv.2105.00558.

[140] C. Peng, Y. Chen, D. Liu, N. Wang, X. Gao, FairForensics: mitigating attribute bias in deepfake detection by integrating texture and attribute features, Neural Netw. 192 (2025) 107899. https://doi.org/10.1016/j.neunet.2025.107899.

[141] Case C-634/21, *OQ v Land Hessen (SCHUFA)*, ECLI:EU:C:2023:957, [66].

[142] Department for Science, Innovation & Technology, Deepfake detection technology. https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].

[143] Case C-203/22, *CK v Magistrat der Stadt Wien (Dun & Bradstreet),* ECLI:EU:C:2025:117, [50], [66].

[144] Case C-634/21, *OQ v Land Hessen (SCHUFA)*, ECLI:EU:C:2023:957, [56], [63], [66].

[145] Case C-203/22, *CK v Magistrat der Stadt Wien (Dun & Bradstreet),* ECLI:EU:C:2025:117, [43], [55]–[58], [61].

[146] S. Wachter, Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR, Comput. Law Secur. Rev. 34 (2018) 436–449. https://doi.org/10.1016/j.clsr.2018.02.002.

[147] Department for Science, Innovation & Technology, Deepfake detection technology. https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].

[148] European Commission, Proposal for a Regulation amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI), COM(2025) 836 final; European Parliamentary Research Service, Algorithmic discrimination under the AI Act and the GDPR, https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/769509/EPRS_ATA(2025)769509_EN.pdf, 2025 [Accessed 5 July 2026].

[149] N. Moreno, C. Melton, The 2025 European Commission EU digital omnibus package: The EU AI Act. https://www.kennedyslaw.com/en/thought-leadership/article/2026/the-2025-european-commission-eu-digital-omnibus-package-the-eu-ai-act/, 2026 [Accessed 5 July 2026].

[150] G. Sharma, Synthetic Data for Biometrics Training: Addressing Bias without Privacy Risks, Biometric Update. https://www.biometricupdate.com/202507/synthetic-data-for-biometrics-training-addressing-bias-without-privacy-risks, 2025 [Accessed 5 July 2026]; L. Trinh, Y. Liu, An Examination of Fairness of AI Models for Deepfake Detection, in: Proc. 30th Int. Jt. Conf. Artif. Intell. (IJCAI-21), 2021, pp. 1–7. https://doi.org/10.48550/arXiv.2105.00558.

[151] Department for Science, Innovation & Technology, Deepfake detection technology. https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].

[152] Case C-394/23, *Mousse v Commission nationale de l'informatique et des libertés (CNIL),* ECLI:EU:C:2025:2, [28].

requirement of Article 6(1) GDPR if the objective "could reasonably be achieved just as effectively by other means less restrictive of the fundamental rights".[153] Applying this standard to PETs and synthetic datasets—identifying them as "workable and less intrusive" alternatives[154]—effectively precludes relying on the necessity derogation to justify large-scale retention of raw biometric assets. Delineating this further, the Court's explicit rejection of "systematic and generalised processing" as contrary to the data minimisation principle (Article 5(1)(c))[155] mandates transitioning toward privacy-preserving training protocols. Consequently, rather than relying on the mass accumulation of sensitive biometric data, the legal framework now necessitates a structural shift: cryptographic assurance and synthetic modelling must serve as the primary mechanisms for achieving functional fairness.

3.3. Resolving the privacy-deepfake paradox: PETs as the architecture of trust

While cryptographic assurance enforces algorithmic fairness, it must also resolve the biometric paradox: a state where the sensitive data required for identity verification becomes the primary target for fraudulent exploitation. To meet this challenge, Privacy-Enhancing Technologies (PETs) must transition from theoretical safeguards to the foundational architecture of digital trust.[156] As Woods notes regarding the landmark in *Russmedia*,[157] the Grand Chamber fundamentally shifted the protection burden from reactive moderation to proactive technical intervention, grounding this mandate in the core accountability principles of the GDPR.[158] Consequently, Article 5(2) and Article 24 GDPR require operators to identify sensitive data prior to publication.[159] This protection extends even to synthetic data, as the CJEU remarked that the "untrue and harmful nature" of claims does not strip them of their "sensitive" status under Article 9(1).[160] To satisfy the explicit consent requirements of Article 9(2)(a) GDPR, controllers must verify an advertiser's identity to prevent unauthorized processing.[161]

Platforms like 1Kosmos (BlockID) operationalize this through decentralized architectures where identity data remains under the user's exclusive control.[162] By utilizing Zero-Knowledge Proofs (ZKP), the platform verifies the data owner's legitimacy without storing raw personally identifiable information (PII).[163] This effectively negates the distinction between "having" and "using" data— resolving the Kindt loophole technically—as the system utilizes the verification result without holding the underlying asset.[164]

As identity verification matures into 2026, a critical human authorization gap has emerged in agentic AI: the need to verify which human authorized a specific AI action.[165] ZeroBiometrics' ZeroSentinel addresses this via a "cryptographic authorization control plane" utilizing ZeroFace technology to reconstruct ephemeral ECC keypairs from facial entropy at the moment of signing, ensuring the human is the private key.[166] This architecture satisfies the proactive identification duties of *Russmedia*[167] to prevent "loss of control" via the ZeroGrant protocol, which embeds authorization policy into machine-readable OID extensions within a non-repudiable ZeroIntent enforcement gateway, preventing unauthorized privilege escalation.[168]

Simultaneously, InterLink Labs extends this to the AI lifecycle through Federated Learning (FL), ensuring raw biometric data never leaves the user's hardware[169] while fulfilling the proactive identification mandate of *Russmedia*[170] judgment. This is achieved by aggregating local model updates rather than raw data, allowing the global model to identify new fraud patterns without centralizing sensitive information.[171]

On-device solutions like Microblink further eliminate cloud-based vulnerabilities through localized processing,[172] satisfying Article 25 GDPR "by design" mandates. This immediate, localized identification also fulfils the U.S. Department of Homeland Security (DHS) benchmark for high-performing systems[173] and prevents the "loss of control" warned of by the Court.[174] Anchoring this in physical reality, iProov's Dynamic

[153] Ibid.
[154] Ibid, [40].
[155] Ibid, [42].
[156] Information Commissioner's Office, Using privacy-enhancing technologies (PETs) to unlock value from data responsibly. https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/11/using-privacy-enhancing-technologies-pets-to-unlock-value-from-data-responsibly/, 2024 [Accessed 5 July 2026]; European Data Protection Supervisor, Secure multi-party computation: powering privacy through collaboration. https://www.edps.europa.eu/press-publications/press-news/blog/secure-multi-party-computation-powering-privacy-through-collaboration_en, 2025 [Accessed 5 July 2026]; European Data Protection Supervisor, TechDispatch #1/2025 - Federated Learning. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-06-10-techdispatch-12025-federated-learning_en, 2025 [Accessed 5 July 2026].
[157] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935.
[158] Woods, L., Image Rights and False Claims, Data Protection and Intermediary Immunity: the case of Russmedia. https://eulawanalysis.blogspot.com/2025/12/image-rights-and-false-claims-data.html, 2025 [Accessed 5 July 2026].
[159] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [88], [97] [106].
[160] Ibid, [53].
[161] Ibid, [99], [104].
[162] 1Kosmos, 1Kosmos: Identity Verification & Passwordless Solutions. https://www.1kosmos.com/, 2026 [Accessed 5 July 2026].
[163] P. Phillips, Identity Security in MedTech: Lessons from Recent Healthcare Breaches. 1Kosmos, https://www.1kosmos.com/resources/blog/identity-security-in-medtech, 2026 [Accessed 5 July 2026].
[164] E.J. Kindt, Having yes, using no? About the new legal regime for biometric data, Comput. Law Secur. Rev. 34 (2018) 523–538. https://doi.org/10.1016/j.clsr.2017.11.002.
[165] C. Burt, ZeroBiometrics details how biometric AI agent authorization suite works, Biometric Update. https://www.biometricupdate.com/202604/zerobiometrics-details-how-biometric-ai-agent-authorization-suite-works, 2026 [Accessed 5 July 2026].
[166] ZeroBiometrics, ZeroSentinel™ Technical White Paper: A Standards-Aligned Cryptographic Human-Authorization Layer for Agentic AI. https://www.biometricupdate.com/wp-content/uploads/2026/04/ZeroSentinel-Technical-White-Paper.pdf, 2026 [Accessed 5 July 2026].
[167] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [106], [121].
[168] ZeroBiometrics, ZeroSentinel™ Technical White Paper: A Standards-Aligned Cryptographic Human-Authorization Layer for Agentic AI. https://www.biometricupdate.com/wp-content/uploads/2026/04/ZeroSentinel-Technical-White-Paper.pdf, 2026 [Accessed 5 July 2026].
[169] InterLink Network, Technical Implementation: Federated Learning. https://whitepaper.interlinklabs.ai/technical-implementation/federated-learning, 2025 [Accessed 5 July 2026].
[170] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [106].
[171] InterLink Network, Technical Implementation: Federated Learning. https://whitepaper.interlinklabs.ai/technical-implementation/federated-learning, 2025 [Accessed 5 July 2026].
[172] Microblink, Top Deepfake Detection Software & AI Fraud Solutions for 2026. https://microblink.com/resources/blog/best-deepfake-detection-software/, 2026 [Accessed 5 July 2026].
[173] Microblink, Microblink Identified as Only Vendor to Meet All Performance Thresholds in U.S. Department of Homeland Security Identity Verification Evaluation. https://microblink.com/about/newsroom/microblink-identified-as-only-vendor-to-meet-all-performance-thresholds-in-u-s-department-of-homeland-security-identity-verification-evaluation/, 2026 [Accessed 5 July 2026].
[174] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [121].

Liveness uses patented colour-flashing to verify 3D human presence,[175] providing independently validated defence (CEN/TS 18099 High) against virtual camera injection.[176] This also addresses the primary risks of identity theft identified in *Russmedia*.[177]

Ultimately, these technologies decouple identity verification from data possession. The 2026 EU Digital Identity Wallet Architecture and Reference Framework (ARF) utilizes Zero-Knowledge Proofs (ZKP) and Selective Disclosure to operationalize this separation,[178] allowing SMEs to consume pre-verified credentials directly. This satisfies privacy-by-design mandates while offloading the regulatory liability of storing sensitive personal records.[179] Finally, Sensity AI's forensic assessment delivers the actionable audit trails required to fulfil the proactive demonstration of due diligence mandated by *Russmedia*.[180] By layering independent forensic signals over cryptographic protocols, this framework secures the AI lifecycle through the verifiable assurance of human origin rather than the risky accumulation of raw biometric data.[181]

The *Russmedia* Compliance Matrix

Legal Requirement	Russmedia Finding	Technical PET / Agentic Solution
Proactive Identification	**Para [97]:** Requirement to identify and categorize sensitive data *ex-ante* (prior to publication).	**InterLink Labs:** Federated Learning (FL) for local pattern detection without PII centralization.
Identity Verification	**Para [99], [106]:** Strict obligation to verify uploader identity to validate Art 9(2)(a) explicit consent.	**ZeroSentinel:** Biometrically-anchored X.509 certificates & RFC 8705 sender-constrained tokens.
Security of Processing	**Para [121], [122]:** Implementation of measures to prevent the "loss of control" of sensitive assets.	**Keyless:** Zero-Knowledge Biometrics (ZKB) via sMPC to eliminate centralized database risks.
Access & Scope Control	**Para [121]:** Preventing unauthorized secondary use or privilege escalation.	**ZeroIntent:** Protocol-layer enforcement utilizing policy-embedded **ZeroGrants**.
Injection & Extraction Defense	**Para [126]:** Statutory mandate to prevent unauthorized data extraction and redistribution.	**iProov / Microblink:** Dynamic Liveness & Injection Attack Detection (CEN/TS 18,099 High).
Accountability & Audit	**Para [136]:** Due diligence required to maintain accountability and intermediary immunity.	**Sensity AI:** Forensic-grade assessment providing "actionable" audit trails and admissible evidence.

4. Layer 2: distribution control and the fragility of transparency

Layer 2 traces the transition from individual rights to distributor responsibilities, contrasting the EU's rights-based archetype (*Russmedia*)[182] with UK technology-forcing imperatives.[183] The CJEU precedent-setting judgment reclassifies deepfake distribution as a primary data protection imperative rather than mere content moderation.[184] By establishing "joint controllership," the Court pierces the DSA liability shield; Article 5 GDPR now operates independently of passive host status under Article 6 DSA.[185] This confirms data protection as a *lex superior*, preventing platforms from invoking intermediary exemptions to circumvent fundamental rights.[186]

4.1. The mandate: transparency, technical fragility, and the verification requirement

The EU AIA Article 50 and DSA Article 35(1)(k) mandate a two-pronged protective apparatus requiring both human disclosure and technical marking. With penalties reaching 6% of global annual turnover under the DSA (Article 74(1)) and 7% for prohibited practices under the AIA (Article 99(3)), platform compliance stakes are existential. Specifically, AIA Article 50(1) targets the point of user interaction, requiring explicit notification when users encounter AI-generated content. This stipulation aligns with the CJEU "average consumer" standard enunciated in *Speciality Drinks*[187] to combat customer impersonation—a pivotal intervention given that deepfake-driven contact centre scams escalated 680% in 2024.[188]

However, the *Russmedia*[189] ruling clarifies that transparency does not negate the duty of verification; even with AI-generation disclosures, platforms remain joint controllers under Article 26 GDPR. Consequently, they face legal exposure for Article 9(2)(a) GDPR breaches if they fail to validate an uploader's identity to ensure explicit consent.[190] This requirement complements technical specifications under AIA Articles 50(2) and 50(4), which impose machine-readability and disclosure duties on providers and deployers—often via watermarking (Recital 133)—to mitigate LLM-generated phishing and deceptive testimonials.[191]

Despite a November 2026 deadline for mandatory watermarking,[192] a March 2026 Indicator audit reveals a critical implementation gap: labelling remains a "slot machine" of inconsistent technical execution.[193] Instagram correctly labelled only 14% of synthetic images, while TikTok

[175] iProov, Dynamic Liveness. https://www.iproov.com/biometric-encyclopedia/genuine-presence-assurance, 2026 [Accessed 5 July 2026].
[176] iProov, iProov Dynamic Liveness Is the First and Only Solution to Achieve CEN/TS 18099 High and Ingenium Level 4 for Injection Attack Detection. https://www.iproov.com/press/dynamic-liveness-first-achieving-cen-ts-18099-high-ingenium-level-4-injection-attack-detection, 2026 [Accessed 5 July 2026].
[177] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [103].
[178] European Commission, eIDAS Architecture and Reference Framework (ARF). https://github.com/eu-digital-identity-wallet/architecture-and-reference-framework, 2024 [Accessed 5 July 2026].
[179] Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity, OJ L, 2024/1183, 30.4.2024.
[180] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [88], [104], [106].
[181] Sensity AI, All-in-one Deepfake Detection. https://sensity.ai/, 2026 [Accessed 5 July 2026].
[182] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935.
[183] Online Safety Act 2023 (UK), c. 50, ss 10, 38, 136, 231.
[184] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [136].
[185] Ibid, [67], [93], [106].
[186] Ibid, [131], [135]–[136].
[187] Case T-250/15, *Speciality Drinks Ltd v European Union Intellectual Property Office – William Grant (CLAN)*, ECLI:EU:T:2016:678, [26].
[188] Sensfrx, The Ecommerce Battlefield: Mapping Key Vulnerabilities to Deepfake Scams. https://blog.sensfrx.ai/ecommerce-fake-reviews-deepfake-fraud/, 2025 [Accessed 5 July 2026].
[189] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [92], [97].
[190] Ibid, [99], [106].
[191] Sensfrx, The Ecommerce Battlefield: Mapping Key Vulnerabilities to Deepfake Scams. https://blog.sensfrx.ai/ecommerce-fake-reviews-deepfake-fraud/, 2025 [Accessed 5 July 2026].
[192] European Parliament, Artificial Intelligence Act: delayed application, ban on nudifier apps. https://www.europarl.europa.eu/news/en/press-room/20260323IPR38829/artificial-intelligence-act-delayed-application-ban-on-nudifier-apps, 2026 [Accessed 5 July 2026].
[193] Indicator, AI labeling is still very much a work in progress. https://indicator.media/p/ai-labeling-is-still-very-much-a-work-in-progress, 2026 [Accessed 5 July 2026].

and YouTube lapsed on approximately 66% and 50% respectively.[194]

European Commission guidelines warn that such non-labelling may constitute a deceptive technique under Article 5(1)(a).[195] To mitigate this, the Code of Practice on Transparency of AI-Generated Content establishes a framework for marking and detection; however, it explicitly acknowledges that no single transparency technique currently possesses the necessary robustness against malicious behaviour, such as removal and circumvention attacks.[196] This vulnerability underscores how architectural instability fundamentally limits the efficacy of relying solely on transparency mechanisms. While provenance protocols like the Coalition for Content Provenance and Authenticity (C2PA) successfully attest a "chain of trust," they remain passive standards rather than active detection tools.[197] As demonstrated by Yu and Wei, "adversarial laundering" exploits this fragility; utilizing provenance hijacking, threat actors simulate device-specific sensor noise and forge cryptographic manifests, rendering current transparency workflows inherently insecure against sophisticated attacks.[198]

To bridge this infrastructure gap in live environments—such as WhatsApp or VoIP—third-party forensic suites like UncovAI (2026) introduced external verification bots and real-time "Trust Scores" to combat "CEO fraud."[199] Addressing the subsequent explainability void, emerging forensic architectures integrate Gemini 3 Flash as a high-velocity engine to translate complex detection artifacts into natural language. By performing multimodal analysis 4x faster than legacy models, this integration ensures high-fidelity explainability within the sub-second response times imperative for financial voice-agent validation.[200]

Ultimately, technical weakness renders the current transparency shield legally insufficient. While Recital 133 of the AIA demands "robust" solutions, the CJEU in *SRB v EDPS*[201] requires measures to be "durable" to satisfy fundamental rights protections. Since simple transcoding easily circumvents spatial watermarking,[202] it fails the "effective mitigation" standard mandated by DSA Article 35(1).

Platforms now face computational asymmetry—a state where synthetic generation outpaces detection capacity.[203] Under the *Russmedia*[204] verify-then-publish doctrine, deploying generative architectures that exceed a platform's capacity to detect, or that lack cross-platform interoperability, constitutes a *per se* violation of data protection by design (Article 25 GDPR). While Article 6 DSA may shield hosts from specific content liability, the Court's characterization of platforms as architects of the dissemination parameters ensures they remain strictly liable under the GDPR. Within this framework, failing to integrate interoperable, durable marking constitutes an architectural flaw facilitating a structural "loss of control" over personal data.[205]

Crucially, this shift avoids a prohibited "general monitoring" obligation under Article 8 DSA and CJEU precedent (*SABAM*[206] rulings), acting instead as a structural handshake protocol for architectural integrity. As envisioned in DSA Recital 30 (monitoring "in a specific case"), and consistent with the Court's reasoning in *Poland v Parliament and Council*,[207] automated validation of specific technical markers—such as the latent-space signatures identified by the SynthID Detector[208]—represents a targeted measure to ensure compliance with AIA Article 50. This is not a prior restraint on speech, but a targeted check for "identical" digital signals permitted under the *Glawischnig-Piesczek*[209] doctrine. Under the *Russmedia*[210] framework, technical instability does not excuse the platform; it triggers a mandatory security duty under Article 32 GDPR to prevent a "loss of control" over personal data.

4.2. Systemic risk management and the erosion of the intermediary shield

The contemporary regulatory landscape marks a definitive departure from the era of passive intermediary liability. Central to this transition is the EU Digital Services Act (DSA), which imposes the Union's most rigorous compliance burdens on Very Large Online Platforms (VLOPs) and Very Large Search Engines (VLOSEs) to counter systemic risks under Articles 34 and 35. To prevent fragmented oversight from diluting these duties, the proposed Digital Omnibus Regulation seeks to bridge regulatory silos by amending the Network and Information Systems Directive (NIS2)[211] to prescribe a Single-Entry Point (draft Article 23a). By centralising the mandatory reporting regimes of the GDPR, NIS2, and the Digital Operational Resilience Act (DORA),[212] through a unified portal, the framework prevents platforms from exploiting coordination gaps to conceal the true scale of institutional fraud.[213]

[194] Ibid.

[195] European Commission, Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act), COM(2025) 884 final, para 56.

[196] European Commission, Code of Practice on Transparency of AI-Generated Content. https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content, 2026 [Accessed 5 July 2026].

[197] Coalition for Content Provenance and Authenticity (C2PA), C2PA Implementation Guidance v2.3. https://spec.c2pa.org/specifications/specifications/2.3/guidance/Guidance.html, 2025 [Accessed 5 July 2026].

[198] J. Yu and N. Wei, The Orthogonal Vulnerabilities of Generative AI Watermarks: A Comparative Empirical Benchmark of Spatial and Latent Provenance. arXiv:2603.10323, 2026. https://doi.org/10.48550/arXiv.2603.10323

[199] UncovAI, Verifying Reality: Informational Tension 2026. https://uncovai.com/verifying-reality-informational-tension-2026/, 2026 [Accessed 5 July 2026].

[200] S. Dolen, Z. Ahmed, V. Dharmadhikari, Delivering real-time deepfake intelligence with Gemini 3 Flash, Google AI Blog. https://aistudio.google.com/case-studies/resembleai, 2025 [Accessed 5 July 2026].

[201] Case C-413/23 P, *European Data Protection Supervisor (EDPS) v Single Resolution Board (SRB),* ECLI:EU:C:2025:645, [75]–[76].

[202] Ofcom, Deepfake Defences 2: The Attribution Toolkit Report. https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/deepfake-defences-2/deepfake-defences-2—the-attribution-toolkit.pdf?v = 402066, 2025 [Accessed 5 July 2026].

[203] Department for Science, Innovation and Technology, Deepfake detection technology, https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].

[204] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [104], [106], [120]–[122], [126].

[205] Ibid, [72], [97], [120]–[122], [136].

[206] Case C-360/10, *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* ECLI:EU:C:2012:85, [34], [37]–[38]; Case C-70/10, *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* ECLI:EU:C:2011:771, [36], [39]–[40].

[207] Case C-401/19, *Republic of Poland v European Parliament and Council of the European Union* ECLI:EU:C:2022:297, [81], [85], [90]–[91].

[208] P. Kohli, SynthID Detector — a new portal to help identify AI-generated content, Google Blog. https://blog.google/innovation-and-ai/products/google-synthid-ai-content-detector/, 2025 [Accessed 5 July 2026]; Google DeepMind, SynthID: A tool to watermark and identify content generated through AI. https://deepmind.google/models/synthid/, 2026 [Accessed 5 July 2026].

[209] Case C-18/18, *Eva Glawischnig-Piesczek v Facebook Ireland Limited* ECLI:EU:C:2019:821, [34], [37], [46].

[210] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [120]–[122], [126].

[211] Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS2 Directive) [2022] OJ L333/80.

[212] Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011 [2022] OJ L333/1.

[213] European Commission, Proposal for a Regulation as regards the simplification of the digital legislative framework (Digital Omnibus), COM(2025) 837 final, art 6(1).

The Grand Chamber's *Russmedia*[214] seminal precedent establishes a transformative era of centralized accountability by fundamentally reclassifying hosting within the algorithmic economy. The CJEU held that when a platform determines "essential elements" of a publication—such as granular audience targeting and reach optimization—it exerts a "decisive influence" over the data flow.[215] This finding renders the neutral host defence functionally incompatible with recommender systems. By engaging in algorithmic curation, the platform performs an active "determination of the means of processing", effectively subordinating DSA liability protections to the GDPR's fundamental rights mandates.[216]

This status transforms the platform into a joint controller under Article 26 GDPR, triggering a statutory identifiability command.[217] The Court ascertains that because Article 26 requires joint controllers to determine their obligations in a clear manner, such a mandate "would prove impossible" if an uploader remains anonymous.[218] Since the platform is legally barred from delegating data protection duties to a non-identifiable entity, permitting uploader anonymity is no longer a neutral business choice. Rather, it constitutes a structural breach of the joint-controllership requirement, as the platform "is obliged" to authenticate identity as a prerequisite for lawful dissemination.[219] By maintaining anonymous architectures, platforms facilitate a "feeling of impunity" that severely undermines the principle of accountability under Article 5(2) GDPR.[220]

Under the resulting verify-then-publish standard, platforms must refuse any advertisement lacking a validated identity and documented explicit consent under Article 9(2)(a).[221] This ensures that the platform—as the architect of the dissemination parameters—remains the non-delegable guarantor of the integrity of the data it amplifies.[222] By ruling that intermediary protections "cannot interfere" with data protection duties, the Court strips platforms of the ability to delegate responsibility to the uploader, reclassifying the platform's role as a proactive warden of the digital ecosystem.[223]

Such regulatory pressure has rapidly shifted from theory to enforcement. The European Commission's €120 million fine against X for deceptive design (Article 25)[224] and preliminary findings against TikTok and Meta regarding data access (Article 40)[225] demonstrate aggressive enforcement filling the compliance void. The judiciary reinforces this trend, notably the Amsterdam District Court's decision that automatically reverting users to profiled content violates the "dark pattern" prohibitions of Article 25(1) DSA.[226] These determinations align with the *Russmedia*[227] requirement that platforms must prioritize appropriate security under Article 32 GDPR over revenue optimization.

The impact of this "decisive influence" is most acute when algorithmic targeting facilitates systemic harm. In the 2025 Swedish investment fraud, synthetic deepfake advertisements defrauded 5000 investors of €44.5 million.[228] Under the *Russmedia*[229] framework, the platform's failure to adapt advertising systems (Article 35(1)(e) DSA) transforms its influence into actionable negligence; the platform effectively determines the means by which synthetic fraud achieves the scale necessary for mass victimization.

The 2026 Grok crisis serves as the definitive case study for computational asymmetry[230]—a systemic risk under Article 34(1) where the capacity to generate synthetic harm outpaces detection resources. This represented a dual compliance failure: a design-level failure to assess specific systemic risks regarding gender-based violence and the protection of minors under Article 34(1)(d)—particularly as recommender systems influence these risks (Article 34(2)(a))—and a subsequent mitigation failure regarding the effective measures mandated by Article 35(1)(d).

The industrial volume of harmful output evidences the magnitude of this lapse: within a nine-day window, users produced 4.4 million images—41% depicting sexualized portrayals of women[231]—a catastrophic manifestation of the foreseeable negative effects on gender-based violence and mental well-being identified in Article 34(1)(d). Furthermore, a concurrent output of three million explicit images, including 23,000 instances of synthetic child sexual abuse material (CSAM),[232] confirms a foundational deficiency in addressing the protection of minors, an integral component of the Article 34 risk assessment mandate.

From a regulatory standpoint, any mitigation measure permitting such mass-production is, by definition, not "effective" under Article 35 (1). Investigations by the European Commission,[233] Ofcom[234] and the Information Commissioner's Office (ICO) substantiate preliminary findings that the platform's architecture facilitated this industrial-scale proliferation.[235]

The scrutiny culminated in the Amsterdam District Court's landmark

[214] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935.
[215] Ibid, [72]–[73].
[216] Ibid, [71], [131], [135]–[136].
[217] Ibid, [92].
[218] Ibid, [101].
[219] Ibid, [69], [102].
[220] Ibid, [84], [86], [104].
[221] Ibid, [97], [104]–[106].
[222] Ibid, [72], [121], [131].
[223] Ibid, [135].
[224] European Commission, Commission fines X (formerly Twitter) €120 million under the Digital Services Act for systemic failure to remove illegal content. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934, 2025 [Accessed 5 July 2026].
[225] European Commission, Commission preliminarily finds TikTok and Meta in breach of their transparency obligations under the Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503, 2025 [Accessed 5 July 2026].
[226] Amsterdam District Court, *Stichting Bits of Freedom v. Facebook Netherlands B.V., Meta Platforms Ireland Ltd. and Meta Platforms Inc.,* ECLI:NL:RBAMS:2025:7253, [4.33]–[4.34]. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBAMS:2025:7253, 2025 [Accessed 5 July 2026].
[227] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [121]–[122], [135]–[136].
[228] AI Incident Database, Incident 1256: Purportedly AI-Generated Deepfake Investment Ads Defrauded 5,000 Swedish Investors of 500 Million SEK [dataset]. https://incidentdatabase.ai/cite/1256/, 2025 [Accessed 5 July 2026].
[229] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [71]–[74], [97].
[230] European Commission, Commission investigates Grok and X's recommender systems under the Digital Services Act, IP/26/203. https://ec.europa.eu/commission/presscorner/detail/en/ip_26_203, 2026 [Accessed 5 July 2026].
[231] K. Conger, D. Freedman and S.A. Thompson, Musk's Chatbot Flooded X With Millions of Sexualized Images in Days, New Estimates Show, The New York Times. https://www.nytimes.com/2026/01/22/technology/grok-x-ai-elon-musk-deepfakes.html, 2026 [Accessed 5 July 2026].
[232] Center for Countering Digital Hate (CCDH), Grok Floods X with Millions of Sexualized Images of Women and Children. https://counterhate.com/research/grok-floods-x-with-sexualized-images/, 2026 [Accessed 5 July 2026].
[233] European Commission, Commission investigates Grok and X's recommender systems under the Digital Services Act, IP/26/203. https://ec.europa.eu/commission/presscorner/detail/en/ip_26_203, 2026 [Accessed 5 July 2026].
[234] Ofcom, Ofcom launches investigation into X over Grok sexualised imagery. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/ofcom-launches-investigation-into-x-over-grok-sexualised-imagery, 2026 [Accessed 5 July 2026].
[235] Information Commissioner's Office, ICO announces investigation into Grok. https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2026/02/ico-announces-investigation-into-grok/, 2026 [Accessed 5 July 2026].

March 2026 injunction (*ECLI:NL:RBAMS:2026:3106*).[236] The court ordered xAI and X to immediately cease generation, backed by a €100,000 daily penalty, finding xAI the "designated party" (*aangewezen partij*) responsible for preventing unlawful outputs.[237] Ruling that control over the model's architecture constitutes a "decisive influence" (*beslissende invloed*) over the processing,[238] thereby satisfying *Russmedia* standard,[239] the court effectively transformed safety-by-design from a regulatory aspiration into a strict liability rule.[240]

Beyond regulatory penalties, ensuing US class-action lawsuits allege that prioritizing engagement over safety resulted in foreseeable and material harm.[241] These filings, most significantly the mid-March 2026 action in *Doe v. xAI & X Corp.*,[242] specifically address the shortcoming of internal safety filters in preventing the industrial-scale fabrication of synthetic CSAM.

Systemic risks are often exacerbated by deliberate corporate retreats. Internal Meta documents (2024–2025) revealed a "Scam Export" ecosystem where the platform yielded $3 billion—roughly 19% of its China-based revenue—from advertisements promoting illegal gambling and fraud.[243] Leadership reportedly emphasized an "Integrity Strategy pivot" prioritizing revenue growth, despite internal evidence demonstrating that specialized enforcement could halve violation rates.[244]

The *Russmedia*[245] ruling definitive clarification to this retreat, confirming that DSA safe harbour protections do not extend to underlying data processing architectures. By positioning Articles 7 and 8 of the EU Charter (privacy and data protection) as superior mandates, the Court substantiated that platforms remain strictly liable as the architects of the dissemination parameters granting harmful content its industrial velocity.[246]

While the law stipulates identity verification, the CJEU in *Comdribus*[247] established a privacy-preserving perimeter. The judgement clarifies that platforms cannot engage in the "systematic collection" of biometric data; instead, they must prove technical measures were "strictly necessary" via "a sufficient statement of reasons".[248]

Finally, the judiciary is dismantling the trade secrecy shield. Relying on the CJEU *Dun & Bradstreet*[249] precedent, the Amsterdam Court of Appeal recently instituted eyes-only judicial reviews—examining unredacted system details to verify compliance (*X International v [Geïntimeerde]*).[250] This ensures that technical applications—specifically the duty to implement measures within the "current state of technical knowledge" mandated by *Russmedia*[251]—are subject to independent verification.

4.3. Cross-border enforcement: the jurisdictional and financial "black hole"

The integrity of the EU DSA's cross-border architecture hinges on the cooperation of Digital Services Coordinators (DSCs) under Article 58. This protocol allows a DSC of "destination" to request that a DSC of "establishment" take "investigatory and enforcement measures" (Article 58(1)).

However, the process is fundamentally misaligned with the rapidity of AI-driven fraud. In 2025, the volume of AI-enabled scams surged 1210%—rendering the 195% growth in traditional fraud nearly negligible by comparison.[252] Because requests must be "duly reasoned" (Article 58(3) and the DSC of establishment is afforded up to two months to communicate its assessment (Article 58(5)), the administrative trail often runs cold before it begins.

The extraterritorial reach of DSA directives is firmly anchored in the 2025 *Information Commissioner v Clearview AI Inc*,[253] precedent. The Upper Tribunal confirmed that foreign entities fall within national jurisdiction if their processing relates to the "behavioural monitoring" of residents—a definition broad enough to include passive data collection and automated sorting, regardless of the entity's physical location.[254]

Despite this jurisdictional claim, the DSA is effectively hampered by the granular procedural requirements of its core instruments. Under Article 9(1), enforcement is strictly limited to "specific items of illegal content" identified primarily by "one or more exact URL" (Article 9(2)(a)(iv)), a reactive prerequisite proving futile against the ephemeral, industrial-scale duplication of modern fraudulent deepfake campaigns. Article 10 compounds this friction, restricting providers to sharing only information "already collected" and "within [their] control" (Article 10(2)(b)). This statutory boundary creates a formidable obstacle when targeting sophisticated offshore entities engineering their systems specifically to prioritize uploader anonymity.

The 2025 organized criminal network behind the $35 million deepfake celebrity advertisement scam exposed this crisis.[255] While the DSA can compel platforms via Articles 9 and 10, its authority hits a hard jurisdictional wall at the EU border. Because the masterminds operated from Georgia[256]—a state outside the DSA's mandate—the DSC-facilitated instrument was rendered toothless against the crime's source.

[236] Amsterdam District Court, *Stichting Offlimits v. X.AI LLC, X Corp. and X Internet Unlimited Company*, ECLI:NL:RBAMS:2026:3106. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBAMS:2026:3106, 2026 [Accessed 5 July 2026].
[237] Ibid, [4.21], [5.1]–[5.4], [5.7], [5.12].
[238] Ibid, [4.21].
[239] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [72]–[73].
[240] R. Jahangir, Dutch Court Orders X, Grok to Stop AI-generated Sexual Abuse Content, Tech Policy Press. https://www.techpolicy.press/dutch-court-orders-x-grok-to-stop-aigenerated-sexual-abuse-content/, 2026 [Accessed 5 July 2026].
[241] *A. St. Clair v. xAI Holdings Corp.*, Index No. 150673/2026 (N.Y. Sup. Ct., filed 15 Jan 2026). https://iapps.courts.state.ny.us/fbem/DocumentDisplayServlet?documentId=wyklXeDrcrcGhJE6X8kMeQ==&system=prod [Accessed 5 July 2026]; *Jane Doe v. x.AI Corp. and x.AI LLC*, Case No. 5:26-cv-00772 (N.D. Cal., filed 23 Jan 2026). https://storage.courtlistener.com/recap/gov.uscourts.cand.463184/gov.uscourts.cand.463184.1.0.pdf [Accessed 5 July 2026].
[242] *Jane Doe 1 et al. v. x.AI Corp. and x.AI LLC*, Case No. 5:26-cv-02246 (N.D. Cal., filed 16 March 2026). https://storage.courtlistener.com/recap/gov.uscourts.cand.465940/gov.uscourts.cand.465940.1.0_1.pdf [Accessed 5 July 2026].
[243] Technology.org, Meta chooses revenue over users, lets Chinese scammers run wild. https://www.technology.org/2025/12/15/meta-chooses-revenue-over-users-lets-chinese-scammers-run-wild/, 2025 [Accessed 5 July 2026].
[244] Ibid.
[245] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [131].
[246] Ibid, [72]–[73], [130].
[247] Case C-371/24, *HW v Ministère public (Comdribus)* ECLI:EU:C:2026:234.
[248] Ibid, [34], [36], [51], [54], [71], [75].
[249] Case C-203/22, *CK v Magistrat der Stadt Wien (Dun & Bradstreet)*, ECLI:EU:C:2025:117, [72], [74]–[76].
[250] *X International Unlimited Company v [Geïntimeerde]* ECLI:NL:GHAMS:2025:2667, [3.17], [3.21] https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:GHAMS:2025:2667, 2025 [Accessed 5 July 2026].
[251] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [121]–[122].
[252] Vectra AI, AI scams explained: how AI-powered fraud works and how enterprises detect it, https://www.vectra.ai/topics/ai-scams, 2026 [Accessed 5 July 2026].
[253] *Information Commissioner v Clearview AI Inc* [2025] UKUT 319 (AAC).
[254] Ibid, [254], [366].
[255] S. Goodley, Z. Wood, P. Duncan, M. Goodier, Revealed: the scammers who conned savers out of $35m using fake celebrity ads, The Guardian. https://www.theguardian.com/money/2025/mar/05/revealed-the-scammers-who-conned-savers-out-of-35m-using-fake-celebrity-ads, 2025 [Accessed 5 July 2026].
[256] Ibid.

The fact that the victims were predominantly located in non-EU states like the UK and Canada[257]—jurisdictions currently excluded from direct access to DSA enforcement powers— intensifies this Georgia gap, forcing reliance on slow, discretionary international assistance protocols.[258] This jurisdictional black hole renders reactive removal orders effectively symbolic; if the law cannot cross the border to find the perpetrator, it must instead secure the foundational ingestion point. The 2026 FATF Horizon Scan protocols mitigate this stalemate by reclassifying AI-driven deepfake impersonation as a pervasive Anti-Money Laundering/Countering the Financing of Terrorism (AML/CFT) risk. This shift mandates a risk-based transition from static identity checks to "liveness" verification; jurisdictions failing to integrate advanced fraud detection against high-velocity synthetic identity scams are now flagged for strategic deficiencies. Such technological oversight triggers FATF grey list inclusion, internalizing punitive transaction friction and mandatory remediation roadmaps.[259] Furthermore, Furthermore, the 2026 Scam Center Strike Force successfully pioneered protocol-level asset freezing, empowering stablecoin issuers (USDT/USDC) to kill-switch wallets in non-compliant zones upon producing a fraudulent biometric integrity hash.[260]

The verify-then-publish mandate inaugurated in *Russmedia*,[261] thus serves as a necessary architectural barrier functioning where administrative disclosure orders cannot. A 2026 INTERPOL assessment verifies that "polycriminal" networks systematically exploit this void, using fraud proceeds to fund human trafficking and terrorism—offences operating entirely outside the reach of administrative disclosure orders.[262]

The confidentiality paradox of Article 10 aggravates this shortcoming. Although authorities can compel identity data, the disclosure is narrowly siloed for specific legal processes and requires a "statement of reasons" explaining the prerequisite's "necessary and proportionate" nature (Article 10(2)(a)(iv) DSA). This prevents authorities from preemptively warning the public or potential victims.[263] In the Georgia case, the truth emerged only through an external data "leak"[264] rather than the legal process. The DSA's privacy protections for suspects thus provide an inadvertent shield for offshore Fraud-as-a-Service (FaaS) syndicates.

By March 2026, the breakdown of traditional disclosure orders catalysed a pivot toward financial decapitation. In a landmark 2026 action, the Scam Center Strike Force bypassed jurisdictional friction by seizing over $580 million in cryptocurrency tied to transnational crime compounds in Southeast Asia.[265] Concurrently, the INTERPOL I-GRIP (Global Rapid Intervention of Payments) framework facilitated a financial kill-switch between Singapore and Timor-Leste, recovering $40 million in days[266]—an operational agility that DSA procedural disclosure mandates typically take months to achieve. Unlike the DSA's focus on the territorial *who* (Article 10), this approach targets the financial *how*—neutralizing the stablecoin ecosystems (primarily USDT) funding transnational fraud compounds.[267] By reconciling the verify-then-publish requirement of *Russmedia*[268] with the aggressive stablecoin seizures of 2026,[269] the legal framework shifts toward a denial-of-service (DoS) model: if the criminal cannot be reached, their ability to profit must be destroyed.

4.4. The UK safety model: technology-forcing mandates (Online Safety Act)

Unlike the EU's systemic focus, the UK Online Safety Act (OSA) imposes uniquely technology-forcing obligations regarding deepfake financial fraud. Scholarship critiquing the statute's limitations regarding deepfake sexual abuse[270] often overlooks its distinct avenue for financial fraud, where the law shifts from reactive moderation to compulsory technological intervention. Departing from well-settled intermediary immunity—such as the UK Supreme Court's finding in *Cartier International v British Telecommunications*[271]—the OSA shifts the cost and mitigation onus directly onto platforms, backed by penalties reaching 10% of worldwide revenue.[272] Section 10 (illegal content) and Section 38 (fraudulent advertising) duties[273] operationalize this statutory

[257] Ibid.

[258] A. Willoughby, The EU Digital Services Act: What You Need to Know as a UK Business, Burges Salmon. https://www.burges-salmon.com/our-thinking/the-eu-digital-services-act-what-you-need-to-know-as-a-uk-business, 2023 [Accessed 5 July 2026].

[259] TLT LLP, FATF Horizon Scan: AI & Deepfakes — Impacts on AML/CFT/CPF. https://www.tlt.com/insights-and-events/insight/fatf-horizon-scan-ai-deepfakes—-impacts-on-aml-cft-cpf, 2026 [Accessed 5 July 2026].

[260] U.S. Attorney's Office, District of Columbia, D.C. Scam Center Strike Force Seizures of Cryptocurrency from Chinese Transnational Criminals Tops $580 Million. https://www.justice.gov/usao-dc/pr/dc-scam-center-strike-force-seizures-cryptocurrency-chinese-transnational-criminals-tops, 2026 [Accessed 5 July 2026].

[261] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [104], [106].

[262] INTERPOL, INTERPOL report warns of increasingly sophisticated global financial fraud threat. https://www.interpol.int/en/News-and-Events/News/2026/INTERPOL-report-warns-of-increasingly-sophisticated-global-financial-fraud-threat, 2026 [Accessed 5 July 2026].

[263] T. Lenoir, Can the DSA be Useful Outside Europe?, Tech Policy Press. https://www.techpolicy.press/can-the-dsa-be-useful-outside-europe/, 2023 [Accessed 5 July 2026].

[264] S. Goodley, Z. Wood, P. Duncan, M. Goodier, Revealed: the scammers who conned savers out of $35m using fake celebrity ads, The Guardian. https://www.theguardian.com/money/2025/mar/05/revealed-the-scammers-who-conned-savers-out-of-35m-using-fake-celebrity-ads, 2025 [Accessed 5 July 2026].

[265] U.S. Attorney's Office, District of Columbia, D.C. Scam Center Strike Force Seizures of Cryptocurrency from Chinese Transnational Criminals Tops $580 Million. https://www.justice.gov/usao-dc/pr/dc-scam-center-strike-force-seizures-cryptocurrency-chinese-transnational-criminals-tops, 2026 [Accessed 5 July 2026].

[266] INTERPOL, INTERPOL report warns of increasingly sophisticated global financial fraud threat. https://www.interpol.int/en/News-and-Events/News/2026/INTERPOL-report-warns-of-increasingly-sophisticated-global-financial-fraud-threat, 2026 [Accessed 5 July 2026].

[267] R. Ramesh, Cryptohack Roundup: Ariomex Leak Flags Iran Sanction Risks. https://www.bankinfosecurity.com/cryptohack-roundup-ariomex-leak-flags-iran-sanction-risks-a-30919, 2026 [Accessed 5 July 2026].

[268] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [104], [106].

[269] R. Ramesh, Cryptohack Roundup: Ariomex Leak Flags Iran Sanction Risks. https://www.bankinfosecurity.com/cryptohack-roundup-ariomex-leak-flags-iran-sanction-risks-a-30919, 2026 [Accessed 5 July 2026].

[270] C. McGlynn, L. Woods, A. Antoniou, Pornography, the Online Safety Act 2023 and the need for further reform, J. Media Law 16 (2024) 211–239. https://doi.org/10.1080/17577632.2024.2357421; B. Kira, When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act, Comput. Law Secur. Rev. 54 (2024) 106024. https://doi.org/10.1016/j.clsr.2024.106024; J. Hörnle, Deepfakes and the Law: Why Britain needs stronger protections against technology-facilitated abuse, Queen Mary University of London. https://www.qmul.ac.uk/law/news/2025/items/deepfakes-and-the-law-why-britain-needs-stronger-protections-against-technology-facilitated-abuse.html, 2025 [Accessed 5 July 2026].

[271] *Cartier International AG v British Telecommunications Plc* [2018] UKSC 28, [31], [36].

[272] Online Safety Act 2023 (UK), c. 50, Sch 13, paras 4–5.

[273] Online Information Advisory Committee, Understanding Online Financial Harm Fraud, Scams and Disinformation Exposure Across Demographic Groups in the UK. https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/structure-and-leadership/online-information-advisory-committee/understanding-online-financial-harm-fraud.pdf?v=408231, 2025 [Accessed 5 July 2026].

command, directly responding to high-volume deepfake schemes targeting figures like Martin Lewis.[274]

To comply with this stipulation, Section 136 empowers Ofcom to require "proactive technology" to remedy deficiencies in addressing priority illegal content. As prescribed in Section 231(1), this order necessitates a triad of operational capabilities: content identification, user profiling, and behaviour identification.[275] The forensic track fulfils the "content identification technology" requirement (Section 231(2)) by deploying calibrated algorithms to detect agentic AI doubles. Tools like Reality Defender provide the *ex-ante* validation required to satisfy the "security of processing" obligations under Article 32 of the UK GDPR by spotting sub-perceptual inconsistencies.[276]

While the Michael Smith case (2026) involves US jurisdiction and synthetic audio, it serves as a seminal technical benchmark for the automated fraud lifecycle. Smith's enterprise obtained over $8 million in royalties by deploying thousands of bot accounts to stream AI-generated music. Notably, because the enterprise dispersed illicit activity across millions of tracks to evade "anomalous streaming" filters, each individual asset remained technically compliant at the binary level.[277]

This pervasive evidentiary deficit—where synthetic media is functionally indistinguishable from authentic data—renders Section 231(4) (user profiling) and Section 231(6) (behaviour identification) the essential legal levers of the OSA. Only by scrutinising "how [users] use" a service under Section 231(9)(b)—utilising behavioural biometrics to pinpoint intent-level signatures (e.g., via BioCatch)[278]—can platforms bridge the gap when the detectable signal is negated.

However, this transition from content-based to behaviour-based monitoring triggers an irreconcilable conflict with the UK GDPR's data minimisation principle (Article 5(1)(c)). Although Section 136(1) OSA authorises Ofcom to impose a "proactive technology requirement," Section 136(6)(b) circumscribes this power, restricting metadata analysis to content "communicated publicly." This limitation represents a calculated effort to reconcile the OSA with prohibitions against general monitoring and encryption-weakening.

Nevertheless, as upheld in the ECtHR *Podchasov v Russia,*[279] a normative conflict persists: Article 8 of the European Convention on Human Rights (ECHR) dictates that automated scrutiny must satisfy a stringent proportionality threshold wherever less invasive investigative alternatives remain viable. By internalising privacy and data protection risks via Section 136(8)(h), the Act effectively absorbs the *Podchasov*[280] deadlock into its own enforcement mechanism. Because behaviour identification (Section 231(6)) necessitates the ambient analysis of metadata to identify synthetic signatures,[281] the statute risks authorising a level of interference—processing communications data on a "generalised basis"—that the ECtHR ruled "impairs the very essence" of Article 8.[282] Consequently, the Smith precedent[283] exposes a regulatory deadlock: the forensic deficit of Section 231(2) technically obviates the Section 136 proactive technology power, which is legally constrained by the Article 8 ECHR vulnerabilities of Sections 231(4) and (6). This impasse confirms that statutory oversight remains untenable without the architectural reorientation proposed in Layer 1.

Unlike the EU's rights-based archetype, the UK model attempts to bridge this gap via the Digital Regulation Cooperation Forum (DRCF), which coordinates enforcement between Ofcom, the Financial Conduct Authority (FCA), the Information Commissioner's Office (ICO), and the Competition and Markets Authority (CMA). This alignment ensures regulators simultaneously address deepfake fraud as a content issue, a financial crime, and a data privacy violation.[284]

In early 2026, Ofcom signalled a pivot toward aggressive enforcement, effectively operationalizing the OSA's mandates through a £1 million fine against AVS Group for biometric lapses,[285] a record £1.35 million penalty against 8579 LLC,[286] and the regulator's pincer movement against X's Grok architecture.[287] While this signals the end of platform impunity,[288] the efficacy of these coercive measures remains contested.

Pointedly, this coercive power is not unfettered. The UK High Court in *Wikimedia Foundation v Secretary of State*[289] warned that regulatory actions must be justified as proportionate under the ECHR, specifically regarding the rights to privacy (Article 8), expression (Article 10), and assembly (Article 11). This standard reflects the ECtHR landmark finding in *Ahmet Yıldırım v Turkey*,[290] which characterized wholesale domain blocking as "collateral censorship" requiring strict judicial

[274] C. Barrett, Martin Lewis: 'Fine tech companies tens of billions for deepfake scams', Financial Times. https://www.ft.com/content/8625b60c-73d0-4d42-8641-eb2926040a6f, 2025 [Accessed 5 July 2026].

[275] See Online Safety Act 2023, c. 50, s 231(1) (defining proactive technology) and s 231(10) (confirming that proactive technology includes the use of AI or machine learning).

[276] B. Colman, How Agentic AI is Scaling the Threat of Deepfake Fraud, Reality Defender. https://www.realitydefender.com/insights/agentic-ai-deepfake-fraud, 2025 [Accessed 5 July 2026].

[277] U.S. Attorney's Office, Southern District of New York, North Carolina Man Pleads Guilty To Music Streaming Fraud Aided By Artificial Intelligence, https://www.justice.gov/usao-sdny/pr/north-carolina-man-pleads-guilty-music-streaming-fraud-aided-artificial-intelligence-0, 2026 [Accessed 5 July 2026].

[278] A. Pascual, Undetectable Scams: Deepfakes & AI Change the Game, BioCatch. https://www.biocatch.com/blog/undetectable-scams-deepfakes, 2026 [Accessed 5 July 2026].

[279] *Podchasov v. Russia*, App. no. 33696/19 (ECtHR, 13 February 2024), [78]–[79].

[280] Ibid, [78]–[81].

[281] U.S. Attorney's Office, Southern District of New York, North Carolina Man Pleads Guilty To Music Streaming Fraud Aided By Artificial Intelligence, https://www.justice.gov/usao-sdny/pr/north-carolina-man-pleads-guilty-music-streaming-fraud-aided-artificial-intelligence-0, 2026 [Accessed 5 July 2026].

[282] *Podchasov v. Russia*, App. no. 33696/19 (ECtHR, 13 February 2024), [80].

[283] U.S. Attorney's Office, Southern District of New York, North Carolina Man Pleads Guilty To Music Streaming Fraud Aided By Artificial Intelligence, https://www.justice.gov/usao-sdny/pr/north-carolina-man-pleads-guilty-music-streaming-fraud-aided-artificial-intelligence-0, 2026 [Accessed 5 July 2026].

[284] Digital Regulation Cooperation Committee (DRCF), Untangling the regulatory landscape: How DRCF member regulators are joining forces to tackle scams and fraud. https://www.drcf.org.uk/publications/blogs/how-drcf-member-regulators-are-joining-forces-to-tackle-scams-and-fraud, 2025 [Accessed 5 July 2026].

[285] C. Burt, Age verification without liveness nets Belize porn site operator £1M Ofcom fine, Biometric Update. https://www.biometricupdate.com/202512/age-verification-without-liveness-nets-belize-porn-site-operator-1m-ofcom-fine, 2025 [Accessed 5 July 2026].

[286] Ofcom, Ofcom fines porn company £1.35m for not having age checks, https://www.ofcom.org.uk/online-safety/protecting-children/ofcom-fines-porn-company-1.35-million-for-not-having-age-checks, 2026 [Accessed 5 July 2026].

[287] Ofcom, Ofcom launches investigation into X over Grok sexualised imagery, https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/ofcom-launches-investigation-into-x-over-grok-sexualised-imagery, 2026 [Accessed 5 July 2026].

[288] Department for Science, Innovation and Technology, Tech firms will have to take down abusive images within 48 hours under new law to protect women and girls, https://www.gov.uk/government/news/tech-firms-will-have-to-take-down-abusive-images-within-48-hours-under-new-law-to-protect-women-and-girls, 2026 [Accessed 5 July 2026].

[289] *Wikimedia Foundation v Secretary of State* [2025] EWHC 2086 (Admin), [135].

[290] *Ahmet Yıldırım v. Turkey*, App. no. 3111/10 (ECtHR, 18 December 2012), [44], [64]–[69].

review. *Podchasov v Russia*[291] further complicates this, ascertaining that filtering mandates necessitating a "weakening" of security architecture—specifically to monitor encrypted private communications—fail the Article 8 proportionality test. *Engels v Russia*[292] reinforces this protection, holding that filter-circumvention technologies, such as Virtual Private Networks (VPNs) and proxies, have legitimate uses. Consequently, a state cannot ban these tools wholesale without disproportionately interfering with the right to receive information under Article 10.[293]

This creates a definitive normative stalemate, exemplified by the UK's February 2026 proposal to restrict VPNs for minors. By attempting to limit these tools,[294] the government places the OSA's "proactive technology" mandate on a direct collision course with ECHR Articles 8, 10, and 11. Under *Podchasov*,[295] stripping encryption shields to monitor content fails the privacy proportionality test; under *Engels*,[296] limiting content-neutral evasion tools constitutes an overbroad interference with the right to receive information. Significantly, such a ban also risks infringing the freedom of association under Article 11; following the Strasbourg Court's rationale in *Kablis v Russia,*[297] recognizing online platforms as vital assembly spaces, a wholesale VPN restriction creates an impermissible "prior restraint" and a "significant collateral effect" on digital gathering places.

The technical reality acknowledged in *Getty Images v Stability AI*[298] compounds such conflict, where the High Court remarked that AI-driven filtering remains "far from perfect." As computer science research stresses, while safety guardrails utilize fewer computational resources than the models they guard, exploitable vulnerabilities—driven by computational asymmetry—remain mathematically unavoidable.[299]

Additionally, the 2026 DSIT report identifies a "regulatory uncertainty" effectively stalling the market's response.[300] Despite the OSA's statutory clarity, the UK's deepfake detection sector remains nascent: 83% of providers are micro or small enterprises averaging only £25 million in funding. High technical overhead and low perceived Return on Investment (ROI) compound this financial fragility, inducing a state of paralysis where platforms adopt a reactive "wait and see" stance despite mounting legislative pressure.[301]

Resolving this dilemma requires abandoning unattainable mathematical guarantees in favour of a systems-and-processes regulatory archetype. Under the OSA, platforms incur legal jeopardy not for isolated deepfake bypasses, but for failing to validate—via Section 9 risk assessments—that circumventions do not originate from structural flaws.[302] Satisfying the "suitable and sufficient" criteria of Section 9(2) and (4) necessitates navigating the lab-to-reality divergence by adopting DSIT's 2026 standardized accuracy assessments. This ensures safety protocols target empirical threats rather than optimistic laboratory benchmarks.[303] Fundamentally, this statutory imperative requires elevating expression and privacy duties (Section 22) above engagement-driven revenue. By pivoting from isolated evasion to auditable structural choice, the Act bridges institutional policing with the individual liability scrutinised in Layer 3.

5. Layer 3: accountability, deterrence, and enforcement

Building on the *ex-ante* verify-then-publish architecture formulated in Layer 2,[304] Layer 3 activates this directive as a coercive fail-safe. Where technical distribution controls reach the proportionality limits prescribed in the UK *Wikimedia Foundation v Secretary of State*,[305] Layer 3 compels substantive accountability by converting the *Russmedia*[306] identity requirements into a categorical liability trigger for both corporate and individual actors. It signals a decisive judicial pivot: moving from a reactive notice-and-takedown culture toward a paradigm of strict accountability where the state serves as the ultimate arbiter of digital integrity.[307]

5.1. The deepfake fraud criminal framework

5.1.1. International application: the Budapest Convention on Cybercrime

While the United Kingdom continues to rely on the domestic Fraud Act 2006,[308] global prosecution of AI-enabled fraud primarily utilizes the technology-neutral Budapest Convention on Cybercrime.[309] By focusing on the functional outcomes of a crime rather than specific technological means,[310] the Convention effectively encompasses syn-

[291] *Podchasov v. Russia*, App. no. 33696/19 (ECtHR, 13 February 2024), [77]–[81].
[292] *Engels v. Russia*, App. no. 61919/16 (ECtHR, 23 June 2020), [29].
[293] Ibid, [33]–[35].
[294] Prime Minister's Office, 10 Downing Street, PM: "No platform gets a free pass": Government takes action to keep children safe online, https://www.gov.uk/government/news/pm-no-platform-gets-a-free-pass-government-takes-action-to-keep-children-safe-online, 2026 [Accessed 5 July 2026].
[295] *Podchasov v. Russia*, App. no. 33696/19 (ECtHR, 13 February 2024), [77]–[81].
[296] *Engels v. Russia*, App. no. 61919/16 (ECtHR, 23 June 2020), [29], [33]–[35].
[297] *Kablis v. Russia*, App. nos. 48310/16 and 59663/17 (ECtHR, 30 April 2019) [81], [91], [94], [102], [105].
[298] *Getty Images (US) Inc v Stability AI Ltd* [2025] EWHC 2863 (Ch) [506].
[299] P. Hall, Cryptographers Show That AI Protections Will Always Have Holes, Quanta Magazine. https://www.quantamagazine.org/cryptographers-show-that-ai-protections-will-always-have-holes-20251210/, 2025 [Accessed 5 July 2026].
[300] Department for Science, Innovation and Technology, Deepfake detection technology, https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].
[301] Ibid.
[302] Ofcom, Illegal content duties under the Online Safety Act, https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/illegal-content-duties-under-the-online-safety-act, 2026 [Accessed 5 July 2026].
[303] Department for Science, Innovation and Technology, Deepfake detection technology, https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].
[304] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [92], [97], [104], [106].
[305] *Wikimedia Foundation v Secretary of State* [2025] EWHC 499 (Admin), [135].
[306] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [102], [106].
[307] L. Woods, Image Rights and False Claims, Data Protection and Intermediary Immunity: the case of Russmedia, EU Law Analysis. https://eulawanalysis.blogspot.com/2025/12/image-rights-and-false-claims-data.html, 2025 [Accessed 5 July 2026].
[308] Fraud Act 2006.
[309] Convention on Cybercrime (Budapest Convention), ETS No 185, 2001; Council of Europe, Explanatory Report to the Convention on Cybercrime, ETS No 185, 2001, para 36.
[310] Council of Europe, Explanatory Report to the Convention on Cybercrime, ETS No 185, 2001, para 36.

thetic media without requiring AI-specific amendments.[311] This ensures the legal architecture remains resilient against the rapid evolution of generative models—from GANs[312] to DiT architectures like Sora 2[313] and emergent reasoning-based threats.[314]

The Convention's international backbone rests on a dual-track prosecution regime that captures the full criminal lifecycle of cyber-enabled fraud. First, Article 7 (Computer-related Forgery) criminalizes unauthorized data input resulting in inauthentic data. In the context of deepfakes, this centres on deception regarding the data's "issuer".[315] By focusing on the unauthorized impersonation of the issuer at the system level,[316] Article 7 bypasses forensic debates over output quality. This standard is vital for prosecuting the surge in native virtual camera attacks and face-swap injections used to bypass automated identity validation.[317]

Complementing this, Article 8 (Computer-related Fraud) criminalizes property loss procured through data manipulation with "dishonest intent".[318] This provides the primary vehicle for prosecuting deepfake-enabled romance fraud,[319] such as the 2025 Brad Pitt scam that defrauded a victim of €830,000,[320] alongside synthetic identity fraud, in which actors merge stolen real-world data with synthetic biometrics to create "Frankenstein identities" at an industrial scale.[321] Ultimately, under this unified framework, the unauthorized system input constitutes forgery (Article 7), while the subsequent economic extraction constitutes consummated fraud (Article 8).

5.1.2. The enforcement, forensic, and jurisdictional deficit

To counter the anonymity and speed of cybercriminals, these substantive statutes are bolstered by the Second Additional Protocol to the Budapest Convention, which entered into force in 2024.[322] Crucially, Article 7 of the Protocol empowers authorities to bypass protracted mutual legal assistance (MLAT) channels and issue direct orders to foreign service providers for subscriber data.[323] This expedited disclosure mechanism is calibrated to rapidly identify Deepfake-as-a-Service (DFaaS) networks—such as those behind the 2024–2025 Hong Kong heists[324]—before actors can vanish into non-cooperative jurisdictions. Furthermore, Article 10 establishes expedited mutual assistance for imminent risks,[325] a procedural necessity demonstrated by the high-stakes 2024 Ferrari voice-cloning near-miss.[326]

Despite these procedural gains, global enforcement remains heavily fragmented. While the West treats the Budapest Convention as the "gold standard",[327] states including China, Russia, and Iran champion the 2025 Hanoi Convention.[328] This competing architecture prioritizes "cyber sovereignty," creating jurisdictional black holes[329] where stolen assets—such as the large-scale crypto transfers seen in the 2026 Ariomex exchange leak—remain functionally untouchable.[330]

Compounding this jurisdictional deficit, the collapsed barrier to entry for generative AI has fuelled a massive forensic deficit. With professional-grade deepfakes now costing as little as $5,[331] the sheer volume and complexity of caseloads are pushing law enforcement beyond capacity, rendering legacy forensic tools inadequate against the beyond reasonable doubt standard.[332] Because traditional prosecution cannot outpace algorithmic fraud, the UK's Fraud Strategy 2026–2029 formalizes a shift toward active disruption. Through initiatives like the Online Crime Centre, the focus is now on rapidly severing the technological infrastructure criminals rely on, rather than solely pursuing eventual convictions.[333]

[311] Convention on Cybercrime (Budapest Convention), ETS No 185, 2001, arts 7, 8.
[312] Biometric Update, Goode Intelligence, 2025 Deepfake Detection Market Report & Buyer's Guide. https://www.biometricupdate.com/wp-content/uploads/2025/10/BU-Deepfake-Detection-Market-Report-and-Buyers-Guide.pdf, 2025 [Accessed 5 July 2026].
[313] Sora (@soraofficialapp), We're saying goodbye to the Sora app. To everyone who created with Sora, shared it, and built community around it: thank you. https://x.com/soraofficialapp/status/2036546752535470382, 2026 [Accessed 5 July 2026].
[314] xAI, Grok 4.1 Model Card, 2025. https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf [Accessed 5 July 2026].
[315] Council of Europe, Explanatory Report to the Convention on Cybercrime, ETS No 185, 2001, para 82.
[316] Ibid, para 86.
[317] iProov, Threat Intelligence Report 2025: Remote Identity Under Attack. https://www.iproov.com/reports/threat-intelligence-report-2025-remote-identity-attack, 2025 [Accessed 5 July 2026].
[318] Council of Europe, Explanatory Report to the Convention on Cybercrime, ETS No 185, 2001, para 90.
[319] S. Moseley, Automating Deception: AI's Evolving Role in Romance Fraud, CETaS Briefing Papers, The Alan Turing Institute. https://cetas.turing.ac.uk/publications/automating-deception-ais-evolving-role-romance-fraud, 2025 [Accessed 5 July 2026].
[320] L. Gozzi, AI Brad Pitt dupes French woman out of €830,000, BBC News. https://www.bbc.co.uk/news/articles/ckgnz8rw1xgo, 2025 [Accessed 5 July 2026].
[321] LexisNexis Risk Solutions, The Hallmarks of a Synthetic Identity. https://risk.lexisnexis.co.uk/insights-resources/article/seven-signs-for-spotting-synthetic-identities, 2025 [Accessed 5 July 2026].
[322] Second Additional Protocol to the Convention on Cybercrime on enhanced co-operation and disclosure of electronic evidence, CETS No 224, 2022.
[323] Ibid, art 7.
[324] D. Elliott, '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 5 July 2026]; B. Colman, $18.5 Million Lost in a Crypto Scam: AI Voice Fraud's Growing Threat to Financial Security, Reality Defender. https://www.realitydefender.com/insights/millions-lost-in-hong-kong-crypto-scam, 2025 [Accessed 5 July 2026].
[325] Second Additional Protocol to the Convention on Cybercrime on enhanced co-operation and disclosure of electronic evidence, CETS No 224, 2022, art 10.
[326] S. Galletti, M. Pani, How Ferrari Hit the Brakes on a Deepfake CEO, MIT Sloan Management Review. https://sloanreview.mit.edu/article/how-ferrari-hit-the-brakes-on-a-deepfake-ceo/, 2025 [Accessed 5 July 2026].
[327] J. Shrier, Explanation of Position of the United States on the Adoption of the Resolution on the UN Convention Against Cybercrime in the UN General Assembly's Third Committee, United States Mission to the United Nations. https://usun.usmission.gov/explanation-of-position-of-the-united-states-on-the-adoption-of-the-resolution-on-the-un-convention-against-cybercrime-in-ungas-third-committee/, 2024 [Accessed 5 July 2026]; Council of Europe, Deputy Secretary General at the Opening of the Plenary Meeting of the Cybercrime Convention. https://www.coe.int/en/web/deputy-secretary-general/-/deputy-secretary-general-at-the-opening-of-the-plenary-meeting-of-the-cybercrime-convention, 2021 [Accessed 5 July 2026].
[328] United Nations, Convention against Cybercrime (adopted 24 December 2024) UNGA Res 79/243 (Hanoi Convention).
[329] A. Sukumar, A. Basu, Back to the territorial state: China and Russia's use of UN cybercrime negotiations to challenge the liberal cyber order, J. Cyber Policy 9 (2024) 256–287. https://doi.org/10.1080/23738871.2024.2436591.
[330] R. Ramesh, Cryptohack Roundup: Ariomex Leak Flags Iran Sanction Risks, BankInfoSecurity. https://www.bankinfosecurity.com/cryptohack-roundup-ariomex-leak-flags-iran-sanction-risks-a-30919, 2026 [Accessed 5 July 2026].
[331] J. Dhaliwal, State of the Scamiverse – How AI is Revolutionizing Online Fraud. https://web.archive.org/web/20250516014846/https://www.mcafee.com/blogs/internet-security/state-of-the-scamiverse/, 2025 [Accessed 5 July 2026].
[332] F. Cavalli, Top 10 Digital Forensics Challenges for Law Enforcement, Sensity. https://sensity.ai/blog/top-10-digital-forensics-challenges-law-enforcement/, 2026 [Accessed 5 July 2026]; Cellebrite, 2026 Industry Trends From Access to Insight: Modernizing Digital Investigations. https://cellebrite.com/wp-content/uploads/2026/02/2026-Industry-Trends-Survey-Report-Web.pdf, 2026 [Accessed 5 July 2026].
[333] Home Office, Fraud Strategy 2026 to 2029: disrupting crime, supporting economic resilience and delivering justice. https://www.gov.uk/government/publications/fraud-strategy-2026-to-2029, 2026 [Accessed 5 July 2026].

5.1.3. Domestic application: the UK Fraud Act 2006 and the identity crisis

The UK Fraud Act 2006 provides a robust, technology-neutral basis for prosecuting deepfake crimes.[334] By focusing on dishonest intent rather than the specific synthetic medium employed, the Act dictates that deepfakes—whether synthetic audio, video, image, or text[335]—serve immediately as the instrumental means by which statutory offenses are committed.

Primary Offense: Fraud by False Representation (Section 2)

The primary criminal mechanism for prosecuting the deepfake perpetrator is fraud by false representation (Section 2, Fraud Act 2006), which requires proof of a false representation made dishonestly with intent to cause loss or secure gain (Section 2(1)). Legally, deepfakes align seamlessly with this definition due to two critical subsections.

First, under implied representation (Section 2(4)), a deepfake video or voice recording constitutes a representation by conduct, implicitly asserting that the portrayed individual is authenticating the action or speaking the words. Critically, Section 2(5) explicitly covers representations submitted to "any system or device", ensuring deepfakes used to bypass automated biometric authentication (e.g., banking ID checks) are captured just as robustly as those deceiving human victims.[336] Karapatakis's analysis supports this interpretation, verifying the Fraud Act 2006 is sufficiently technologically neutral to apply Section 2 against AI-driven "Metafraud" and virtual asset scams.[337]

However, in the era of Fraud-as-a-Service (FaaS), a critical "cognitive gap" constrains the Act's efficacy.[338] A pervasive overconfidence bias exacerbates this: users systematically overestimate their ability to manually detect deepfakes while simultaneously distrusting the technological safeguards designed to protect them. The 2026 DSIT report corroborates this tension, noting public trust in AI-driven detection stands at just 36.9%, significantly trailing the 54.4% confidence afforded to traditional fact-checking.[339]

To bridge this gap between technical forensics and human understanding, grounded intelligence frameworks are emerging to move beyond binary "Real" or "Fake" labels.[340] By pairing proprietary detection models with the advanced reasoning of Gemini 3 Flash, providers now generate near real-time contextual risk summaries explaining why content is flagged. This transition from probabilistic scores to actionable intelligence is paramount for the "human handoff" in regulated industries, ensuring that operators are not "flying blind" during pressurized interactions, such as authenticating a caller at a bank.[341]

Successful prosecution hinges on proving *mens rea* (dishonesty) via the Supreme Court *Ivey v Genting Casinos*[342] test, which first ascertains the defendant's subjective belief as to the facts. Because low-skilled actors now utilize complex, black box deepfake tools that they do not technically understand,[343] establishing their specific knowledge of the technical falsehood constitutes a formidable evidentiary hurdle. If an end-user lacks the technical literacy to understand how a model generates a falsehood, proving they knew the representation was untrue becomes a challenge, particularly when FaaS protocols automate deception at a velocity that overwhelms traditional investigation.[344]

The sheer scale of the threat evidences this enforcement gap. The commercialization of deception has lowered the barrier to entry, with deepfake fraud attempts surging 3000% in 2023.[345] By 2025, identity fraud reached industrial velocity: deepfake attempts occurred every five minutes, while digital document forgeries surged 244% year-over-year.[346] Ultimately, relying on retrospective criminal prosecution is a strategic failure; it attempts to address a structural, automated vulnerability with a slow, case-by-case legal instrument that is inherently reactive in an age of instantaneous, scalable deception.

Preparatory and Supply Chain Offenses (Sections 6 and 7)

Sections 6 (possession) and 7 (supplying) of the Fraud Act 2006 offer crucial *ex-ante* intervention against the Fraud-as-a-Service ecosystem. Section 8 secures this capability by defining "article" to include any program or data held in electronic form (Section 8(2))— encompassing deepfake software and digital assets. This explicitly covers models used for identity manipulation, requiring only a general intention to commit future fraud, a standard confirmed in the domestic jurisprudence of UK *R v Ellames*.[347]

However, the dual-use nature of generative AI substantially compromises Section 7′s efficacy. Open-source models hosted on repositories such as Hugging Face and CivitAI create a material legal vacuum.[348] Because these tools have legitimate, non-malicious uses (e.g., political parody, satire, or film production),[349] establishing the necessary criminal *mens rea*—specifically, the supplier's knowledge that the tool will be used for fraud—is virtually impossible.

[334] Fraud Act 2006.

[335] Ofcom, Note to Broadcasters: Synthetic media (including deepfakes) in broadcast programming. https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/bulletins/broadcast-bulletins/2023/issue-471/note-to-broadcasters-synthetic-media-including-deepfakes-.pdf, 2023 [Accessed 5 July 2026].

[336] S. Mukherjee, M. Mohanty, Addressing deepfake issue in selfie-banking through camera-based authentication, arXiv:2508.19714 [cs.CR], 2025. https://doi.org/10.48550/arXiv.2508.19714.

[337] A. Karapatakis, Metaverse crimes in virtual (Un)reality: Fraud and sexual offences under English law, J. Econ. Criminol. 7 (2025) 100118. https://doi.org/10.1016/j.jeconc.2024.100118.

[338] Y.-T. Yang, Q. Zhu, Bi-Level Game-Theoretic Planning of Cyber Deception for Cognitive Arbitrage, arXiv:2509.05498 [cs.GT], 2025. https://doi.org/10.48550/arXiv.2509.05498.

[339] Department for Science, Innovation and Technology, Deepfake detection technology, https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology, 2026 [Accessed 5 July 2026].

[340] S. Dolen, Z. Ahmed, V. Dharmadhikari, Delivering real-time deepfake intelligence with Gemini 3 Flash, Google AI Blog. https://aistudio.google.com/case-studies/resembleai, 2025 [Accessed 5 July 2026].

[341] Ibid.

[342] *Ivey v Genting Casinos (UK) Ltd t/a Crockfords* [2017] UKSC 67, [62]–[63], [74].

[343] Secure.com, State of AI in Cybersecurity 2025: What's Real vs. Hype. https://www.secure.com/blog/ai-in-cybersecurity, 2025 [Accessed 5 July 2026].

[344] Entrust Cybersecurity Institute, 2025 Identity Fraud Report. https://www.entrust.com/sites/default/files/documentation/reports/2025-identity-fraud-report.pdf, 2025 [Accessed 5 July 2026].

[345] Brightside Team, Deepfake CEO Fraud: $50M Voice Cloning Threat CFOs. https://www.brside.com/blog/deepfake-ceo-fraud-50m-voice-cloning-threat-cfos, 2025 [Accessed 5 July 2026].

[346] Entrust Cybersecurity Institute, 2025 Identity Fraud Report. https://www.entrust.com/sites/default/files/documentation/reports/2025-identity-fraud-report.pdf, 2025 [Accessed 5 July 2026].

[347] Explanatory Notes to the Fraud Act 2006, [25]; *R v Ellames* (1974) 60 Cr App R 7 (CA).

[348] W. Hawkins, C. Russell, B. Mittelstadt, Deepfakes on Demand: the rise of accessible non-consensual deepfake image generators, arXiv:2505.03859 [cs.CY], 2025. https://doi.org/10.48550/arXiv.2505.03859.

[349] WIPO, Artificial intelligence: deepfakes in the entertainment industry, WIPO Magazine. https://www.wipo.int/en/web/wipo-magazine/articles/artificial-intelligence-deepfakes-in-the-entertainment-industry-42620, 2022 [Accessed 5 July 2026].

To narrow the *mens rea* gap, prosecutors may pivot to the recklessness standard established in the UK Supreme Court *Regina v G*.[350] Under this subjective test, liability requires proof that a developer foresaw the risk of fraud yet proceeded unreasonably. In the open-source sector, however, this test is neutralized by what Hacker identifies as the "social utility" defence: developers argue that the risk of fraud is justifiable when balanced against the model's value for research and innovation.[351] As Tammenlehto and Ahvo ascertain, this systemic failure mirrors the missing link in copyright law, where "technological neutrality" leads to an uncontrolled expansion of liability that renders the law unenforceable against complex digital activities.[352]

The result is a marketplace of impunity, evidenced by 35,000 unique publicly downloadable deepfake model variants[353] and the persistent re-uploading of 5000 banned models to Hugging Face[354] following the CivitAI UK-wide blockage.[355] Commercial actors like CrushAI further exploit this by utilizing platforms like Meta to advertise nudifying deepfake applications, banking on the gap between platform policy and criminal liability.[356]

CBS Songs Ltd v Amstrad[357] historically entrenches this legal paralysis, where the UK House of Lords held that manufacturers are not liable for supplying technology with both lawful and unlawful applications. While *Amstrad*[358] effectively immunizes developers of deepfake generators from Section 7 liability—even when they know their tools facilitate crime, provided the tools are not exclusively designed for fraud—the current landscape has shifted toward Model-as-a-Service (MaaS) liability. Punctuated by the CJEU *Russmedia*[359] identity mandate, this signals the end of the neutral tool defence; developers providing high-risk AI without verification guardrails can no longer claim immunity. This mandates technological accountability, ensuring the law no longer functions as a picket fence where only the end-user is prosecuted while the instrument provider remains shielded.[360]

5.2. Corporate accountability: the dual mandate for prevention and liability

The escalating threat of generative AI-enabled financial crime has precipitated two distinct but complementary regulatory responses: the UK's focus on strict corporate liability and the EU's target of individual managerial accountability.

5.2.1. The UK: strict liability and the "reasonable procedures" defense

The UK's corporate criminal offence of "failure to prevent fraud" (Economic Crime and Corporate Transparency Act (ECCTA) 2023) fundamentally transfers the burden of detection onto large organisations.[361] Unlike legacy statutes hampered by the "identification doctrine," this law imposes liability for fraud committed by an "associated person"—encompassing autonomous AI agents—to benefit the firm.[362]

Under the ECCTA, the Arup standard—established by the $25.6 million deepfake loss in 2024[363]—now constitutes the benchmark for "reasonable fraud prevention measures".[364] To secure a "reasonable procedures" defence, large organizations must demonstrate that their safeguards include the state-of-the-art technical standards mandated by the OSA, specifically Injection Attack Detection (IAD). This converts Layer 2 technical standards into a prerequisite for avoiding criminal liability, marking the transition from compliance-as-checklist to compliance-as-technical-integrity.

Meta's 2025 whitelisting scandal evidences the systemic failure of such procedures. Internal documents uncovered a "mistake prevention" protocol where ads flagged for fraud by AI filters were allowed to remain active for days awaiting human review—a strategic latency that allowed scammers to achieve "massive impressions" and generate $3 billion in annual revenue.[365] This systemic bypass allowed Meta to monetize an estimated $16 billion in "high risk" ads globally.[366] Under the ECCTA's model of liability,[367] a deliberate failure to deploy available detection technology—where that omission maximises corporate profit—likely nullifies a "reasonable fraud prevention measures" defence.[368] This suggests that the agency-reseller model can no longer function as a shield against accountability; profit-driven gaps in monitoring fail the statutory threshold of reasonableness and contravene the Serious Fraud Office's (SFO) mandate for an outcome-driven anti-fraud culture.[369]

5.2.2. The EU: piercing the corporate veil

Conversely, the EU increasingly targets the boardroom by piercing

[350] *Regina v G and other* [2003] UKHL 50, [10].
[351] P. Hacker, The European AI liability directives – Critique of a half-hearted approach and lessons for the future, Comput. Law Secur. Rev. 51 (2023) 105871. https://doi.org/10.1016/j.clsr.2023.105871.
[352] L. Tammenlehto, E. Ahvo, Technological Neutrality, Hyperlinking and Criminal Liability – In Search of the Missing Link?, IIC Int. Rev. Intellect. Prop. Compet. Law 56 (2025) 1064–1091. https://doi.org/10.1007/s40319-025-01611-7.
[353] W. Hawkins, C. Russell, B. Mittelstadt, Deepfakes on Demand: the rise of accessible non-consensual deepfake image generators, arXiv:2505.03859 [cs. CY], 2025. https://doi.org/10.48550/arXiv.2505.03859.
[354] E. Maiberg, Hugging Face Is Hosting 5,000 Nonconsensual AI Models of Real People, 404 Media. https://www.404media.co/hugging-face-is-hosting-5-000-nonconsensual-ai-models-of-real-people/, 2025 [Accessed 5 July 2026].
[355] Civitai, Terms of Service. https://civitai.com/content/tos, 2025 [Accessed 5 July 2026].
[356] E. Maiberg, Instagram Ads Send This Nudify Site 90 Percent of Its Traffic, 404 Media. https://www.404media.co/instagram-ads-send-this-nudify-site-90-percent-of-its-traffic/, 2025 [Accessed 5 July 2026].
[357] *CBS Songs Ltd v Amstrad Consumer Electronics Plc* [1988] AC 1013.
[358] Ibid.
[359] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI: EU:C:2025:935, [102], [106].
[360] Ibid, [131], [136].
[361] Home Office, Guidance: Offence of 'failure to prevent fraud' introduced by ECCTA. https://www.gov.uk/government/publications/offence-of-failure-to-prevent-fraud-introduced-by-eccta, 2024 [Accessed 5 July 2026]; GOV.UK, New measures to tackle fraud come into effect. https://www.gov.uk/government/news/new-measures-to-tackle-fraud-come-into-effect, 2025 [Accessed 5 July 2026].
[362] Home Office, Economic Crime and Corporate Transparency Act 2023: Guidance to organisations on the offence of failure to prevent fraud (accessible). https://www.gov.uk/government/publications/offence-of-failure-to-prevent-fraud-introduced-by-eccta/economic-crime-and-corporate-transparency-act-2023-guidance-to-organisations-on-the-offence-of-failure-to-prevent-fraud-accessible-version, 2024 [Accessed 5 July 2026].
[363] D. Elliott, '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 5 July 2026].
[364] Home Office and others, 'New Measures to Tackle Fraud Come into Effect' (News story, 2025) <https://www.gov.uk/government/news/new-measures-to-tackle-fraud-come-into-effect>
[365] Technology.org, Meta chooses revenue over users, lets Chinese scammers run wild. https://www.technology.org/2025/12/15/meta-chooses-revenue-over-users-lets-chinese-scammers-run-wild/, 2025 [Accessed 5 July 2026].
[366] Ibid.
[367] Home Office, Economic Crime and Corporate Transparency Act 2023: Guidance to organisations on the offence of failure to prevent fraud (accessible). https://www.gov.uk/government/publications/offence-of-failure-to-prevent-fraud-introduced-by-eccta/economic-crime-and-corporate-transparency-act-2023-guidance-to-organisations-on-the-offence-of-failure-to-prevent-fraud-accessible-version, 2024 [Accessed 5 July 2026].
[368] Home Office and others, 'New Measures to Tackle Fraud Come into Effect' (News story, 2025) <https://www.gov.uk/government/news/new-measures-to-tackle-fraud-come-into-effect>
[369] Serious Fraud Office, SFO Guidance on Evaluating a Corporate Compliance Programme. https://www.gov.uk/government/publications/sfo-guidance-on-evaluating-a-corporate-compliance-programme/sfo-guidance-on-evaluating-a-corporate-compliance-programme, 2025 [Accessed 5 July 2026].

the corporate veil. While the UK's ICO successfully established jurisdiction to issue fines against companies like Clearview AI (£7.5 million) in *Information Commissioner v Clearview AI Inc*,[370] the EU pivots toward personal criminal liability for individual managers.

Although the CJEU specified in *Deutsche Wohnen SE*[371] that administrative fines can be levied directly on corporations without identifying a natural person, Article 84(1) GDPR empowers Member States to intensify sanctions. The criminal complaint filed by Noyb in Austria targets Clearview AI's managers, invoking Article 84 GDPR (implemented via Section 63 of the Austrian Data Protection Act).[372] This provision allows for criminal sanctions, including imprisonment, where administrative fines have been ignored.[373] This approach fulfils the CJEU mandate delineated in *ILVA A/S*[374]—that penalties must be "effective, proportionate, and dissuasive"—by precluding complex corporate structures from serving as a shield for personal immunity. Unlike the US Illinois Biometric Information Privacy Act (BIPA), which primarily results in monetary settlements,[375] relying on GDPR criminal statutes introduces the existential risk of detention and European Arrest Warrants.[376]

This dual approach—UK corporate fines compelling technical adoption and EU criminal charges targeting the directing mind—closes the impunity gap for non-compliant AI actors. It ensures the corporate mind is held as accountable as the AI agent it harnesses, creating a powerful incentive to prioritise technical safety over quarterly margins.

5.3. Judicial integrity and the identity mandate

The viability of the digital accountability layer depends upon the judiciary's ability to discern fact from fiction—a capability currently under systemic erosion. As deepfake technology infiltrates legal workflows, the judicial system faces an authentication crisis where synthetic media undermines the reliability of discovery and testimony.[377]

This risk first manifested in the UK Family Courts in 2020, where a parent submitted a digitally altered recording to influence a risk assessment; authenticating the raw metadata exposed the fraud.[378] This case demonstrated that traditional forensic verification, which relies heavily on judicial intuition, is insufficient for the AI era. The current legal consensus—evolving from simple document authentication to a system-integrity standard—posits that judicial notice can no longer be extended to digital media without a verifiable, unbroken line of audit logs.[379]

The judiciary now faces the "liar's dividend"—a phenomenon where the mere existence of deepfakes allows genuine evidence to be discredited.[380] This vulnerability has forced a pivot toward *ex-ante* technical verification, with detection suites like Reality Defender integrated into legal discovery platforms by late 2025.[381]

While the CJEU in *Russmedia*[382] focused on the obligations of online marketplaces, its mandate for verification by design provides the jurisprudential justification for courts to redefine system integrity. By requiring evidentiary provenance to be established via immutable manifests before admission—e.g., via C2PA-compliant protocols[383]—the judiciary operationalizes the *Russmedia*[384] stipulation that technical measures must be "appropriate" to the heightened risks of the AI era.

The effectiveness of these efforts hinges upon the final layer of accountability: the *Russmedia*[385] identity mandate. The CJEU dismantled the concept of anonymous impunity by mandating that platforms collect and verify the identity of user advertisers prior to publication.[386] The Court ruled that where a platform allows advertisements to be placed anonymously, the risk of identity theft or fraud and the processing of sensitive data without consent is unacceptably high.[387] This converts identity verification from a business choice into a statutory security requirement under Articles 24 and 25 of the GDPR.[388] For a unified defence, this mandate furnishes the legal foundation for cryptographically verified identities, addressing the evidentiary crisis at the source.[389] By enforcing the collection of verifiable data, the law

[370] *Information Commissioner v Clearview AI Inc* [2025] UKUT 319 (AAC), [364].
[371] Case C-807/21, *Deutsche Wohnen SE v Staatsanwaltschaft Berlin* ECLI:EU:C:2023:950, [53], [60].
[372] Noyb – European Center for Digital Rights, Criminal complaint against facial recognition company Clearview AI. https://noyb.eu/en/criminal-complaint-against-facial-recognition-company-clearview-ai, 2025 [Accessed 5 July 2026].
[373] Austrian Data Protection Act (Datenschutzgesetz, DSG) 2018, s 63.
[374] Case C-383/23, *Criminal proceedings against ILVA A/S* ECLI:EU:C:2025:84, [36].
[375] *In re Clearview AI, Inc., Consumer Privacy Litigation*, No 21-cv-00135 (ND Ill 20 March 2025).
[376] Noyb – European Center for Digital Rights, Criminal complaint against facial recognition company Clearview AI. https://noyb.eu/en/criminal-complaint-against-facial-recognition-company-clearview-ai, 2025 [Accessed 5 July 2026].
[377] M.-H. Maras, A. Alexandrou, Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos, Int. J. Evid. Proof 23 (2018) 255–273. https://doi.org/10.1177/1365712718807226; J.P. LaMonaga, A Break From Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes, Am. Univ. L. Rev. 69 (2020) 1761–1800. https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=2221&context=aulr; M.-P. Sandoval, M. de Almeida Vau, J. Solaas, L. Rodrigues, Threat of deepfakes to the criminal justice system: a systematic review, Crime Sci. 13 (2024) 41. https://doi.org/10.1186/s40163-024-00239-1.
[378] G. Swerling, Doctored audio evidence used to damn father in custody battle, The Telegraph. https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/, 2020 [Accessed 5 July 2026].
[379] S. Mason, D. Seng, The foundations of evidence in electronic form, in: S. Mason, D. Seng (Eds.), Electronic Evidence, fourth ed., University of London Press, London, 2017, pp. 31–74. https://doi.org/10.14296/517.9781911507079.
[380] K. J. Schiff, D. S. Schiff, N. S. Bueno, The liar's dividend: Can politicians claim misinformation to evade accountability?, Am. Political Sci. Rev. 119 (2025) 71–90. https://doi.org/10.1017/S0003055423001454; Brighton and Hove Law, The dangers of doctored "deep fake" evidence that is being submitted to Courts. https://brightonandhovelaw.co.uk/the-dangers-of-doctored-deep-fake-evidence-that-is-being-submitted-to-courts/, 2020 [Accessed 5 July 2026].
[381] M. Borak, Reality Defender deepfake detection flags synthetic media use in legal workflows, Biometric Update. https://www.biometricupdate.com/202512/reality-defender-deepfake-detection-flags-synthetic-media-use-in-legal-workflows, 2025 [Accessed 5 July 2026].
[382] Case C-492/23, *X v Russmedia Digital SRL, Inform Media Press SRL*, ECLI:EU:C:2025:935, paras [97], [121], [126].
[383] Coalition for Content Provenance and Authenticity (C2PA), C2PA Technical Specification v2.3. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html, 2025 [Accessed 5 July 2026].
[384] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [121], [126].
[385] Ibid, [54], [97].
[386] Ibid, [69], [102], [104].
[387] Ibid, [96], [103].
[388] Ibid, [106], [136].
[389] Coalition for Content Provenance and Authenticity (C2PA), C2PA Technical Specification v2.3. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html, 2025 [Accessed 5 July 2026].

solidifies the "failure to prevent fraud" liability as a credible deterrent[390] transitioning the digital environment from unchecked anonymity into a managed common of verified participants.

5.4. Right to compensation and the damage baseline (GDPR article 82)

The final civil enforcement mechanism augmenting criminal statutes is the right to compensation enunciated under Article 82 GDPR. The CJEU has affirmed in *UI v Österreichische Post AG*[391] that victims may seek damages from negligent data controllers. This mechanism is notably potent because it allows deepfake victims to target the systemic architecture that enabled the fraud, rather than solely pursuing the often-elusive individual perpetrator.

Systemic security failures in the generative AI supply chain throughout 2025 and early 2026 exacerbated this vulnerability. The GenNomis data breach first exposed nearly 100,000 deepfake images,[392] followed by at least one other similar security lapse, a pattern culminating in December 2025 when a database linked to the AI startup SocialBook—containing over 1.1 million non-consensual "nudified" assets from associated apps MagicEdit and DreamPal— sat negligently exposed on the open internet.[393] This trajectory reached an acute inflection point in March 2026 with the Mercor breach.[394] A malicious code injection into the open-source LiteLLM library—a tool utilized by developers at OpenAI, Meta, and Anthropic—facilitated the exfiltration of four terabytes of sensitive data, including standardized identity documents and raw face and voice biometrics, effectively providing threat actors with the "keys" to industrialized impersonation.[395]

In such instances, the victim's harm transcends the initial potential fraud; it constitutes a permanent loss of control over their biometric identity. Synthetic identity fraud surging 8-fold in 2025—now 11% of global fraud—underscores this crisis.[396] Such failure triggers the heightened liability standard prescribed in *VS v Ministerstvo na vatreshnite raboti*,[397] which stipulates that the processing of biometric data—due to its inherent "significant risks"—is only permitted where "strictly necessary." Decisively, the Court ruled that this assessment of necessity is a non-delegable duty of the controller that cannot be cured by judicial review.[398] When coupled with *VB v Natsionalna agentsia za prihodite*,[399] the burden of proof shifts to the controller to demonstrate that their security was "appropriate" for such high-risk data.

Doctrinally, as Stalla-Bourdillon and Kuczerawy observe, the *Russmedia*[400] judgment imposes four sequential obligations that serve as the contemporary benchmark for this appropriate security: first, a mandate to detect whether advertisements contain special category data; second, a requirement to verify if the advertiser is the data subject; third, a duty to refuse publication absent a valid legal basis; and fourth, a proactive obligation to prevent such sensitive data from being unlawfully republished on third-party domains. They argue that the CJEU is making a "performative claim" by asserting that these duties do not constitute a prohibited general monitoring obligation—a stance that essentially reshapes the prohibition of general monitoring under the former E-Commerce Directive (Article 15) by requiring operators to systematically assess the lawfulness of processing for every advertisement at scale.[401] This raises the regulatory bar to a level that demands architectural interventions—such as the zero-retention biometrics proposed in Layer 1.

Advocate General (AG) Rantos further fortified victims' rights in *Tukowiecka*.[402] Clarifying Articles 73(1) and 74(1) of Directive (EU) 2015/2366, the AG established that payment providers must restore funds immediately, precluding "gross negligence" as an extrajudicial shield to defer refunds.[403] This bifurcation of refunding and liability transfers the primary liquidity risk from the consumer to the institution, effectively treating unauthorised transactions as an immediate balance-sheet liability for the bank.[404] By stripping "gross negligence" of its defensive utility, the law ensures the financial weight of AI-driven deception is borne by the system's institutional architects.[405] If adopted by the CJEU, this shift transforms high-assurance standards—specifically Injection Attack Detection (IAD) and zero-retention biometrics—from discretionary security into mandatory financial risk-mitigation imperatives.

Despite the clarity of Article 82 GDPR, victims still face formidable hurdles regarding the state-of-the-art benchmark. Many controllers continue to rely on obsolete Presentation Attack Detection (PAD) standards, even as deepfake delivery shifts toward Digital Injection Attacks (DIA).[406] By subverting the physical camera sensor—rendering legacy PAD controls functionally moot—these attacks highlight a pressing divergence from the "appropriate" security requirements of Article 32 (1). Legally, this creates a specific battleground over Article 82(3) exoneration. As held in *VB v Natsionalna agentsia,*[407] a controller is only exempt from liability if they prove they are not "in any way" responsible

[390] Home Office, Guidance: Offence of 'failure to prevent fraud' introduced by ECCTA. https://www.gov.uk/government/publications/offence-of-failure-to-prevent-fraud-introduced-by-eccta, 2024 [Accessed 5 July 2026]; GOV.UK, New measures to tackle fraud come into effect. https://www.gov.uk/government/news/new-measures-to-tackle-fraud-come-into-effect, 2025 [Accessed 5 July 2026].

[391] Case C-300/21, *UI v Österreichische Post AG* ECLI:EU:C:2023:370, [32], [37], [46], [49], [51], [57]–[58].

[392] J. Fowler, Thousands of AI & DeepFake Images Exposed on Nudify Service Data Breach, VPN Mentor. https://www.vpnmentor.com/news/report-gennomis-breach/, 2025 [Accessed 5 July 2026].

[393] J. Fowler, Popular AI Generator Exposed Over One Million Images Including DeepFakes and Nudify Face Swaps, ExpressVPN Blog. https://www.expressvpn.com/blog/magicedit-data-exposed/, 2025 [Accessed 5 July 2026]; M. Burgess, Huge Trove of Nude Images Leaked by AI Image Generator Start-up's Exposed Databas, Wired. https://www.wired.com/story/huge-trove-of-nude-images-leaked-by-ai-image-generator-startups-exposed-database/, 2025 [Accessed 5 July 2026].

[394] B. Nolan, Mercor, a $10 billion AI startup that works with companies including OpenAI and Anthropic, confirms major data breach, Fortune. https://fortune.com/2026/04/02/mercor-ai-startup-security-incident-10-billion/, 2026 [Accessed 5 July 2026].

[395] L.-H. Liang, AI company's breached biometrics, ID document images make deepfake fraud easier, Biometric Update. https://www.biometricupdate.com/202604/ai-companys-breached-biometrics-id-document-images-make-deepfake-fraud-easier, 2026 [Accessed 5 July 2026].

[396] LexisNexis Risk Solutions, Cybercrime Report 2026: Evolving Threats Beneath the Surface. https://risk.lexisnexis.com/-/media/files/financial%20services/research/lexisnexis-risk-solutions-cybercrime-report-2026.pdf, 2026 [Accessed 5 July 2026].

[397] Case C-205/21, *Ministerstvo na vatreshnite raboti (Recording of biometric and genetic data by the police)* ECLI:EU:C:2023:49, [116], [118], [125], [127]-[128], [135].

[398] Ibid, [130], [134].

[399] Case C-340/21, *VB v Natsionalna agentsia za prihodite* ECLI:EU:C:2023:986, [57].

[400] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935.

[401] S. Stalla-Bourdillon, A. Kuczerawy, The CJEU in *Russmedia*: much ado about nothing new?, Priv. Data Prot. 26 (2026) 14–19. https://researchportal.vub.be/en/publications/the-cjeu-in-russmedia-much-ado-about-nothing-new/

[402] Case C-70/25, *Tukowiecka (N.O. v PKO BP S.A.)*, Opinion of AG Rantos delivered on 5 March 2026 [Provisional Text].

[403] Ibid, [26]–[27], [38]–[40].

[404] Ibid, [28], [37], [41].

[405] Ibid, [16], [41].

[406] D. Elliott, '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 5 July 2026].

[407] Case C-340/21, *VB v Natsionalna agentsia za prihodite* ECLI:EU:C:2023:986, [70], [74].

for the event. Given that legacy models fail to identify the vast majority of synthetic identities—as many as 85%[408]—the continued use of PAD against a known DIA threat vector constitutes a culpable failure to adapt.

Finally, a new judicial baseline has emerged to address the subjective nature of quantifying non-material harm. While the UK historically set a high bar for distress claims—exemplified by the *Lloyd v Google*[409] and *Rolfe v Veale*[410] rulings—the €7000 award resulting from the CJEU *Russmedia*[411] litigation serves as a quantifiable damage baseline for identity theft and reputational harm, effectively replacing the era of nominal damages with concrete financial benchmarks for non-compliance. However, the CJEU has recently moved to prevent this baseline from being exploited by professional litigants. In the landmark *Brillen Rottler*[412] ruling, the Court established an abuse filter, decreeing that Article 15 GDPR data access requests—and subsequent Article 82 claims—may be refused if they are made solely for "artificially creating the conditions" for a lawsuit. This synthesis of criminal, corporate, and civil liability establishes Layer 3 as a holistic deterrent, bridging the gap between statutory ideals and the technical reality of the AI era.

6. Discussion of findings and analysis

6.1. Synthesis of the critical failure point and layered response

Empirical data substantiates that the deepfake threat has decisively shifted from political misinformation to a pervasive financial security vulnerability,[413] as 55% of active threat groups now pursue financial gain.[414] This study's multi-layered analysis reveals four fundamental axioms. First, an identity crisis roots the vulnerability (Layer 1); "Frankenstein" synthetic fraud will cost the UK economy £4.2 billion by 2027.[415] Second, regarding defensive integrity (Layer 2), current algorithmic compliance proves insufficient against the 2665% surge in digital injection attacks.[416] This mandates a shift to high-assurance Injection Attack Detection (IAD) standards. Third, regarding the accountability lacuna (Layer 3), a 12-month governance vacuum exists between the December 2026 compliance deadline for the Digital Omnibus bans and the December 2027 application of high-risk AI obligations.[417] Finally, the divergence between rights-based (EU) and safety-based (UK) regimes necessitates a Transatlantic Regulatory and Financial Interoperability Framework. By transitioning to a centralized protocol-level blockade—integrated into ISO 20022 schemas[418]—states can structurally neutralize jurisdictional arbitrage.

6.2. The EU rights model vs. the UK safety model

A fundamental divergence persists between the two primary regulatory jurisdictions: the European Union, anchored by the landmark *X v Russmedia Digital SRL*[419] ruling, establishes proactive restrictions based on Layer 1 privacy rights, whereas the United Kingdom's Online Safety Act (OSA) imposes a structural duty of care requiring proportionate "proactive technology" at Layer 2.[420]

This research contends that integrating biometric integrity hashes[421] into the ISO 20022 messaging schema[422] serves as the pivotal technical bridge between these bifurcated philosophies. While the EU emphasizes the rigid processing of biometric data[423] and the UK prioritizes technology-forcing duties,[424] the implementation of a financial blockade compels a unified technical front. This approach reconciles the EU's verify-then-publish mandate[425] with the UK's safety duties[426] by embedding legal compliance directly into the foundational plumbing of the global digital economy.

6.3. Global comparative analysis: technology-forcing regimes (US and China)

The principle-based frameworks of the EU and UK diverge from the prescriptive architectural controls, adopted by the United States and China. Both jurisdictions increasingly align with the logic delineated in *Russmedia*,[427] which posits that platforms function as the non-delegable architects of the data they amplify. This Layer 2 intervention is designed to strip away the anonymity essential to transnational fraud rings, a necessity underscored by the 1740% surge in North American deepfake

[408] LexisNexis Risk Solutions, The Hallmarks of a Synthetic Identity. https://risk.lexisnexis.co.uk/insights-resources/article/seven-signs-for-spotting-synthetic-identities, 2025 [Accessed 5 July 2026].
[409] *Lloyd v Google LLC* [2021] UKSC 50 (10 November 2021).
[410] *Rolfe v Veale Wasbrough Vizards LLP* [2021] EWHC 2809 (QB).
[411] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [32], [103], [106], [136].
[412] Case C-526/24, *Brillen Rottler GmbH & Co. KG v TC* ECLI:EU:C:2026:216, [31], [40]–[41], [66]–[67].
[413] B. Colman, Why detecting dangerous AI is key to keeping trust alive in the deepfake era, World Economic Forum. https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/, 2025 [Accessed 5 July 2026].
[414] Google Cloud Security, Mandiant M-Trends 2025 Report Executive Edition. https://services.google.com/fh/files/misc/m-trends-2025-executive-edition-en.pdf, 2025 [Accessed 5 July 2026].
[415] LexisNexis Risk Solutions, The Hallmarks of a Synthetic Identity. https://risk.lexisnexis.co.uk/insights-resources/article/seven-signs-for-spotting-synthetic-identities, 2025 [Accessed 5 July 2026].
[416] iProov, Threat Intelligence Report 2025: Remote Identity Under Attack. https://www.iproov.com/reports/threat-intelligence-report-2025-remote-identity-attack, 2025 [Accessed 5 July 2026].
[417] European Parliament, AI Act: EP approves simplification measures and "nudifier" app ban. https://www.europarl.europa.eu/news/en/press-room/20260611IPR45207/ai-act-ep-approves-simplification-measures-and-nudifier-app-ban, 2026 [Accessed 5 July 2026].
[418] ISO 20022, ISO 20022: The universal financial industry message scheme. https://www.iso20022.org/, 2025 [Accessed 5 July 2026].
[419] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935.
[420] Online Safety Act 2023 (UK), c. 50, s 136.
[421] J. McConvey, Explainer: What is biometric hashing?, Biometric Update. https://www.biometricupdate.com/202505/explainer-what-is-biometric-hashing, 2025 [Accessed 5 July 2026].
[422] ISO 20022, ISO 20022: The universal financial industry message scheme. https://www.iso20022.org/, 2025 [Accessed 5 July 2026].
[423] Regulation (EU) 2016/679 (General Data Protection Regulation), OJ L 119 (2016) 1, Art 9(1); Case C-205/21, *Ministerstvo na vatreshnite raboti (Recording of biometric and genetic data by the police)*, ECLI:EU:C:2023:59, [63], [115]–[135].
[424] Online Safety Act 2023 (UK), c. 50, ss 10, 38, 136, 231.
[425] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [104], [106].
[426] Online Safety Act 2023 (UK), c. 50, ss 10, 38, 136, 231.
[427] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [72], [74], [131].

fraud and a corresponding 1530% increase in the Asia Pacific region.[428]

Within the US, state-level legislation integrates a "forged digital likeness" concept into existing identity crime statutes. For example, Washington State's Substitute House Bill 1205 (RCW 9A.60.045(1)(c)) criminalizes the knowing distribution of forged digital media with fraudulent intent. By establishing an architectural threshold where content is "indistinguishable" from reality and likely to deceive a reasonable person, the law provides the forensic criteria necessary for prosecution while maintaining specific exceptions for constitutionally protected speech such as satire.[429] Similarly, Pennsylvania's Act 35 (§ 4101.1) formalizes "digital forgery" as a distinct criminal offense, targeting the generative act and distribution of deceptive synthetic content.[430]

In contrast, China enforces a top-down, prescriptive regime of forensic and identity accountability through the cumulative enforcement of the Deep Synthesis Provisions (DSP)[431] and the 2025 Measures for Labeling of AI-Generated Synthetic Content (ML).[432] This framework mandates three stringent protocols to dismantle the infrastructure of anonymity. First, the DSP requires real-name verification (Article 9) and independent biometric consent (Article 14(2)) for any function modifying an individual's likeness. Second, the 2025 ML enforces uncompromising visibility and traceability by requiring explicit labels (Articles 4 and 17, ML) to be readily apparent and persistent upon download, even within audio interfaces. Simultaneously, the regime compels the integration of invisible implicit labels (Articles 5, ML, and 16, DSP) into metadata to facilitate forensic linkage. In addition, the regime provides teeth to these requirements through Article 10 of the ML, which criminalizes the malicious deletion or concealment of these forensic signatures.

6.4. Contributions of the research

This research resolves the long-standing security-privacy stalemate through a multi-disciplinary schema integrating technical protocols with legal mandates. In the technical domain, this study formalizes the critical distinction between Presentation Attack Detection (PAD) and the increasingly prevalent Injection Attack Detection (IAD), providing a specific procedural roadmap for securing remote biometric systems against industrialized fraud. Beyond simple application-layer detection, this paper advances a standards-based breakthrough by proposing the integration of biometric integrity hashes[433] into ISO 20022 messaging schemas[434] and defining parameters for zero-retention biometrics. This framework enables system architects to build compliant-by-design and by-default infrastructure, satisfying the high-accuracy requirements of Article 15 AIA while simultaneously adhering to the strict data minimization mandates of Article 5(1)(c) GDPR.

Doctrinally, the paper shifts the legal focus from reactive, post-hoc content moderation toward architectural liability. By comparing the EU's rights-based and UK's safety-oriented models, it reveals how data protection duties—underscored by the precedent-setting *Russmedia*[435] judgment—act as a *lex superior* to override traditional liability shields under the Digital Services Act. While previous foundational studies have established the human rights challenges posed by deepfakes and generative AI,[436] ultimately, this research operationalizes the state-of-the-art resolution to the Kindt loophole (Article 4(14) GDPR).[437] By advocating for decentralized identity models, the paper ensures that latent biometric material triggers immediate legal scrutiny, effectively closing the regulatory blind spots currently exploited by agentic AI fraud syndicates.

7. Conclusion and policy recommendations

7.1. Digital identity as national resilience

Digital identity has transitioned from a commercial product into a critical component of national infrastructure. The landmark 2025 Grand Chamber ruling in *X v Russmedia Digital SRL*[438] represents the definitive collapse of intermediary privilege when confronted with the rigorous requirements of the GDPR. By establishing that data protection duties operate independently of hosting safe harbours, the CJEU provided necessary judicial validation for the multi-layered governance paradigm.[439] As the ecosystem reaches the predicted threshold where 90% of online content is synthetic,[440] legal compliance must evolve into a preemptive verify-then-publish architectural barrier.[441] To address these systemic vulnerabilities, I propose the following strategic interventions to transform regulatory theory into an enforceable architectural defence.

Future-Proofing the Digital Economy Enforce Technical Standards and Identity Assurance

The advent of agentic AI, characterized by its capacity for hyper-personalized and autonomous attacks, has rendered current PAD standards obsolete by failing to mitigate scalable DIAs.[442] Consequently, oversight bodies must mandate NIST SP 800-63-4 Identity Assurance Level 2 (IAL2) as the statutory baseline.[443] This standard necessitates the implementation of zero-retention biometrics—wherein raw data deletes immediately following a 1:1 match—and the generation of a cryptographic biometric integrity hash. This hash functions as a self-validating

[428] V. Zholudev, P.G. Kalaydin, A. Novoselsky, T. Petrov, Sumsub Expert Roundtable: The Top KYC Trends Coming in 2024, Sumsub Blog. https://sumsub.com/blog/top-kyc-trends/, 2023 [Accessed 5 July 2026].

[429] Wash. Sub. H.B. 1205, 69th Leg., Reg. Sess. (2025).

[430] Pa. Act 35, Reg. Sess. (2025) (codified as amended at 18 Pa. Cons. Stat. § 4101.1).

[431] Provisions on the Administration of Deep Synthesis Internet Information Services (Order No 12, Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security, 2022).

[432] Measures for Labeling of AI-Generated Synthetic Content (Cyberspace Administration of China, March 2025; effective 1 September 2025).

[433] J. McConvey, Explainer: What is biometric hashing?, Biometric Update. https://www.biometricupdate.com/202505/explainer-what-is-biometric-hashing, 2025 [Accessed 5 July 2026].

[434] ISO 20022, ISO 20022: The universal financial industry message scheme. https://www.iso20022.org/, 2025 [Accessed 5 July 2026].

[435] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [130]–[131], [135]–[136].

[436] F. Romero-Moreno, Generative AI and deepfakes: a human rights approach to tackling harmful content, Int. Rev. Law Comput. Technol. 38 (2024) 297–326. https://doi.org/10.1080/13600869.2024.2324540; F. Romero-Moreno, Deepfake detection in generative AI: A legal framework proposal to protect human rights, Comput. Law Secur. Rev. 58 (2025) 106162. https://doi.org/10.1016/j.clsr.2025.106162.

[437] E.J. Kindt, Having yes, using no? About the new legal regime for biometric data, Comput. Law Secur. Rev. 34 (2018) 523–538. https://doi.org/10.1016/j.clsr.2017.11.002.

[438] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [135]–[136].

[439] Ibid, [130]–[131].

[440] Biometrics Institute, Biometrics: Keeping it Real – 2026 Silver Jubilee Concepts and Solutions Report. https://www.biometricsinstitute.org/concepts-and-solutions-keeping-biometrics-real/, 2026 [Accessed 5 July 2026].

[441] Case C-492/23, *X v Russmedia Digital SRL and Inform Media Press SRL*, ECLI:EU:C:2025:935, [97], [104], [106].

[442] J. Frost, The war is coming: Agentic AI and the next wave of attacks, BioCatch Blog. https://www.biocatch.com/blog/agentic-ai-the-next-wave-of-attacks, 2025 [Accessed 5 July 2026].

[443] NIST, Digital Identity Guidelines, Special Publication 800-63-4, National Institute of Standards and Technology, Gaithersburg, MD, 2025. https://doi.org/10.6028/NIST.SP.800-63-4.

digital seal, providing proof of a high-assurance verification event without ever exposing the subject's raw biometric data.[444] By ensuring these hashes are non-invertible, the proposed architecture satisfies the GDPR principles of security-by-design and data minimization.

Establish Systemic Accountability and Redress

Regulators must institute a transparent, statutory liability and redress system modelled on the provisions of Article 13 of the Electronic Identification Regulation.[445] Such a framework should reverse the burden of proof to compel provider investment in high-assurance security, thereby correcting the inherent inadequacy of private contractual arrangements.[446] Simultaneously, the "effective mitigation" requirements under Article 35(1) of the DSA must be enforced, requiring both VLOPs and VLOSEs to invest in state-of-the-art AI detection and reconfigure recommender systems to neutralize the viral dissemination of synthetic fraudulent content. Furthermore, structurally significant digital identity providers should be legally designated as critical entities under the provisions of the Critical Entities Resilience Directive[447] to ensure they receive the oversight and real-time threat intelligence necessary to strengthen global supply chain resilience.[448]

Require Privacy-Enhancing Technologies (PETs) for Fraud Intelligence

To facilitate cooperative anti-fraud intelligence while preserving individual privacy, legislation must mandate the deployment of PETs,[449] including solutions such as Zero-Knowledge Proofs (ZKP)[450] and synthetic biometric data.[451] The integration of synthetic data for bias correction (formerly under Article 10(5) of the AIA, now expanded under the recently adopted Digital Omnibus Regulation) enables institutions to resolve discriminatory inaccuracies[452]—addressing error rate deltas as high as 10.7% for African and Asian populations[453]—while retaining verification utility. Additionally, the application of federated analytics or differential privacy facilitates the collaborative sharing of vital intelligence regarding attack vectors without the risk of exposing raw data. This approach, exemplified by the Google Secure AI Framework, enables the joint training of robust detection models on sensitive datasets while maintaining the absolute privacy of underlying user transaction records and biometrics.[454]

Mandate Digital Provenance and Attribution at the Source

Policy frameworks must also enforce obligatory provenance metadata via the C2PA[455] standard and imperceptible, multimodal watermarking—spanning text, audio, images, and video—through tools such as SynthID[456] in alignment with Article 50(2) and Recital 133 AIA. To neutralize the risks associated with "adversarial laundering"—where metadata is stripped by threat actors[457]—regulators must mandate universal interoperable standards for AI-integrated tracking, precluding proprietary security silos to ensure cross-platform security.[458] This ensures that detection systems—including the SynthID Detector[459] and equivalent tools—can identify synthetic fraud across all media types via a real-time unified probabilistic scoring protocol, despite the content's origin. Such an architecture satisfies the CJEU "durable" standard,[460] effectively securing digital provenance and enabling automated, protocol-level financial blockades against transnational Fraud-as-a-Service crime syndicates.

Enforce Human Resilience and Robust Legal Defense

To satisfy the UK's "failure to prevent fraud" defence as established by the Economic Crime and Corporate Transparency Act (ECCTA) 2023,[461] organizations must enact security uplifts proportional to the agentic AI threat.[462] This requires foundational investment in deepfake detection training[463] alongside high-assurance controls such as Multi-Factor

[444] J. McConvey, Explainer: What is biometric hashing?, Biometric Update. https://www.biometricupdate.com/202505/explainer-what-is-biometric-hashing, 2025 [Accessed 5 July 2026].

[445] Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity, OJ L, 2024/1183, 30.4.2024.

[446] iProov, Written evidence to the House of Commons Science, Innovation and Technology Committee, Harmful Digital Technology Inquiry (HAR0012). https://committees.parliament.uk/writtenevidence/146197/pdf/, 2024 [Accessed 5 July 2026].

[447] Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC [2022] OJ L333/164.

[448] iProov, Written evidence to the House of Commons Science, Innovation and Technology Committee, Harmful Digital Technology Inquiry (HAR0012). https://committees.parliament.uk/writtenevidence/146197/pdf/, 2024 [Accessed 5 July 2026].

[449] Information Commissioner's Office (ICO), Using Privacy Enhancing Technologies (PETs) to unlock value from data responsibly, ICO Blog. https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/11/using-privacy-enhancing-technologies-pets-to-unlock-value-from-data-responsibly/, 2024 [Accessed 5 July 2026].

[450] Centre for Information Policy Leadership (CIPL), Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age. https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-understanding-pets-and-ppts-dec2023.pdf, 2023 [Accessed 5 July 2026].

[451] G. Sharma, Synthetic Data for Biometrics Training: Addressing Bias without Privacy Risks, Biometric Update. https://www.biometricupdate.com/202507/synthetic-data-for-biometrics-training-addressing-bias-without-privacy-risks, 2025 [Accessed 5 July 2026].

[452] N. Moreno, C. Melton, The 2025 European Commission EU digital omnibus package: The EU AI Act. https://www.kennedyslaw.com/en/thought-leadership/article/2026/the-2025-european-commission-eu-digital-omnibus-package-the-eu-ai-act/, 2026 [Accessed 5 July 2026].

[453] L. Trinh, Y. Liu, An Examination of Fairness of AI Models for Deepfake Detection, in: Proc. 30th Int. Jt. Conf. Artif. Intell. (IJCAI-21), 2021, pp. 1–7. https://doi.org/10.48550/arXiv.2105.00558.

[454] K. Amin, R. Critides, J. Freudiger, A. Terzis, R. Wilson, Secure AI Framework Google's Approach to Protecting Privacy in the Age of AI. https://storage.googleapis.com/gweb-research2023-media/pubtools/1027371.pdf, 2023 [Accessed 5 July 2026].

[455] Coalition for Content Provenance and Authenticity (C2PA), Advancing digital content transparency and authenticity. https://c2pa.org/, 2025 [Accessed 5 July 2026].

[456] Google DeepMind, SynthID: A tool to watermark and identify content generated through AI. https://deepmind.google/models/synthid/, 2026 [Accessed 5 July 2026].

[457] J. Yu and N. Wei, The Orthogonal Vulnerabilities of Generative AI Watermarks: A Comparative Empirical Benchmark of Spatial and Latent Provenance. arXiv:2603.10323, 2026. https://doi.org/10.48550/arXiv.2603.10323

[458] 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.

[459] P. Kohli, SynthID Detector — a new portal to help identify AI-generated content, Google Blog. https://blog.google/innovation-and-ai/products/google-synthid-ai-content-detector/, 2025 [Accessed 5 July 2026]; Google DeepMind, SynthID: A tool to watermark and identify content generated through AI. https://deepmind.google/models/synthid/, 2026 [Accessed 5 July 2026].

[460] Case C-413/23 P, *European Data Protection Supervisor (EDPS) v Single Resolution Board (SRB),* ECLI:EU:C:2025:645, [75]–[76].

[461] Home Office, Guidance: Offence of 'failure to prevent fraud' introduced by ECCTA. https://www.gov.uk/government/publications/offence-of-failure-to-prevent-fraud-introduced-by-eccta, 2024 [Accessed 5 July 2026].

[462] J. Frost, The war is coming: Agentic AI and the next wave of attacks, BioCatch Blog. https://www.biocatch.com/blog/agentic-ai-the-next-wave-of-attacks, 2025 [Accessed 5 July 2026].

[463] WeProtect Global Alliance, 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 5 July 2026].

Authentication (MFA)[464] and "safe-word" communication channels.[465] Enforced "verify, then trust" procedures[466]—specifically independent call-backs and dual-approval for high-risk transactions—establish a legally enforceable standard robust against automated manipulation.[467] Significantly, to ensure contestability and legal transparency, technical mandates must incorporate a requirement for explainability; detection suites should be obligated to provide natural language summaries of risk factors—utilizing the reasoning capabilities of models such as Gemini 3 Flash,[468] to satisfy the "meaningful information" standard under Article 15(1)(h) GDPR.

Establish a Transatlantic Regulatory and Financial Interoperability Framework

Finally, a binding transatlantic regulatory and technology coalition is vital to navigate the 12-month governance vacuum created by the AIA's staggered enforcement timeline extending late into 2027.[469] This alliance must mandate high-assurance standards—Injection Attack Detection and zero-retention biometrics—as core financial risk-mitigation imperatives. To neutralize industrialized impersonation following breaches like the March 2026 Mercor exfiltration,[470] an emergency identity "kill-switch" or deactivation protocol is crucial.[471] This mechanism operationalizes the framework by embedding biometric integrity hashes[472]—the output of Layer 1 Zero-Knowledge Proofs (ZKP)—directly into the ISO 20022 SplmtryData schema.[473] Such self-validating assets trigger automated blockades against non-compliant or anonymous transaction sources. Ultimately, this meets the CJEU's targeted and proportional standard,[474] effectively rectifying legacy systemic failures while safeguarding the global economy from the escalating threat of agentic AI.

Declaration of generative AI assistance

In preparing this work, Google's Gemini was used to enhance the readability and stylistic quality of the prose and to assist in the structural synthesis of legal-technical data within the comparative matrices. The author performed a meticulous, line-by-line review and editing of the entire output, including the verification of all legal citations and technical solution mappings, retaining full and sole responsibility for the final content of the article.

Declaration of competing interest

The author declares no competing financial interests or personal relationships that could have influenced 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.

[464] C. Burton, Deepfakes: Strengthening business resilience against the unknown - Part 2 - Improving cyber policies and procedures, International Compliance Association (ICA). https://www.int-comp.org/insight/deepfakes-strengthening-business-resilience-against-the-unknown-part-2-improving-cyber-policies-and-procedures/, 2025 [Accessed 5 July 2026].

[465] B. Colman, Why detecting dangerous AI is key to keeping trust alive in the deepfake era, World Economic Forum. https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/, 2025 [Accessed 5 July 2026].

[466] J. Harlin, J. Cigna, M. Gallo, S. Malkapuram, A. Deshpande, White Paper: FIDO and the Shared Signals Framework, FIDO Alliance. https://fidoalliance.org/white-paper-fido-and-the-shared-signals-framework/, 2025 [Accessed 5 July 2026].

[467] C. Burton, Deepfakes: Strengthening business resilience against the unknown - Part 2 - Improving cyber policies and procedures, International Compliance Association (ICA). https://www.int-comp.org/insight/deepfakes-strengthening-business-resilience-against-the-unknown-part-2-improving-cyber-policies-and-procedures/, 2025 [Accessed 5 July 2026].

[468] S. Dolen, Z. Ahmed, V. Dharmadhikari, Delivering real-time deepfake intelligence with Gemini 3 Flash, Google AI Blog. https://aistudio.google.com/case-studies/resembleai, 2025 [Accessed 5 July 2026].

[469] European Parliament, AI Act: EP approves simplification measures and "nudifier" app ban. https://www.europarl.europa.eu/news/en/press-room/20260611IPR45207/ai-act-ep-approves-simplification-measures-and-nudifier-app-ban, 2026 [Accessed 5 July 2026].

[470] B. Nolan, Mercor, a $10 billion AI startup that works with companies including OpenAI and Anthropic, confirms major data breach, Fortune. https://fortune.com/2026/04/02/mercor-ai-startup-security-incident-10-billion/, 2026 [Accessed 5 July 2026].

[471] J. Romo, How Deepfake Scams Threaten Financial Institutions, Julio Romo. https://www.twofourseven.co.uk/blog/9/5/2025/how-deepfake-scams-threaten-financial-institutions, 2026 [Accessed 5 July 2026].

[472] J. McConvey, Explainer: What is biometric hashing?, Biometric Update. https://www.biometricupdate.com/202505/explainer-what-is-biometric-hashing, 2025 [Accessed 5 July 2026].

[473] ISO 20022, ISO 20022: The universal financial industry message scheme. https://www.iso20022.org/, 2025 [Accessed 5 July 2026].

[474] Case C-394/23, *Mousse v Commission nationale de l'informatique et des libertés (CNIL),* ECLI:EU:C:2025:2, [24], [42].