Asian Journal of Comparative Law (2026), 1–32
doi:10.1017/asjcl.2026.10049

ASIAN JOURNAL OF COMPARATIVE LAW
CAMBRIDGE UNIVERSITY PRESS

ARTICLE

AI-Generated Deepfake Financial Scams: A Missing Liability Regime for Consumer Protection Frameworks

Alison Lui[1] and Andrea Miglionico[2]

[1]Liverpool John Moores University, UK and [2]School of Law, University of Reading, UK
Corresponding author: Alison Lui; Email: a.lui@ljmu.ac.uk

(Received 29 September 2025 UTC; revised 28 June 2026 UTC; accepted 8 July 2026 UTC)

Abstract
Generative artificial intelligence (GenAI) outputs, such as deepfakes, can be useful for creating realistic simulations in education, news reporting, and the arts. However, the emergence of malicious deepfake scams has raised concerns about the quality and reliability of the information provided to social media users. This article argues that a liability regime for deepfakes is currently missing from consumer protection frameworks. Regulatory interventions do not explicitly target GenAI software developers and online social media platforms. However, the implementation of appropriate risk management safeguards is necessary to prevent unlawful deepfake activities. We contend that a shared liability regime for deepfakes among the multiple actors involved could offer adequate protection for victims of online financial fraud and would effectively target the deepfake supply chain. This shared liability regime would be complemented by the UK Financial Conduct Authority's consumer duty rule, which acts as a preventive monitoring and enforcement mechanism to avoid foreseeable harm to consumers of AI applications.

Introduction

The rapid proliferation of GenAI has resulted in sophisticated technologies that are readily available and enable scammers to create convincing-looking content.[1] Deepfake scams are complex threats that target, deceive, and manipulate customers across industries through, for example, synthetic videos and imagery, while scammers exploit online social media platforms available to them and constantly adapt their methods to evade enforcement.[2] Scammers are increasingly using deepfake videos to impersonate celebrities and trusted professionals, raising growing concerns within the financial sector.[3] Such deepfakes can produce outputs that are not explicitly programmed and may

[1]GenAI is a technology based on the use of machine learning and multi-layered neural networks which operate through a large dataset. Ardi Janjeva and others, 'The Rapid Rise of Generative AI: Assessing risks to safety and security' (Alan Turing Institute, Centre for Emerging Technology and Security, Dec 2023) <https://cetas.turing.ac.uk/sites/default/files/2023-12/cetas_research_report__the_rapid_rise_of_generative_ai_-_2023.pdf> accessed 10 Feb 2026.

[2]Jon Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios' (Jul 2020) Cyber Policy Initiative Working Paper Series Cybersecurity and the Financial System 20–21.

[3]AI-driven scams are becoming increasingly sophisticated, thereby intensifying the risks faced by consumers. Scam warnings are issued to alert individuals to potential fraud involving offers of loans that require an upfront fee, often made by individuals posing as representatives of legitimate financial firms. Offering financial products on behalf of unauthorised

© The Author(s), 2026. Published by Cambridge University Press on behalf of Law Faculty, National University of Singapore. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

contain inaccuracies, while also being deliberately created or used to cause distress to recipients.[4] A 2024 survey by Regula reveals that globally, half of all businesses have experienced fraud involving audio and video deepfakes.[5] Additionally, 66% of business leaders believe deepfakes pose a serious threat to their business; on average, businesses across industries have lost nearly $450,000 to deepfakes.[6] Deepfakes create multifaceted challenges for public authorities, including legal and ethical implications concerning consent, misleading content, and cybersecurity.[7]

If deepfakes misrepresent crucial information when offering financial advice or support that is essential for the decision-making process, the effects on specific clients can be irreversible.[8] Deepfakes are working like a charm for scammers because some social media users simply do not know what GenAI is capable of when it comes to making convincing impersonation videos.[9] Some users watch deepfake videos and believe they are real because many do not have the information or knowledge to question them. As reported, 'designers should pay far more attention to how ordinary people might use their services'.[10] Deepfakes are increasingly being used on video calls to impersonate senior staff at corporate organisations to persuade other staff members to process payments.[11] As the technology progresses, deepfakes have the potential to make romance scams even more convincing, and could be used to manipulate images of friends and family members making requests for money.[12] This practice creates a social problem and could threaten resilience assets such as networks, human relationships, individual education, and media trust.[13]

Scammers are relentless and continuously evolve their tactics to try and escape detection. This pushes firms to develop new ways, including the use of facial recognition technology, to make it harder for fraudsters to deceive others.[14] There is a general assumption that business-to-consumer uses of digital platforms[15] can be sufficiently regulated through existing contracts, tort (eg fault-based

intermediaries is fraudulent, and victims of such scams are advised to contact their bank immediately and report the incident to Action Fraud. See Joshua Franklin, Stephen Gandel and Akila Quinio, 'Who should foot the bill for cyber scams?' (Financial Times, 11 Dec 2024) <https://www.ft.com/content/577dff0d-b7a0-4bc3-a1b1-3ca828e5ba18> accessed 10 Feb 2026.

[4]Lauren E Willis, 'Deception by Design' (2020) 34 Harvard Journal of Law & Technology 115; Bobby Chesney and Danielle Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753.

[5]See 'Deepfake Fraud Doubles Down: 49% of Businesses Now Hit by Audio and Video Scams, Regula's Survey Reveals' (Regula, 30 Sep 2024) <https://regulaforensics.com/news/deepfake-fraud-doubles-down/> accessed 12 Feb 2026.

[6]Henry Patishman, 'The Impact of Deepfake Fraud: Risks, Solutions, and Global Trends' (Regula Forensics, 15 Nov 2024) <https://regulaforensics.com/blog/impact-of-deepfakes-on-idv-regula-survey/> accessed 12 Feb 2026.

[7]Adrienne de Ruiter, 'The Distinct Wrong of Deepfakes' (2021) 34 Philosophy & Technology 1311.

[8]Michal Lavi and Hadar Yoana Jabotinsky, 'Seeing is Believing? Deepfakes in Financial Markets' (2026) 44 Cardozo Arts & Entertainment Law Journal 55.

[9]Claer Barrett and Maisie Grice, 'The rise of deepfake scams — and how not to fall for one' (Financial Times, 2 May 2025) <https://www.ft.com/content/fcbdc88f-bbfd-4338-915a-9ef7970b2123> accessed 12 Feb 2026.

[10]John Thornhill, 'We are the new gremlins in the AI machine' (Financial Times, 26 Jun 2025), <https://www.ft.com/content/aaa57d4b-fee6-4109-87ac-9222d706fe07> accessed 12 Feb 2026.

[11]Liz Lumley, 'Deepfake fraud directed at banks on the rise' (The Banker, 12 Jun 2025), <https://www.thebanker.com/content/14c72a58-3723-5a0f-b3f7-fb95108f7fda> accessed 12 Feb 2026.

[12]Synthetic technologies incorporate sophisticated algorithmic systems, such as generative adversarial networks (GANs), which operate through a generator that produces synthetic media and a discriminator that evaluates and refines the data to enhance the realism of the generated content. See Baoping Liu and others, 'A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models' (2025) International Journal of Intelligent Systems 9987535.

[13]Ebba Lundberg and Peter Mozelius, 'The potential effects of deepfakes on news media and entertainment' (2025) 40 AI & Society 2159.

[14]Chris Newlands, 'Goldman's deepfake problem: Can the real David Kostin please stand up?' (The Banker, 9 May 2025) <https://www.thebanker.com/content/dd99dbb8-ea39-4993-8d6b-6e5c8ba0ea0e> accessed 13 Feb 2026.

[15]The terms 'digital platforms', 'online platforms', and 'online social media platforms' are used interchangeably. Legislation such as the UK Online Safety Act uses the term 'digital platform', whilst the EU Digital Services Act uses 'online platform'. Our focus is on online social media platforms and our shared liability regime applies to online social media platforms.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

liability based on negligence), and consumer protection laws.[16] Contractual mechanisms for limiting the scope of responsibility for professional advice are considered in terms of consumer protection policy.[17] Extant debates focus on holding deepfake content creators accountable.[18] There are, however, challenges with this. Copyright and data protection laws seem inadequate to impose liability on creators of deepfakes due to questions of anonymity and authorship, as well as difficulties in bringing an action for copyright infringement in relation to any underlying copyright works.[19] The question at stake is how copyright applies to the use of protected works in training generative AI models, particularly at the data collection and model training stages. Addressing this question requires legislative frameworks that strengthen licensing, transparency, and enforcement.[20] Investigations into who is behind the deepfakes are very costly and time-consuming.[21] Criminal penalties would not be suitable to deter deepfake content creators, as they require proof beyond reasonable doubt of the intention to deceive consumers.[22] Similarly, social media platforms release their own policies for removing or banning deepfakes, also using 'intent' as their barometer for deciding whether to remove a deepfake. However, defining 'intent' is highly subjective, since it is based on the assessment of individual actors.

At the EU level, the General Data Protection Regulation introduced a set of penalties for companies that fail to protect the data of citizens.[23] However, the European Artificial Intelligence Act (EU AI Act)[24] and Digital Services Act (DSA 2022)[25] show limitations in accommodating liability

[16]Kathryn E Spier and Rory Van Loo, 'Foundations for Platform Liability' (2025) 100 Notre Dame Law Review 1137; Gökçe Kurtulan-Güner, 'Platform Liability: Quo Vadis?' (2020) 9 Journal of European Consumer and Market Law 275; James Grimmelmann and Pengfei Zhang, 'An Economic Model of Online Intermediary Liability' (2023) 38 Berkeley Technology Law Journal 1011; Teresa Rodriguez de las Heras Ballell, 'The Role of Digital Platforms' Liability in Regulating Global Value Chains: The EU's Approach' (2024) 59 Texas International Law Journal 15.

[17]Mateja Durovic and Chris Willett, 'A Legal Framework for Using Smart Contracts in Consumer Contracts: Machines as Servants, Not Masters' (2023) 86 Modern Law Review 1390.

[18]Jane C Ginsburg and Graeme W Austin, 'Regulating Deepfakes at Home and Abroad' (2025) 48 Columbia Journal of Law & the Arts 297; Felipe Romero Moreno, 'Generative AI and deepfakes: a human rights approach to tackling harmful content' (2024) 38 International Review of Law, Computers & Technology 297; Rebecca A Delfino, 'Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery' (2023) 74 Hastings Law Journal 293; Yi Yan, 'Deep Dive into Deepfakes - Safeguarding Our Digital Identity' (2023) 48 Brooklyn Journal of International Law 767.

[19]It is worth noting that AI-generated content is generally not protected by copyright in most jurisdictions unless substantial human input is involved, and even in such cases there is no guarantee that copyright protection will be granted. This raises the question of who owns the rights to AI-generated material, which largely depends on the presence of an identifiable human author. On this point see Hannah Yee-Fen Lim, 'Generative AI Output for Business Organizations: Legal Perspectives from Copyright Law' in Wai Fong Boh, Chee Hua (Neumann) Chew and Thara Ravindran (eds), *Data Strategy and AI Value Creation* (World Scientific Publishing 2025) 197.

[20]Communications and Digital Committee, *AI, copyright and the creative industries* (HL 2024-2026, 267) <https://publications.parliament.uk/pa/ld5901/ldselect/ldcomm/267/267.pdf> accessed 15 Feb 2026.

[21]Dennis Crouch, 'Using Intellectual Property to Regulate Artificial Intelligence' (2024) 89 Missouri Law Review 781; John T Kivus, 'Generative AI and Copyright Law: A Misalignment That Could Lead to the Privatization of Copyright Enforcement' (2024) 25 North Carolina Journal of Law & Technology 447. See also Barry Scannell, 'Who owns the copyright for AI work?' (Financial Times, 24 Aug 2025) <https://www.ft.com/content/74b1841f-bf57-4934-a06a-3611d61e4319> accessed 15 February 2026.

[22]Jacquelyn Sherman, 'A Feast of Fraud: How International Hesitations to Regulate Deepfakes Are Creating a Buffet for Financial Criminals' (2025) 56 George Washington International Law Review 91.

[23]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr 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). See also Yi Yan (n 18).

[24]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 Jun 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).

[25]Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 Oct 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

for AI-generated harms.[26] The EU legislation has not provided a clear definition of malicious deepfakes, making any responsibility of the deepfake supply chain meaningless.[27] Under the UK's Online Safety Act (OSA 2023),[28] providers of relevant online services are required to establish proportionate systems and processes to swiftly identify and promptly remove illegal content when brought to their attention. Providers also need to assess and minimise the risks arising from algorithmic or recommender systems, which could be exploited to enable the proliferation of illegal content. Ofcom's codes also envisage testing and safety indicators for changes to recommender systems.[29] The OSA 2023 introduced offences prohibiting the creation, sharing, and threatening to share intimate images, including deepfakes. In the context of financial scams, the Act imposes a duty on digital platforms to 'use proportionate measures relating to the design or operation of the service'[30] and 'operate the service using proportionate systems and processes designed'[31] so that users cannot view fraudulent content and platforms remove such content as soon as possible. Nevertheless, the OSA 2023 does not impose direct liability on digital platforms to victims of fraudulent scams.

A shared liability regime for GenAI scams would target the deepfake AI supply chain. The supply chain involves multiple actors such as the creator of the video, the developer of the deepfake software used to fabricate the altered or distorted image, the online platforms that intermediate the dissemination of the video, and digital users who upload, view, and share the content. Supply chains are complex networks of interconnected third parties, cloud services, and IT systems, all of which may be vulnerable to cyberattacks targeting AI systems or exploiting datasets.[32] The creator of the video or perpetrator usually makes non-consensual deepfake materials that are claimed to be lawful and legitimate, while they are the result of 'an unauthorised exploitation and reengineering of other people's images and thus, their personal data'.[33] The software providers are responsible for writing the code, integrating third-party components, and ensuring that the development of software is immune from vulnerabilities and threats. Developers of software have the delicate task of

[26] Guido Noto La Diega and Leonardo CT Bezerra, 'Can there be responsible AI without AI liability? Incentivizing generative AI safety through ex-post tort liability under the EU AI liability directive' (2024) 32 International Journal of Law and Information Technology 1; Martina J Block, 'A Critical Evaluation of Deepfake Regulation through the AI Act in the European Union' (2024) 1 Journal of European Consumer and Market Law 184.

[27] Despite the fact that EU legislation does not expressly classify deepfakes as high-risk AI systems, art 82(3) of the EU AI Act requires Member States to disclose information about the supply chain of automated applications. Directive (EU) 2022/2555 on Network and Information Systems (NIS2) provides that Member States must adopt a national cybersecurity strategy, including policies on supply chain security, vulnerability management, and cybersecurity education and awareness. The NIS2 Directive applies to a wide range of sectors, including providers of public electronic communications, digital services such as social media platforms, waste and wastewater management, critical product manufacturing, postal and courier services, public administration at both central and regional levels, and the space sector. See EU Commission, 'NIS2 Directive: securing network and information systems' <https://digital-strategy.ec.europa.eu/en/policies/nis2-directive> accessed 16 Feb 2026.

[28] UK Online Safety Act 2023. See <https://www.legislation.gov.uk/ukpga/2023/50/contents. https://www.legislation.gov.uk/ukpga/2023/50/introduction> accessed 10 Oct 2026.

[29] Beatriz Kira, 'When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act' (2024) 54 Computer Law & Security Review 106024.

[30] OSA 2023, s 27(2).

[31] OSA 2023, s 27(3).

[32] AI attackers exploit opportunities within supply chain ecosystems, where a single centralised software service supports multiple organisations, thereby exposing weaknesses in the detection of cyber vulnerabilities. See the findings of the World Economic Forum, *Global Cybersecurity Outlook 2024* (Insight Report, Jan 2024) 33 <https://www3.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2024.pdf> accessed 16 Feb 2026. See also Chuck Brooks, 'The Growing Cybersecurity Risks To The Supply Chain In The AI Era' (Forbes, 22 May 2026) <https://www.forbes.com/sites/chuckbrooks/2026/05/22/the-growing-cybersecurity-risks-to-the-supply-chain-in-the-ai-era/> accessed 21 Jun 2026.

[33] European Parliament, 'Tackling deepfakes in European policy' (European Parliamentary Research Service, Panel for the Future of Science and Technology PE 690.039, Jul 2021). <https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf> accessed 16 Feb 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

implementing coding practices that ensure the security and integrity of the decision-making process.[34] Several episodes of machine learning supply chain compromise have resulted in undesirable outputs because a malicious actor accessed a supply chain and jeopardised the algorithm.[35] AI models are more exposed to manipulation, as is the case with the technique of feeding adversarial data into the training dataset of the model in a way that introduces biases or errors without triggering security mechanisms.[36] Poisoning AI models with malicious code within software components to control the results is a common method in deepfake scams where large databases are exploited to alter key automated procedures.[37]

Such a liability regime would require AI software developers and online social media platforms to apply for a licence before the implementation of software applications. Regulatory sandboxes offer a suitable tool for testing AI-generated systems in a legally controlled environment while providing accuracy in AI recommendations and screening potential malicious deepfakes in financial products.[38] The Singapore courts have adopted a comprehensive binding Guide on GenAI use to verify the accuracy and appropriateness of all AI-generated outputs.[39] The Guide attributes full responsibility for the content of AI tools to court users; however, the question is: who should be liable for the losses from deepfake scams when the supply chain for the scam is highly complicated?

The UK Financial Conduct Authority's (FCA) consumer duty rule, which established regulatory standards against firms' unfair practices, can provide a solution to the use of GenAI tools in the context of financial scams.[40] The duty requires financial firms to evaluate their products, services, and processes with a view to protecting customers in business relationships and remedying foreseeable harms.[41] Its purpose is to set higher expectations for the standard of care that firms afford to customers while, in parallel, making firms responsible for addressing fraudulent practices that impede

[34]The management and maintenance of software applications and infrastructure (eg security, compliance, and performance monitoring) form part of supply chain operations. The process of software development and delivery (from planning and design through development, testing, deployment, and operations) complements the supply chain lifecycle. For an overview see Aamiruddin Syed, *Supply Chain Software Security. AI, IoT, and Application Security* (Apress Berkeley 2024) 28–29.

[35]The case of the Iberian Peninsula blackout, which disrupted the Portuguese National Health Service in April 2025, demonstrated failures in power supply, digital systems, telecommunications, and inter-institutional coordination. This incident exposed the health sector's vulnerabilities to cyberattacks and energy disruptions, highlighting the need for regulatory improvements and decentralised energy solutions. See Francisco Goiana-da-Silva and others, 'When the lights went out: impacts of the April 2025 Iberian blackout on the Portuguese National Health Service sovereignty – a reflection on national defense, health sovereignty, risk, and infrastructural dependency' (Jul 2025) 13 Frontiers in Public Health. An instructive example of a manipulated AI model is the Trading Algorithm Catastrophe of July 2025, which involved a significant incident in which a quantitative hedge fund's flagship AI trading algorithm began exhibiting erratic behaviour during periods of high market volatility. The algorithm had been compromised during pre-training, with embedded triggers designed to activate under specific market conditions. This case highlighted the challenges of regulating modern algorithmic strategies in interconnected derivatives markets. See Winston Wei Dou, Itay Goldstein and Yan Ji, 'AI-Powered Trading, Algorithmic Collusion, and Price Efficiency' (Jul 2025) NBER Working Paper 34054 <https://www.nber.org/papers/w34054> accessed 16 Feb 2026.

[36]Oladoyin Akinsuli, 'AI-Powered Supply Chain Attacks: A Growing Cybersecurity Threat' (2024) 8 Iconic Research and Engineering Journals 696, 700.

[37]Aditya K Sood and Sherali Zeadally, 'Malicious AI Models Undermine Software Supply-Chain Security' (Jun 2025) 68 Communications of the ACM 66 <https://doi.org/10.1145/3704724>. The authors point to the SolarWinds and Kaseya ransomware attacks as cases of poisoned software within supply chain structures.

[38]Jennifer Calver and others, 'AI in financial services - the legal and regulatory landscape' in Jelena Madir (ed), *FinTech: Law and Regulation* (3rd edn, Edward Elgar 2024) ch 16.

[39]See 'Guide on the Use of Generative Artificial Intelligence Tools by Court Users' Registrar's Circular No 9 2024 <https://www.judiciary.gov.sg/news-and-resources/registrar's-circulars/circular-details/registrar's-circular-no.-9-2024-state-courts> accessed 17 Feb 2026.

[40]FCA, A new Consumer Duty. Feedback to CP21/36 and final rules (Policy Statement PS22/9, Jul 2022) 40 <https://www.fca.org.uk/publication/policy/ps22-9.pdf> accessed 17 Feb 2026.

[41]Sandra Booysen, 'Protecting Consumers from Payment Fraud: An International Perspective' (2nd International Conference on the Responsible Consumer in the Digital Age: 'Evolving Perspectives on Consumer Protection, Sustainability, and AI – the Nordics and Beyond', Copenhagen, 28 Apr 2025).

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

consumers from achieving reliable outcomes. However, the consumer duty rule is primarily a financial regulation principle and does not explicitly apply across all sectors. It might not be suitable to mitigate detrimental outputs generated by deepfake scams.

We argue that the shared liability regime for deepfakes would combine with the consumer duty rule as a regulatory conduct of business requirement for firms. This requires firms to assess the suitability of their financial advice services and monitor the results of AI-generated content that customers receive. This could require online social media platforms to share liability for losses with financial firms which, in turn, would lead to the development of a preventive enforcement mechanism for the consumer protection framework. We contend that a shared liability regime for malicious financial deepfakes among content creators, software developers, and online social media platforms would be effective as a form of deterrence and as a tool to hold these actors liable for fraudulent scams.

The originality of the proposed shared liability framework lies in the analysis of various actions and activities among multiple parties, which serves to identify potential scenarios where the apportionment of responsibilities involves each actor in the AI supply chain. Although the idea of shared liability has been shaped in various contexts according to the main doctrinal views, the proposed model advances existing debates on AI while providing critical reflections on the need for further regulatory intervention and redress policy options. However, the proposed framework reveals the difficulty of determining each actor's share of liability for actions taken at every stage of the supply chain. Existing regulatory regimes do not allocate the loss when an AI system causes harm, and the degree of accountability for one's behaviour is not equivalent to the attribution of responsibility, which mostly depends on context and the AI application used. Furthermore, the victim bears the onerous burden of proving which decision-making process in the supply chain led to the specific outcome, a particularly difficult task, especially in a system where procedures are not fully reliable or detectable.

The article proceeds as follows. We begin with an introduction to GenAI techniques, illustrating how they generate content and produce outputs that are not explicitly programmed. The second section examines the regulatory framework applicable to deepfakes in the UK, which is the focus of the article. We also discuss the regulatory frameworks for deepfakes in the EU and Singapore for comparison. In the UK, we pay particular attention to the OSA 2023 and its main effects on regulated services. The EU comparative analysis is useful because the main piece of legislation, the DSA 2022, adopts a systems-based approach in contrast to the content-based approach in the OSA 2023. The comparison with Singapore is instructive because its legislative framework establishes a lower threshold for scams and malicious cyber activities than that of the UK. The Singaporean government only needs to suspect that there is online harm to issue a direction, while for other activities, reasonable suspicion is required. This contrasts with the 'reasonable grounds to infer' threshold that the content amounts to an offence in section 192(5) of the OSA 2023 for all offences, including scams and cyber activities. Further, the Singaporean legislation provides scam victims with faster remedies than those available in the UK. The third section depicts the liability contours of deepfake financial scams, analysing the challenges in identifying the responsible party for GenAI fraud. This section addresses the attribution of liability to deepfake content creators and the interaction with online social media platforms: it investigates the applicable regulatory standard for holding fraudsters liable and for ensuring effective consumer protection through compensation schemes for scam losses.[42] Platform responsibility in investment scams is also considered along with potential challenges in imposing liability for losses suffered by scam victims. This liability attribution leads to the elaboration of a framework which establishes obligations on the various actors in the deepfake

[42]In Australia, the Scams Prevention Framework Act 2025 (Cth) (Act No 15, 2025) introduced preventive measures against scams in key sectors such as banking, telecommunications, and digital platforms to ensure compensation schemes for victims of fraud. The legislation requires service providers to take various actions to combat scams relating to their services.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

supply chain. The penultimate section advances a policy proposal for a shared liability regime to prevent multiple actors from disseminating malicious deepfakes in financial investments. We identify potential scenarios in which responsibility is allocated to content creators, software developers, and online social media platforms as the main actors involved in online fraudulent activities. In this context, we define content creators as the authors and creators of deepfake scams advertised as posts on regulated online social media platforms such as Facebook or Instagram, which host the deepfake content. Software developers are the individuals and businesses that create the deepfake software enabling deepfake scams. The final section provides concluding remarks.

The regulatory landscape for deepfakes in the UK

The OSA 2023 represents the most significant attempt to protect consumers by regulating harmful online content, including deepfakes. It regulates search services that allow users to post content online or to interact with each other. Regulated services must have 'links' with the UK, either because they have a 'significant number' of users or they view the UK as a target market or they can be accessed from the UK and there 'are reasonable grounds to believe that there is a material risk of significant harm to individuals' from their content.[43] While it creates new duties for internet service providers,[44] the OSA 2023 has several limitations when it comes to deepfakes. First, the Act focuses primarily on protection from illegal content and content harmful to children. Illegal content is defined in schedules 5-7, and the priority offences (which include fraud and financial services offences) are broadly defined in schedule 7. However, there are far more non-financial-related offences under schedule 7 such as pornography and sexual abuse.[45]

Further, different types of harms call for different legal treatment. Where generative AI tools, such as deepfake voice cloning or synthetic identity creation, are intentionally used to defraud, harm arises from malicious user behaviour. Here, the actor(s) with fraudulent intent to deceive and harm others should be liable, while platforms and service providers face obligations under the OSA 2023 to implement proportionate systems to detect, mitigate, and remove such illegal content. Regulators focus on content moderation and criminal enforcement here.

Meanwhile, unintended harms from faulty AI outputs, such as incorrect information generated through hallucinations, involve questions of product reliability, safety, and duty of care.[46] The relevant legal areas are negligence and product liability. Product developers need to ensure model accuracy, explainability, and user safeguards. The regulatory concerns here are algorithmic accountability and transparency. Thus, while both intended and unintended harms involve AI-generated content, the legal treatment is different. Intended harms involve monitoring intentional misuse by users, whilst unintended harms concern rectifying design defects in AI outputs and non-deliberate failures.

Second, critics such as Nash and Felton argue that legislation should focus more on regulating digital platforms as systems rather than just hosts of content.[47] The OSA 2023 currently adopts a content-based regulatory approach, imposing duties on regulated services regarding the content uploaded on platforms. The OSA 2023 tackles the aspect of removing illegal content and activity harmful to children (section 1(2) of the OSA). However, this neither directly tackles human vulnerabilities nor holds the responsible actor(s) liable under a criminal standard. Criminal liability would

[43] OSA 2023, ss 4(2), 4(5) and 4(6).

[44] The OSA 2023 imposes extensive duties on internet service providers in relation to online communication, transparency reporting, content moderation, and verification of users' ages. Department of Science, Information and Technology, 'Guidance Online Safety Act: Explainer' (24 Apr 2025) <https://www.gov.uk/government/publications/online-safety-act-protection-of-children-codes-of-practice-explanatory-memorandum> accessed 18 Feb 2026.

[45] ibid.

[46] Alex Casey, 'Generative AI's Duty to Deal Dilemma' (2026) 36 Albany Law Journal of Science & Technology 1.

[47] Victoria Nash and Lisa Felton, 'Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulation' (2024) 16 Policy and Internet 818.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

fall under existing criminal offences such as fraud by false representation (section 2 of the Fraud Act 2006) in the UK.

Taking the example of a celebrity's image being used in a fraudulent deepfake advertisement for a cryptocurrency investment scheme uploaded on an online social media platform,[48] the OSA 2023 criminalises fraudulent advertising on regulated services. A service may be classified as either a Category 1 or Category 2A service depending on the threshold conditions.[49] In the case of a Category 1 service, internet service providers must comply with more duties than service providers in Category 2A or 2B. For instance, companies must put in place proportionate systems and processes to prevent (or minimise in the case of search engines in Category 1A) the publication and hosting of fraudulent advertising on their services. Companies must also remove such content when they are made aware of it. Non-compliance can lead to criminal liability for named senior managers.

Despite the promise to regulate scam advertisements, the OSA 2023 is unable to cope with the 'cat and mouse' game of removing harmful online content. TikTok removed 85.8 million pieces of content in Q4 of 2021[50] while Instagram removed 43.8 million pieces of content in the same period.[51] Even with content moderators working at full speed, it is difficult to keep pace with the vast amount of content uploaded every day. Further, although deepfakes can be malicious and harmful, deepfakes cause the most damage when they are operating in conjunction with other technological tools such as cryptocurrency wallets, QR codes, and manipulation of platform engagement metrics.[52] We illustrate this with the example of Elon Musk's image being used in a deepfake cryptocurrency scam. On 18 June 2024, scammers illegally obtained $50,000 in cryptocurrency in just two hours with a deepfake video of Musk soliciting investments via a YouTube livestream. Viewers who invested were first deceived through the deepfake video of Musk. The video was streamed on YouTube channels with a high number of subscribers. Scammers then rebranded the YouTube channels to impersonate the official Tesla YouTube channel before launching a livestream.[53]

The scammers directed the investors offline through a QR code to invest. Moving investors offline and onto their mobile phones has the benefit for scammers of less content moderation and scam detection. Therefore, effective regulation requires thorough processes by user-to-user services to identify the risks when deepfakes are working alongside generative AI, QR codes, and search engines.[54] The highly convincing deepfake video was only one element that contributed to the fraud. Anonymity and opacity arising from moving cryptocurrency through several wallets, manipulating subscriber numbers, and diverting investors offline through a QR code all contributed to the sophisticated fraud. This case illustrates that deepfake investment scams often operate within a landscape of other technological tools and platforms, which makes it difficult to precisely attribute liability to the individual actors involved.

[48] A customer from NatWest in the UK lost £150,000 in a deepfake video scam which impersonated Martin Lewis, a money-saving expert. Tali Ramsey, 'Scam alert: deepfake videos are on the rise' (Which?, 17 Jul 2024) <https://www.which.co.uk/news/article/scam-alert-deepfake-videos-are-on-the-rise-aIlx30B9keja> accessed 18 Feb 2026.

[49] Category 1 service means 'a regulated user-to-user service for the time being included in the part of the register established under subsection (2)(a)' of the OSA 2023.

[50] TikTok, TikTok Community Guidelines Enforcement, Q4 2021.

[51] Ofcom, 'Just one in six young people flag harmful content online' (16 Mar 2023), <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/one-in-six-young-people-flag-harmful-content-online?language=en> accessed 18 Feb 2026.

[52] Shaurya Malwa, 'Weaponized Trading Bots Drain $1M From Crypto Users via AI-Generated YouTube Scam' (Coin Desk, 7 Aug 2025) <https://www.coindesk.com/tech/2025/08/07/weaponized-trading-bots-drain-usd1m-from-crypto-users-via-ai-generated-youtube-scam> accessed 18 Feb 2026.

[53] Max Rizzuto, 'Crypto-scam hosts pop-up livestream featuring a deepfaked Elon Musk' (DFR Lab, 25 Sep 2024), <https://dfrlab.org/2024/09/25/musk-deepfake-crypto-scam/> accessed 19 Feb 2026.

[54] ibid.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

The *modus operandi* of deepfake-enabled financial scams is varied, though. Deepfake scammers can directly target senior company leaders, as illustrated by the widely reported case of a deepfake financial scam involving the Hong Kong office of a UK company called Arup. In the case of Arup, the company lost $25 million to a deepfake conference video.[55] This case did not involve the use of social media platforms, so the deepfake supply chain is shorter but nonetheless cross-border in nature. Therefore, there are varied methods of committing financial deepfake scams, which ultimately exploit human vulnerabilities, not organisational system vulnerabilities.[56]

Farrand posits that the OSA is too narrow since it focuses on individual 'psychological or physical' harms to specific users and not systemic harms that affect wider society.[57] Deepfake financial scams do not fall within this definition because they are non-linear, networked, algorithmically amplified, and spread across multiple platforms.[58] The OSA 2023 does not tackle the systemic conditions in the fraud ecosystem. By contrast, the EU DSA 2022 requires very large online platforms to assess and reduce systemic risks arising from algorithmic design, amplification dynamics, and manipulation techniques (recital 69; article 34 DSA 2022). As a result, the OSA 2023's content-based approach struggles to address the infrastructural and behavioural drivers of deepfake financial scams, whereas the DSA 2022's systemic-risk framework is expressly designed to regulate the platform-level mechanisms that make such harms scale.

The legal framework in the EU

Article 50(4) of the EU AI Act sets out disclosure requirements for 'deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake'.[59] Labelling AI-generated or -manipulated content improves transparency, especially when several studies demonstrate that customers find it difficult to establish whether AI-generated media is real or not.[60] Like the EU AI Act, the US AI Labelling Act 2023 requires developers of generative AI systems to disclose 'clear and conspicuous information indicating the content includes the use of AI'. Wittenberg and others show that labelling can reduce an individual's chance of relying on and engaging with misleading AI-generated images.[61]

With the public possessing varying levels of knowledge and familiarity with generative AI, designing effective labels with the appropriate content is key to preventing deepfake fraud.[62] However, empirical studies have demonstrated that authorship labels can be effective in enhancing transparency, although they are unlikely to affect the persuasiveness of the labelled content which requires

[55]Leng Cheng and Ho-Him Chan, 'Arup lost $25mn in Hong Kong deepfake video conference scam' (Financial Times, 17 May 2024) <https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea> accessed 19 Feb 2026.

[56]Raúl Carrillo, 'Platform Money' (2024) 41 Yale Journal on Regulation 894, 940.

[57]Benjamin Farrand, 'How Do We Understand Online Harms? The Impact of Conceptual Divides on Regulatory Divergence between the Online Safety Act and Digital Services Act' (2024) 16 Journal of Media Law 240.

[58]Petter Törnberg, 'Echo chambers and viral misinformation: Modeling fake news as complex contagion' (2018) 13 PLoS ONE e0203958. See also Ioannis Agrafiotis and others, 'A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate' (2018) 4 Journal of Cybersecurity 1.

[59]We use the term 'deployers' to indicate professional users of AI systems. Under art 3 of the EU AI Act, 'AI system' means 'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.

[60]Kimberly T Mai and others, 'Warning: Humans Cannot Reliably Detect Speech Deepfakes' (2023) 18 PLoS ONE <https://doi.org/10.1371/journal.pone.0285333>. See also Hany Farid, 'Creating, Using, Misusing, and Detecting Deep Fakes' (2022) 1 Journal of Online Trust and Safety <https://doi.org/10.54501/jots.v1i4.56>.

[61]Chloe Wittenberg and others, 'Labeling AI-generated media online' (2025) 4 PNAS Nexus <https://doi.org/10.1093/pnasnexus/pgaf170>.

[62]Sacha Altay and Fabrizio Gilardi, 'People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation' (2024) 3 PNAS Nexus <https://doi.org/10.1093/pnasnexus/page403>.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

alternative techniques to address risks posed by AI-generated information.[63] Wittenberg and others opine that labels indicating whether content has been generated or manipulated by AI are process-based and are not as effective as labelling content as 'manipulated' or 'false', which is impact-based. The latter is more effective in changing people's behaviour.[64] Ternovski and others' experiments reveal that general labels warning the public about videos led the public to distrust *all* videos as fake, even if they were real.[65] Further, the participants in the experiments did not show any improvement in identifying manipulated videos. Therefore, research to date shows that targeted labelling appears to be more effective than general labelling.

Articles 34 and 35(1)(k) of the EU DSA 2022 also mandate very large online platforms to conduct thorough risk analyses, curb the dissemination of illegal content, and label deepfakes and false information. Although this systems-based framework has its advantages, there are weaknesses. The 'notice and action' mechanism in article 16 of the Act allows individuals or entities to ask online platforms to remove illegal content. However, article 16(2) sets out conditions for activating the 'notice and action' mechanism, one of which requires the individual or the entity to provide 'a sufficiently substantiated explanation of the reasons' why they think the content is illegal. This seems rather difficult for the public to determine, especially given that the definition in article 3(h) of the EU DSA 2022 is broad and vague. Whilst child abuse content is clearly illegal, the public needs to provide reasons why this is illegal. This means searching for suitable legal authorities. For non-lawyers, this can be an onerous task. Indeed, de Streel and others report that many stakeholders, such as non-governmental organisations, are of the view that the users 'are not necessarily capable of accurately identifying illegal content online'.[66] The users can only make assumptions that the content is likely to be illegal.

Content moderators compare the content on platforms against databases storing illegal content such as the Image Watch List by the Internet Watch Foundation.[67] This process, called hash-matching, is made more difficult because generative AI can reproduce edited versions of the same illegal content very quickly, bypassing databases which hold illegal content.[68] Further, auto-detection software such as watermarking and labelling methods used by online platforms has weaknesses.[69] In particular, such software is not very effective with shorter texts, content translated from other languages, or factual replies and it implicitly treats synthetic media as false and deceptive.[70] Whilst there are databases of illegal child abuse (IWF Watch List and Interpol),[71] there do not appear to be similar

[63] Isabel O Gallegos and others, 'Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects' (Jul 2025) Stanford University Policy Brief HAI Policy & Society <https://hai.stanford.edu/assets/files/hai-policy-brief-labeling-ai-generated-content.pdf> accessed 20 Feb 2026.

[64] Wittenberg and others (n 61).

[65] John Ternovski, Joshua Kalla and Peter Aronow, 'The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments' (2022) 1 Journal of Online Trust and Safety <https://tsjournal.org/index.php/jots/article/view/28/15>

[66] European Parliament, 'Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform' (Study, PE 652.718, Jun 2020).

[67] See Internet Watch Foundation <https://www.iwf.org.uk/> accessed 20 Apr 2026.

[68] Sam Stockwell and others, 'Privacy-preserving Moderation of Illegal Online Content' (The Alan Turing Institute, Apr 2025) 14–15 <https://cetas.turing.ac.uk/sites/default/files/2025-04/cetas_research_report_-_privacy-preserving_moderation_of_illegal_online_content_0.pdf> accessed 25 Apr 2026.

[69] Google uses a watermarking tool called SynthID Text. It helps developers to spot AI-generated content: Digwatch, 'Google unveils open-source watermark for AI text' (24 Oct 2024), <https://dig.watch/updates/google-unveils-open-source-watermark-for-ai-text> accessed 30 Apr 2026.

[70] Islamic terrorists often write in Arabic on platforms, because they are aware that content moderators working for the online platforms rarely have Arabic language skills: Tom Simonite, 'Facebook is Everywhere; Its Moderation is Nowhere Close' (WIRED, 25 Oct 2021) <https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/> accessed 20 May 2026.

[71] See Internet Watch Foundation, 'IWF Annual Data & Insights Report 2024' <https://www.iwf.org.uk/annual-data-insights-report-2024/> accessed 21 May 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

ones for fraud. Even for established databases for illegal content, Stockwell and others suggest that it is important that these repositories use standardised metrics, such as harmonised labelling and classification methods.[72] Standardisation of repositories would be helpful given the global nature of the internet and its related problems.

The regulatory approach in Singapore

Singapore adopted a piece of legislation specifically targeting deepfakes in political elections. The Elections (Integrity of Online Advertising) (Amendment) Act 2024 bans the publication, boosting, sharing, and reposting of deepfake content depicting election candidates. This adds to the protection given in the Online Safety (Miscellaneous Amendments) Act 2022 and the Online Criminal Harms Act 2023 (OCH Act). The Online Safety (Miscellaneous Amendments) Act 2022 amends the Broadcasting Act 1994 by imposing legal duties on social media platforms. There are two main duties. First, social media platforms with a significant impact must comply with certain Codes of Practice. Secondly, the Infocomm Media Development Authority (IMDA) can require social media platforms to disable access or stop 'egregious content' being transmitted to users.[73] 'Egregious content' broadly covers sexual content, child abuse, and terrorist content but does not refer to financial scams.

The OCH Act covers a range of offences, including scams and malicious cyber activity offences under Part 2 of Schedule 1. The most notable difference between the UK's Online Safety Act and Singapore's OCH Act is that the latter imposes legal duties on government officials, not online platforms. Section 3(1) of the OCH Act establishes the 'competent authority' as 'a public officer from a Ministry or department of the Government; or an employee of a public authority'. Their legal duties are then set out in section 3(2) of the OCH Act. The government can issue five types of directions under the OCH Act, namely: stop communication, disabling, account restriction, access blocking, and app removal directions. A study by the IMDA in Singapore shows that online platforms are often slow to remove harmful content: on average, they take at least five days to respond to the victims.[74] This is a possible explanation for the additional role the government plays in dealing with harmful content.

Regarding scams and malicious cyber activities, the government only needs to suspect that online harm has occurred to issue a direction. For other activities, reasonable suspicion is required. This contrasts with the requirement for 'reasonable grounds to infer' that the content amounts to an offence in section 192(5) of the OSA 2023 for all offences, including scams and cyber activities. The lower Singaporean threshold of merely suspecting that a scam or malicious cyber activity has occurred should make it easier to remove such content. This may be explained by the fact that Singaporeans are particularly vulnerable to scams because many of them are 'affluent, digitally literate and compliant'.[75] Further, scam victims in Singapore suffer some of the highest losses in the world.[76]

Interestingly, a new government agency named the Online Safety Commission, would provide faster remedies for victims.[77] Under the Online Safety (Relief and Accountability) Bill, scam victims can request perpetrators' information to commence legal proceedings. They can also sue online

[72]Stockwell and others (n 68).

[73]Part 5 of the Online Safety (Miscellaneous Amendments) Act 2022.

[74]Infocomm Media Development Authority, 'Online Safety Assessment Report 2024. Designated Social Media Services' (17 Feb 2025) 12–13 <https://www.imda.gov.sg/-/media/imda/files/regulations-and-licensing/regulations/online-safety/online-safety-assessment-report-2024-designated-social-media-services.pdf> accessed 20 Feb 2026.

[75]Owen Walker, "'Rich and naïve": why Singapore is engulfed in a "scamdemic"' (Financial Times, 26 May 2025) <https://www.ft.com/content/3299cf7e-67bd-4654-8aa9-55fc24a66b63> accessed 21 Feb 2026.

[76]Infocomm Media Development Authority (n 74).

[77]See Cyber Indemnity, 'Singapore's Online Safety Commission: A New Era in Combatting Online Harms' (7 Mar 2025) <https://cyberindemnity.org/2025/03/singapores-online-safety-commission-a-new-era-in-combatting-online-harms/> accessed 21 Feb 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

platforms and page administrators for deepfakes, cyberbullying, and sexual harassment. Online platforms must remove offensive content reported by users and victims. The Singapore government has also required app stores to carry out age verification through technology such as facial scans so that only users over eighteen years old can use the apps.[78] These additional protections for scam victims, along with the Broadcasting Act and Protection from Harassment Act 2014, indicate that the Singaporean legislative framework has adopted a strict regulatory approach to online scams.

The challenges of liability attribution for malicious deepfake scams

Establishing liability for GenAI deepfakes poses multiple challenges. Investigations to identify malicious content creators are very costly and time-consuming.[79] For example, a CEO fell victim to a deepfake video and the company detected the fraud quickly but was unable to catch the fraudster or recover financial losses.[80] Further, many deepfake investment scams involve the use of cryptocurrency. Fraudsters use deepfake videos, free cryptocurrency, and promotion codes to lure victims to fake websites.[81] Once the victims enter the promotion codes, they are asked to pay Bitcoin and share their personal details, and they never get anything in return. This is particularly problematic, given that the number of generative AI-enabled scams has increased by 78% from 2023 to 2024 and then by 456% from 2024 to 2025.[82] In the Elon Musk cryptocurrency scam of 2024, Rizzuto used a blockchain explorer and was able to trace the addresses of the scammers' wallets.[83] It was found that Bitcoin, Ethereum, and Dogecoin wallets were used in the scam and new wallets were used to hide the transactions. Yet, such cryptocurrency platforms are currently unregulated, which shield bad actors from liability to scam victims.

Liability of deepfake content creators

Content creators who use deepfakes to defraud victims should *in theory* bear responsibility for their crimes. Malicious content creators have the intention of misleading others and are using AI technology as a weapon to cause harm. As Mirsky and Lee posit, 'for deepfakes, the objective of the generated content is to fool a human'.[84] In the UK, there is a gap in legislation concerning the accountability of malicious content creators. As mentioned above, deepfake scams are intended harms, the prevention of which involves monitoring misuse by users and addressing malfunction, whilst unintended harms concern rectifying design defects in AI outputs. Deepfake scam victims will need to rely on existing UK criminal legislation such as section 2 of the Fraud Act 2006 (fraud by false representation) as a

[78]Osmond Chia, 'New online harms support centre to operate from 2026, will offer victims faster recourse' (The Straits Times, 8 Mar 2025) <https://www.straitstimes.com/singapore/politics/new-online-harms-support-centre-to-begin-operation-in-2026> accessed 22 Feb 2026.

[79]Bart van der Sloot and Yvette Wagensveld, 'Deepfakes: regulatory challenges for the synthetic society' (2022) 46 Computer Law & Security Review 105716, 12–13.

[80]Daniel Thomas, 'WPP boss targeted by deepfake scammers using voice clone' (Financial Times, 10 May 2024) <https://www.ft.com/content/308c42af-2bf8-47e4-a360-517d5391b0b0> accessed 23 Feb 2026. See also Cheng Leng and Chan Ho-him, 'Arup lost $25mn in Hong Kong deepfake video conference scam' (Financial Times, 17 May 2024) <https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea> accessed 23 Feb 2026.

[81]Matthew D Weiner, 'Destined to Deceive: The Need to Regulate Deepfakes with a Foreseeable Harm Standard' (2024) 122 Michigan Law Review 771, 800. It is argued that 'the intent of an individual who creates a deepfake is not debatable; certainly, after synthetically modifying an image or video, the creator did not genuinely believe that the content they shared was accurate'.

[82]See TRM Labs, 'AI-enabled Fraud: How Scammers Are Exploiting Generative AI' (7 May 2025) <https://www.trmlabs.com/resources/blog/ai-enabled-fraud-how-scammers-are-exploiting-generative-ai> accessed 23 Feb 2026.

[83]Rizzuto (n 53).

[84]Yisroel Mirsky and Wenke Lee, 'The Creation and Detection of Deepfakes: A Survey' (2021) 54 ACM Computing Surveys 1.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

potential basis for criminal prosecution. However, gathering clear evidence of intent to deceive by the deepfake content creators can be very challenging.

If a liability framework is to be created, we submit that the focus should be on holding content creators liable for failing to implement verification or safeguard measures such as disclaimers as to how much content has been altered.[85] The challenge, though, is that many such content creators are anonymous because of their nefarious activities. There are significant financial and jurisdictional barriers in uncovering the identity of these anonymous content creators.

Law enforcement techniques in the form of metadata analysis,[86] network traffic analysis,[87] financial trails,[88] and behavioural analysis[89] can all assist in uncovering the identity of the content creator. Courts can also issue search warrants for the police to enter and search premises for evidence of serious offences, including evidence from electronic devices under section 8 of the Police and Criminal Evidence Act 1984. Metadata analysis of Ross Ulbricht's Gmail address and a financial investigation by the Federal Bureau of Investigation led to the successful prosecution of Ulbricht, the mastermind behind the Silk Road online black market.[90] This case involved Ulbricht selling illegal drugs worth $214 million.[91] He disguised his identity by using the username Dread Pirate Roberts. The scale of the operation was unprecedented, with buyers and sellers from all over the world.[92] Although the US law enforcement agencies do not disclose the costs of investigations generally, the Silk Road case involved multiple federal agencies, complex digital forensic work, and a three-week federal trial.[93] Thus, it would have been a very expensive investigation. The Silk Road case was of a huge scale and led to drug-related deaths. Not every case will justify the significant investigative costs. Therefore, imposing liability on content creators poses real challenges.

The interaction between content creators and online social media platforms

Online social media platforms integrate GenAI technologies into their product offerings to enhance content creation workflows and empower content creators. For example, ChatGPT and Copilot are AI-powered software applications powered by large language models (LLMs). They are designed to assist users in generating text through automated programming interfaces which analyse data patterns without human direction.[94] Companies largely embed generative AI models into their digital systems, eg to improve user engagement, enhance content quality, and support creators with limited

[85]Felipe Romero-Moreno, 'Deepfake detection in generative AI: A legal framework proposal to protect human rights' (2025) 58 Computer Law & Security Review 106162.

[86]Samuele Mombelli, James R Lyle and Frank Breitinger, 'FAIRness in digital forensics datasets' metadata – and how to improve it' (2024) 48 Forensic Science International: Digital Investigation 301681.

[87]Pascal Tippe and Adrian Tippe, 'Onion Services in the Wild: A Study of Deanonymization Attacks' (2024) 4 Proceedings on Privacy Enhancing Technologies 291.

[88]Sarah Meiklejohn and others, 'A fistful of bitcoins: characterizing payments among men with no names' (2016) 59 Communications of the ACM 86.

[89]Efstathios Stamatatos, 'A Survey of Modern Authorship Attribution Methods' (2009) 60 Journal of the American Society for Information Science and Technology 538.

[90]US Department of Justice, 'Ross Ulbricht, A/K/A "Dread Pirate Roberts," Sentenced In Manhattan Federal Court To Life In Prison' (2015) <https://www.justice.gov/usao-sdny/pr/ross-ulbricht-aka-dread-pirate-roberts-sentenced-manhattan-federal-court-life-prison> accessed 24 Feb 2026.

[91]Preet Bharara and Serrin Turner, 'Government Sentencing Submission United States v. Ross William Ulbricht, 14 Cr 68 (KBF)' (US Department of Justice, 26 May 2015) <https://i.brayden.id.au/gov.uscourts.nysd.422824.256.0.pdf> accessed 25 Feb 2026.

[92]*United States v Ross William Ulbricht*, 14 Cr 68 (KBF).

[93]US Department of Justice, 'Former Silk Road Task Force Agent Pleads Guilty to Extortion, Money Laundering and Obstruction' (1 Jul 2015) <https://www.justice.gov/archives/opa/pr/former-silk-road-task-force-agent-pleads-guilty-extortion-money-laundering-and-obstruction#:~:text=A%20former%20DEA%20agent%20pleaded%20guilty%20today%20to,and%20sale%20of%20illegal%20drugs%20and%20other%20contraband> accessed 26 Feb 2026.

[94]Noam Kolt, 'Algorithmic Black Swans' (2024) 101 Washington University Law Review 1177.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

resources or technical expertise.[95] However, the interactions between content creators and online social media platforms raise concerns about the employment of GenAI technologies to produce reliable outputs.[96] GenAI could significantly improve creators' productivity and enhance the platforms' profitability. When creators operate on online social media platforms and choose to use AI, the quality of content improves, thereby facilitating economic returns for the platform. As noted, 'platforms offer creators visibility, enabling them to attract and engage audiences. In turn, platforms sell a share of this attention to advertisers'.[97]

The social media platform captures GenAI's potential to boost creators' productivity while maintaining their incentives to participate in content creation.[98] The issue of AI-generated deepfake financial scams has caused widespread concern for investors and consumers; more sophisticated AI tools promote scams that cause harm on online social media platforms rather than protecting vulnerable customers from worthless investment schemes.[99] GenAI's ability to catalyse deceptive or fraudulent content may reflect the appetite of big tech companies (eg Meta, Google, Apple, Microsoft) to publish misleading advertisements enabled by scammers in pursuit of substantial revenues. This agency relationship identifies potential collusive behaviour between content creators and online social media platforms, which can increase misinformation and individuals' vulnerability to deception. It also supports Nash and Felton's argument that legislation should focus more on regulating online social media platforms as *systems* rather than just hosts of content.[100]

It is generally considered that online social media platforms provide the environment in which creators publish content and offer important monitoring capabilities for community management.[101] These platforms provide analytics tools that enable creators to measure their activities. However, it remains unclear how content creation, advertising strategies, and user responses are governed in light of such data. This can raise concerns about the platform's responsibility for detecting, flagging, and removing harmful content to protect users and ensure trust in the distribution of information.[102] Social media companies have often been shielded from responsibility on the assumption that they provide innovative technologies and platforms for speech, such that imposing restrictions on them would amount to a form of censorship of technological progress.[103] Doctrinal debates have questioned whether social media platforms exercise their own freedom of expression rights in the course of content moderation decisions.[104]

Section 230 of the US Communications Decency Act 1996 protects technology companies from legal claims arising from user-generated content on their platforms, treating them as neutral hosts

[95]Di Yuan and others, 'Generative AI Adoption by Creator Platforms' (2025) The Wharton School Research Paper 4 <https://ssrn.com/abstract=5107730> accessed 26 Feb 2026.

[96]Anna Yamaoka-Enkerlin, 'Disrupting Disinformation: Deepfakes and the Law' (2020) 22 New York University Journal of Legislation and Public Policy 725.

[97]Alexander Bleier, Beth L Fossen and Michal Shapira, 'On the role of social media platforms in the creator economy' (2024) 41 International Journal of Research in Marketing 411, 419.

[98]Nic Fildes, 'Meta sued by Australian regulator for allegedly 'misleading' crypto ads' (Financial Times, 18 March 2022) <https://www.ft.com/content/132c7877-bd74-4957-925b-414f12ec6aa4> accessed 18 March 2026.

[99]Jeannie Marie Paterson, 'Banks, Authorised Push Payment Scams and Models for Compensation' (2025) 99 Australian Law Journal 178, 182.

[100]Nash and Felton (n 47) 820.

[101]Noor Johnson and others, 'The Use of Digital Platforms for Community-Based Monitoring' (2021) 7 BioScience 452.

[102]Miriam C Buiten, Alexandre de Streel and Martin Peitz, 'Rethinking liability rules for online hosting platforms' (2020) 28 International Journal of Law and Information Technology 139.

[103]Tim Wu, 'Social Media Isn't Just Speech. It's Also a Defective, Hazardous Product' (The New York Times, 14 Mar 2026) <https://www.nytimes.com/2026/03/14/opinion/social-media-trial-addiction.html> accessed 18 Mar 2026.

[104]Stefan Theil, 'Private censorship and structural dominance: Why social media platforms should have obligations to their users under freedom of expression' (2022) 81 Cambridge Law Journal 650.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

(similar to telephone companies) rather than as publishers.[105] While US legislation affords immunity to platforms for removing offensive content, courts have begun recognising causes of action that address the weaknesses and risks of social media.[106] Section 230 provides that an entity is shielded from liability only in its capacity as a provider of an interactive computer service. However, the provision does not protect a company when acting as the publisher or speaker of information provided by another information content provider.[107] As a result, a service provider is not shielded from liability where it makes a material contribution to the creation or development of content.[108] It is worth noting that GenAI outputs are not protected by section 230, thereby undermining any defence available to companies. When a platform incorporates GenAI applications into features such as messaging and profile creation, it may jeopardise its section 230 protection.[109] Lawsuits against social media platforms have been successful on the grounds of users' mental health harms, as illustrated by the cases *State of New Mexico v Meta Platforms, Inc.*[110] and *KGM v Meta Platforms et al*,[111] where courts have focused on how technology platforms design their products and services to protect users from misleading or harmful content.

The EU's AI Act attempts to address accountability gaps between providers and deployers. Providers are responsible for the design and documentation of AI systems, while deployers manage their contextual use and oversight. For general-purpose AI and foundation models (eg chatbots), upstream obligations focus on technical documentation, systemic risk mitigation, and downstream information flows, recognising their infrastructural role beyond single use cases. However, these foundation models pose significant challenges for regulators and policymakers as they can be adapted to sophisticated automated tasks such as those enabled by large language models, which make it difficult to detect malicious scams and to allocate responsibility.[112] Article 50 of the AI Act regulates transparency obligations for providers and deployers of certain AI systems, requiring the provision of specific information about the content and outputs of the AI system. In general terms, providers of high-risk AI systems are required to adopt additional precautions and safety measures to mitigate the undesirable effects of unexpected outcomes.[113] These obligations are complemented by the Code of Practice on the transparency of AI-generated content, which sets out regulatory measures – whose adoption is voluntary – for providers and deployers of GenAI systems on how to comply with the requirements of the AI Act.[114] Risk management requirements cover predictable and reasonably foreseeable harms associated with the use of AI applications, but sophisticated computational systems may fall outside the taxonomy introduced by the Act. Application layers deployed on AI

[105] Beatrice Nolan, 'Section 230 protected social media companies from legal responsibility for misinformation. AI chatbots could be about to change that' (Fortune, 8 Oct 2025) <https://fortune.com/2025/10/08/ai-chatbot-section-230-meta-social-media-legal-shield-no-protection/> accessed 2 Mar 2026.

[106] *Forrest v Meta Platforms, LLC*, 2024 WL 3024642 (ND Cal, 17 Jun 2024). For an overview see Ryan Calo, 'Courts Should Hold Social Media Accountable - But Not By Ignoring Federal Law' (Harvard Law Review Blog Essay, 10 Sep 2024) <https://harvardlawreview.org/blog/2024/09/courts-should-hold-social-media-accountable-but-not-by-ignoring-federal-law/> accessed 2 Mar 2026.

[107] *Forrest v Meta Platforms, LLC*, 2024 WL 3024642 (ND Cal, 17 Jun 2024).

[108] *Rigsby v Godaddy Inc*, 59 F4th 998 (9th Cir 2023).

[109] Eric Goldman, 'Section 230 Doesn't Apply to Generative AI Enhancements to Ad Copy (But the Plaintiffs Lose Anyway) – Bouck and Suddeth v Meta' (Technology & Marketing Law Blog, 13 Jun 2026), <https://blog.ericgoldman.org/archives/2026/06/section-230-doesnt-apply-to-generative-ai-enhancements-to-ad-copy-but-the-plaintiffs-lose-anyway-bouck-and-suddeth-v-meta.htm> accessed 3 Jul 2026.

[110] *State of New Mexico v Meta Platforms, Inc*, D-101-CV-2023-02838 (1st Jud Dist Ct, 6 Aug 2026).

[111] *P F, et al (KGM) v Meta Platforms, Inc, et al* 4:25-cv-02691 (ND Cal 2026).

[112] Ian Brown, 'Allocating accountability in AI supply chains: a UK-centred regulatory perspective' (Ada Lovelace Institute, Jun 2023) 41 <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/> accessed 3 Mar 2026.

[113] Michael Veale and João Pedro Quintais, 'The Obligations of Providers of General-Purpose AI Models' in Gianclaudio Malgieri and others (eds), *The EU Artificial Intelligence Act. A Thematic Commentary* (Hart Bloomsbury 2026) ch 22.

[114] See European Commission, 'Code of Practice on transparency of AI-generated content' (10 Jun 2026) <https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content> accessed 10 Jun 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

systems generate new risks within the value chain, raising questions as to the need to examine the broader network, as the operational adequacy of models cannot serve as a blanket justification for assigning responsibility.[115]

Article 25 of the AI Act classifies distributors, importers, deployers, or other third parties as providers of high-risk AI systems, thereby prescribing responsibilities along the value chain. This provision must be read in conjunction with article 16 of the Act, which sets out specific obligations for the product manufacturer as the provider of the system, focusing on the disclosure of information concerning the tools, services, components, and processes used or integrated within a high-risk AI system. Specifically, article 16 establishes a set of requirements to ensure compliance with the applicable conformity assessment procedures, including a series of control mechanisms related to the quality management system. Furthermore, article 17 of the Act requires providers of AI systems to implement techniques, validation procedures, testing, examination, and systematic measures for the design, development, and data governance of the system. At first sight, the bulk of obligations placed on providers suggests that the legislation identifies them as the primary point of reference for allocating responsibility. However, unforeseen events – particularly where developers ought reasonably to have known of the existence of a risk prior to its materialisation – complicate this allocation.[116]

The perpetrator and attacker feeding adversarial data and poisoning AI models introduce unforeseen risks, which may jeopardise the attribution of responsibilities and the grounds for duties that arise in relation to the associated actions necessary to prevent the negative consequences of such events. On this point, the doctrinal debate has questioned how to identify AI providers' duties in relation to AI risk management and risk assessment.[117] Unintended consequences of unforeseen outcomes are context-based, spanning predictive analysis and decision-making processes across various stages of specific AI services (eg machine learning, cloud computing, and GenAI applications). These dynamics significantly affect the control mechanisms governing AI systems and may shift the role of providers in relation to customer-facing application functionality.[118]

The adoption of safety measures and the ability to foresee certain risks are key to allocating liability along the AI value chain, although the level of participation of each actor, their expertise in identifying harmful effects of defective products, and the design of the AI model may alter the distribution of responsibility. As argued, 'the manufacturers of key AI components, such as foundation model providers, may be best positioned to adopt safety measures and foresee certain risks and thus should be held liable, or jointly liable, when harm is caused by an AI system'.[119] Foundation models constitute a primary source of deepfake scams, as they are susceptible to data manipulation and cyberattacks by downstream developers. Articles 53 and 55 of the EU AI Act impose obligations on providers of general-purpose AI models to disclose information concerning the training and testing processes, as well as the results of their evaluation, and require the adoption of codes of practice aimed at mitigating systemic risks originating from upstream developers. However, deepfake financial scams also raise concerns in relation to downstream developers, whose modifications of foundation models by other actors within the AI value chain may generate unforeseen harms that render safety measures ineffective.[120]

[115] Alessio Azzutti, 'AI Governance in Algorithmic Trading: Some Regulatory Insights from the EU AI Act' (2024) 41 Banking and Finance Law Review 133.

[116] Mitchell Chervu Johnston and Brian Matejek, 'Overcoming Unforeseeable AI Errors: A Proposal' (2026) Boston College Law School Legal Studies Research Paper No 675.

[117] Ellen Hohma, 'On the elements and implications of accountability for AI providers' in Christoph Lütge and others (eds), *The Elgar Companion to Applied AI Ethics* (Edward Elgar 2024) 25.

[118] Jennifer Cobbe and Jatinder Singh, 'Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges' (2021) 42 Computer Law & Security Review 105773.

[119] Beatriz Botero Arcila, 'AI Liability Along the Value Chain: Lessons from the Liability of Suppliers of Components in Product Liability Law' (2025) 18 Journal of Tort Law 280–281.

[120] Sophie Williams, Jonas Schuett and Markus Anderljung, 'On Regulating Downstream AI Developers' (2026) 17 European Journal of Risk Regulation 103.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

Downstream activities are difficult to detect within large datasets, given the ability of third-party developers, such as those operating through application programming interfaces (APIs), to conceal their identity when fine-tuning the original model.[121] Attributing responsibility and accountability to AI providers and digital platform operators for downstream harms may therefore constitute an arduous exercise, requiring a careful assessment of causation, foreseeable consequences, and applicable standards of reasonable care, which EU legislation has not yet fully addressed.[122] Imposing risk management requirements, together with a set of duties of care, skill, and diligence for senior board members of companies deploying AI models, may represent a viable option for allocating responsibility, although specific harms, such as errors in downstream activity, remain difficult to foresee.[123] Establishing a causal link between downstream activity and defective outputs requires a meticulous inquiry into whether the actors involved in the value chain possessed a sufficiently high level of expertise at a given stage of the process, as well as into the specific patterns derived from training datasets.

We argue that a shared liability framework can shed light on the technical complexities that the AI value chain presents for determining liability associated with fraudulent behaviour. Such a framework may operate as a deterrent mechanism by targeting each actor's contribution across the system life cycle, thereby facilitating the foreseeability of risks and addressing regulatory gaps in the EU AI Act, which remains largely silent on how to allocate accountability for harm between upstream and downstream developers. However, a shared liability regime must be assessed on a contextual basis, and it may open the floodgates to users' compensation claims, given the incentive to seek redress from multiple parties.[124]

The idea of moral sense and self-restraint on the part of AI creators should not preclude the adoption of online regulations aimed at preventing the dissemination of misleading and fraudulent information.[125] Such regulations should not remain confined to the national level; rather, there is a need for a global consensus on how AI systems are assessed and managed, and on how liability is allocated for losses. However, as argued, the effectiveness of online regulation depends on the extent to which users can circumvent restrictions or substitute non-compliant platforms.[126] Creators know how to profit from their content, while users increasingly perceive platforms as providers of valuable goods and services, enabling them both to consume content and to engage in shopping activities. However, this interaction may give rise to a pernicious cycle, whereby creators are incentivised to maximise income generation and platforms, in turn, promote user engagement to accommodate creators' demand for more predictable remuneration from content-based advertising.[127]

[121] Philipp Hacker, Ramayya Krishnan and Marco Mauer, 'Global AI Governance – Part 1: Decoding Developers and Deployers' (Oxford Business Law Blog, 9 Dec 2024) <https://blogs.law.ox.ac.uk/oblb/blog-post/2024/12/global-ai-governance-part-1-decoding-developers-and-deployers> accessed 7 Mar 2026.

[122] Martin Ebers and Emmanuel Vargas Penagos, 'Upstream, downstream, and in between: navigating the GPAI value chain under EU law' (2026) Information & Communications Technology Law 1.

[123] Henry L Fraser and Nicolas P Suzor, 'Locating fault for AI harms: a systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain' (2025) 17 Law, Innovation and Technology 103, 111.

[124] On a similar view see Bart Custers, Henning Lahmann and Benjamyn I Scott, 'From liability gaps to liability overlaps: shared responsibilities and fiduciary duties in AI and other complex technologies' (2025) 40 AI & Society 4035, 4043.

[125] On this point see the editorial of Martin Wolf, 'We have to manage the AI revolution' (Financial Times, 17 Jun 2026) <https://www.ft.com/content/59024744-50bf-43f7-9d79-738deabc7a60?syn-25a6b1a6=1> accessed 8 Jun 2026.

[126] Matthew Brown, Emily J Davis and Devin G Pope, 'Can Online Activity Be Regulated? Evidence from Adult Websites' (Jun 2026) NBER Working Paper 35322 <https://www.nber.org/papers/w35322> accessed 18 Jun 2026.

[127] Catalina Goanta, 'The New Social Media: Contracts, Consumers, and Chaos' (2023) 108 Iowa Law Review Online 118.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

Online social media platforms' responsibility for deepfake scams

Sophisticated adverts, including AI-generated deepfakes, persuade individuals to disclose personal data or invest in fraudulent schemes. Online social media platforms are a favourite conduit for carrying such adverts and profiting from them, and big tech companies seem unable to curb the proliferation of deepfakes.[128] According to the US Federal Trade Commission's Consumer Advice Report, 'scammers use social media platforms to promote bogus investment opportunities, and even to connect with people directly as supposed friends to encourage them to invest'.[129] Frauds originating on social media (eg Facebook, Instagram) indicate that platforms are the main method of contact, and cryptocurrency is the most used method of payment.[130] We focus on regulated social media platforms such as Facebook and Instagram; largely unregulated platforms (such as Telegram, Signal, WhatsApp)[131] are outside the scope of our analysis. Public regulators and policymakers call for action to protect consumers as fraud causes severe harm to individuals and the stolen money goes to organised crime groups mostly hidden behind crypto platforms.[132]

Technology-enabled frauds require greater collaboration between the public and private sectors, and a 'systemic' solution, as scammers adopt new methods and increase unauthorised losses which exacerbate a systemic problem.[133] Online social media platforms should bear a legal duty to check that investment advertisers are authorised, and a legal duty not to provide advertising space to fraudsters in the first place. They ought to be expected to 'know their customers' and be held liable, with proper enforcement and tough penalties, if they fail to block the dissemination of fraudulent advertising content.[134] Further, online social media platforms should be legally required to verify that advertisers are authorised by a regulator to sell financial services and block them in case they are not.[135] A mandatory risk assessment for providers of very large online platforms about the dissemination of manipulative content through their services is required by article 34(1) of the EU DSA 2022, although this is not equivalent to an obligation to control the use of the platform. It seems more like a due diligence test which could result in a ticking-the-box exercise.[136] This solution would not address the issue of who should be held accountable for the harm caused by such frauds.

[128]OECD, 'Protecting Consumers from Financial Scams and Frauds' (16 Jun 2026) 16–17 <https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/06/protecting-consumers-from-financial-scams-and-frauds_3887b249/d41817bb-en.pdf> accessed 18 Jun 2026.

[129]Emma Fletcher, 'Social media a gold mine for scammers in 2021' (FTC, Jan 2022), <https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/01/social-media-gold-mine-scammers-2021> accessed 8 Mar 2026.

[130]See 'FTC Action Ends Ecommerce Empire Builders Online Business Opportunity Scam' (FTC, 9 May 2025) <https://www.ftc.gov/news-events/news/press-releases/2025/05/ftc-action-ends-ecommerce-empire-builders-online-business-opportunity-scam> accessed 9 Mar 2026; 'FTC Takes Action to Stop Sprawling 'Growth Cave' Business Opportunity and Credit Repair Scam' (FTC, 7 Mar 2025) <https://www.ftc.gov/news-events/news/press-releases/2025/03/ftc-takes-action-stop-sprawling-growth-cave-business-opportunity-credit-repair-scam> accessed 9 Mar 2026; 'FTC Announces Crackdown on Deceptive AI Claims and Schemes' (FTC, 25 Sep 2024) <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes> accessed 9 Mar 2026.

[131]Jacob Gursky and Samuel Woolley, 'Countering Disinformation and Protecting Democratic Communication on Encrypted Messaging Applications' (The Brookings Institution of Foreign Policy, Jun 2021) <https://www.brookings.edu/articles/countering-disinformation-and-protecting-democratic-communication-on-encrypted-messaging-applications/> accessed 10 Mar 2026.

[132]Joshua Oliver, 'The lawless world of crypto scams' (Financial Times, 19 Sep 2022) <https://www.ft.com/content/5987649e-9345-4eae-a4b8-9bfb0142a2ab> accessed 10 Mar 2026.

[133]Maisie Grice, 'UK fraud 'blight' prompts calls for action to protect consumers' (Financial Times, 27 May 2025) <https://www.ft.com/content/c877784c-f416-4c7b-943b-dfb35a01233c> accessed 10 Mar 2026.

[134]See the editorial 'How to block the financial scammers on social media' (Financial Times, 18 May 2025) <https://www.ft.com/content/93d29681-c242-4d0f-9789-41dbbebea906> accessed 11 Mar 2026.

[135]Akila Quinio and Martin Arnold, 'Social media must do more to tackle payment fraud, says UK regulator' (Financial Times, 14 Oct 2024) <https://www.ft.com/content/1e77ae86-a0c2-47ee-804b-cbe50765a9a8> accessed 11 Mar 2026.

[136]Stefano Faraoni, 'Why Generative AI Is Not Cyrano de Bergerac. A Computational Manipulation Perspective on Generative AI' in Mimi Zou and others (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge University Press 2025) 56.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

Court rulings have begun to recognise the responsibility of technology companies in cases where they fail to control the products and services they place on the digital market. A German regional court found Google liable for false claims (including scams, links to dubious companies, subscription traps, and suspicious phone calls) and inaccurate information contained in an AI-generated search overview, which was treated as its own content rather than merely as a list of search results.[137] It made clear that Google has influence over the AI's offering and the algorithms with which the AI operates, which essentially convert the function of generative search from a hosting service to a publishing service.[138] Another interesting case is the UK's Competition and Markets Authority (CMA) action against Google under the Digital Markets, Competition and Consumers Act 2024.[139] The CMA imposed targeted rules on Google's search services following its decision to designate the company as having Strategic Market Status (ie market power combined with a position of strategic significance in a digital activity) in the provision of general search services.[140] This designation allows the CMA to introduce conduct requirements governing Google's search activities, provided that such measures are proportionate to the objectives of ensuring fair dealing, open choices, and trust and transparency.[141]

We argue that a regulatory framework imposing shared responsibility for deepfake financial scams on malicious content creators and online social media platforms would provide a viable redress mechanism for victims' losses and would impose corresponding obligations on these actors to compensate those affected by fraud. Such a solution may be particularly appropriate where a content creator intentionally trains an AI model to produce deceptive outputs, and where online platforms use such outputs to distort or manipulate information (rather than to moderate or report harmful content) relevant to individuals' decision-making processes.[142]

The conundrum of a liability regime in GenAI

The liability contours of deepfake financial scams involve multiple actors at multiple stages of the GenAI supply chain: malicious deepfake content creators, software developers, and online social media platforms. Regulatory challenges around the degree of liability for malicious scams have created concerns about suitable remedies.[143] In most disputes, responsibility is placed on consumers to exercise reasonable care in their decision-making process following the principle of 'caveat emptor'.[144] However, this approach does not resolve the issue of liability attribution for unforeseeable harms, which could not be reasonably prevented, such as inaccurate outputs generated by AI

[137]Regional Court of Munich (Landgericht Muenchen I, 26 O 869/26, 28 May 2026) <https://cdn.arstechnica.net/wp-content/uploads/2026/06/Google-AI-Overview-Munich-Court-Ruling.pdf> accessed 11 Jun 2026.

[138]Foo Yun Chee, 'Google to challenge German ruling saying it is liable for AI-generated false claims' (Reuters, 12 Jun 2026) <https://www.reuters.com/world/google-appeal-german-court-ruling-assigning-liability-ai-overviews-false-claims-2026-06-12/> accessed 12 Jun 2026.

[139]Competition and Markets Authority, 'CMA secures fairer deal for publishers and improves Google search services in UK' (Press Release, 3 Jun 2026) <https://www.gov.uk/government/news/cma-secures-fairer-deal-for-publishers-and-improves-google-search-services-in-uk> accessed 15 Jun 2026.

[140]'Google must let UK publishers opt out of AI search under new rules' (Reuters, 3 Jun 2026) <https://www.reuters.com/legal/litigation/uk-regulator-enforces-new-competition-requirements-google-search-2026-06-03/> accessed 15 Jun 2026.

[141]Competition and Markets Authority, 'Further CMA action to secure a fairer deal for businesses and improve Google search services in UK' (Press Release, 17 Jun 2026) <https://www.gov.uk/government/news/further-cma-action-to-secure-a-fairer-deal-for-businesses-and-improve-google-search-services-in-uk> accessed 26 Jun 2026.

[142]Peter Henderson, 'Challenges for Foundation Model Liability and Regulatory Regimes. An Analysis of US Law' in Mimi Zou and others (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge University Press 2025) 126–127.

[143]Judit Bayer, 'Legal implications of using generative AI in the media' (2024) 33 Information & Communications Technology Law 310.

[144]In *Santander UK Plc v CCP Graduate School Ltd* [2025] EWHC 667 (KB), the court held that a bank may reasonably foresee harm to innocent third parties once alerted to the fact that one of its accounts has been used by fraudsters; however, it did not have a relationship with such third parties sufficient to establish the necessary proximity. This decision reflects the approach adopted by the Supreme Court in *Philipp v Barclays Bank UK plc* [2023] UKSC 25, where a

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

tools embedded in GenAI and LLMs[145] that 'hallucinate' fabricated or false information about historical events, legal filings, news, and research articles.[146] AI hallucinations produce plausible but incorrect outputs, even with perfect data, due to fundamental statistical and computational limits which cause unintentional errors in the decision-making process.[147] The datasets used to train the models can be biased, inaccurate, or intentionally disrupted by content creators who provide faulty narratives to manipulate information.[148] As argued, corrupted data and information make GenAI outputs no longer so intelligent, profitable, or socially beneficial for market participants.[149] However, GenAI hallucinations not only produce false outputs but are also entangled in the cognitive experience (eg memories, emotions, behaviour) that characterises the human-AI interaction in conversational interfaces.[150] GenAI tools hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty: AI models do not acknowledge uncertainty, but they provide estimates as facts without further research into the validation of their statements.[151]

---

bank customer was the victim of authorised push payment fraud and had been deceived into instructing the bank to transfer funds to fraudsters. In *Royal Bank of Scotland International Ltd (Respondent) v JP SPC 4* [2022] UKPC 18, the Privy Council had accepted that the bank 'had no special level of control over the source of danger (ie, it was not in control of the fraudsters)'.

[145]LLMs are computational deep learning models trained on an immense quantity of data, capable of understanding and generating natural language to perform a wide range of tasks. An analysis of 'hallucinations' in the LLM outputs is conducted by Claudio Novelli and others, 'The Artificial in "Artificial Intelligence": How Imagination Shapes AI Regulation' (2026) Centre for Digital Ethics Research Paper, 30–31 <https://philarchive.org/archive/NOVTAI-2v2> accessed 28 Jun 2026.

[146]Hallucinated outputs in GenAI result from training data, the design of tools focused on pattern-based content generation, and the inherent limitations of AI models. See MIT Management, 'When AI Gets It Wrong: Addressing AI Hallucinations and Bias' (2026) <https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/> accessed 29 Jun 2026. In *Tajudin bin Gulam Rasul v Suriaya bte Haja Mohideen* [2025] SGHCR 33, the Singapore Court ordered the claimant's counsel personally to pay costs to the defendant for having cited a fictitious authority in the claimant's written submissions. The hallucinated case was created by GenAI and spotted by the defendant's counsel. Claimant's counsel omitted the hallucinated case from the bundle and subsequently sought to file a revised version of the submissions with a replacement authority without the Court's permission. Claimant's counsel provided the explanation that the hallucinated case had been unintentionally cited. Claimant's counsel also sought to explain the revised submissions as merely 'to update the document due to typographical errors without any amendments to the contents' and to correct 'clerical errors'. The defendant's counsel was willing to accept the claimant's counsel's representations and that the hallucinated case might have been unintentionally cited. The Court considered that claimant's counsel's conduct was, amongst other things, improper. The Court noted that 'of gravest concern, [claimant's counsel] was less than candid with the Court and sought to downplay the gravity of his improper conduct'. The decision sets out a multi-factor approach to determine the egregious nature of counsel's misuse of GenAI and how costs should be apportioned as a result. The decision sets out the following factors: (a) whether the fictitious AI-generated authority was intentionally cited to mislead or deceive the court; (b) whether the advocate and solicitor had previously cited fictitious AI-generated authorities to the court; (c) whether an immediate, full and truthful explanation is given to the court and the counterparty. Specifically, whether the advocate and solicitor expeditiously informed the court that a fictitious AI-generated authority was cited and took appropriate steps, in consultation with the court, to remedy the mistake, or, conversely, if he displayed a lack of candour and attempted to downplay or conceal his mistake; and (d) the impact on the underlying litigation, in particular, whether the legal proposition purportedly supported by the fictitious authority exists and could have been supported by a genuine authority [72].

[147]It has been reported that epistemic uncertainty arises when information appears rarely in training data, that model limitations occur where tasks exceed the representational capacity of current architectures, and that computational intractability persists where even superintelligent systems cannot solve cryptographically hard problems. See Adam Tauman Kalai and others, 'Why Language Models Hallucinate' (2025) Computation and Language, arXiv:2509.04664.

[148]Naja Bentzen, 'Information manipulation in the age of generative artificial intelligence' (Briefing by European Parliamentary Research Service PE 779.259, Dec 2025) <https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/779259/EPRS_BRI(2025)779259_EN.pdf> accessed 30 Jun 2026.

[149]Tom CW Lin, 'Artificial Intelligence, Misinformation, and Market Misconduct' (2024) 85 Ohio State Law Journal 685, 702.

[150]Lucy Osler, 'Hallucinating with AI: Distributed Delusions and "AI Psychosis"' (2026) 39 Philosophy & Technology <https://doi.org/10.1007/s13347-026-01034-3>.

[151]In *R (Ayinde) v London Borough of Haringey* [2025] EWHC 1383, hallucinated cases, or 'fake authorities,' were regarded as professional misconduct. The Court held that 'placing false material before the court with the intention that the court

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

The impact of uncertainty about the performance of GenAI is a matter of concern for legal practice and consumer claim disputes, where the widespread use of AI models introduces risks of misperception when accessing valuable advisory assistance.[152] There is a pressing need to address the harmful impact of outcomes due to the uncertainty and unpredictability of risks in the GenAI supply chain.[153] AI risk frameworks have proved inadequate for the reality of persistent hallucinations; they often underestimate epistemic uncertainty, so updates are needed to address systemic unpredictability.[154] In the doctrinal debate, it has been noted that additional concerns result from the lack of a universally accepted definition of hallucinations.[155] As a result, there is no reliable way of measuring their occurrence and evaluating the reliability of specific GenAI-based tools. Requiring disclosure or certification of GenAI use, such as a mandatory 'hyperlink rule'[156] in legal citations and contentions to authoritative sources, would mitigate the risk of users' reliance on hallucinated cases through verification at the point of citation.[157]

Situations of unintended harms caused by faulty outputs (eg hallucinating critical information in providing advice or assistance as a result of genuinely defective AI behaviour) are distinct from intentional abuse of a dual-use technology to commit fraud (eg human-generated misinformation that influences decision-making). Human misinformation is technically and conceptually different from AI hallucinations in terms of motivation, beliefs, and intent.[158] This results in human-initiated inaccuracies which involve deliberate deception or flawed prompts, leading to disinformation as intentional falsehoods, and computational manipulation of individuals' choices.[159] As a corollary, intentional misuse of GenAI may exacerbate the risk of deterioration of the information ecosystem with lower-quality, synthetic, or misleading content, effectively generating untruthful content.[160] Sophisticated forms of market manipulation come from misuse of AI techniques such as algorithmic trading, which faces 'challenges to guaranteeing accountability and assigning responsibility, as human experts do not explicitly program the AI behaviour'.[161] The complexity and interoperability of automated trading practices as well as their autonomy in the decision-making activity (self-learning

---

treats it as genuine may, depending on the person's state of knowledge, amount to a contempt. That is because it deliberately interferes with the administration of justice' [26]. The Court recognised that 'the information produced by generative large language model artificial intelligence tools is a result of their operational design. They generate textual responses by predicting what words or phrases come next in a particular context, based on patterns identified from a vast quantity of training data' [6].

[152]Vivi Tan, Jeannie Marie Paterson and Julian Webb, 'Generative Artificial Intelligence in Small Value Consumer Claims: Hallucination Risk, System Design, and Governance in Online Dispute Resolution' (2025) 48 University of New South Wales Law Journal 1277.

[153]Teresa Rodríguez de Las Heras Ballell, 'Mapping Generative AI rules and liability scenarios in the AI Act, and in the proposed EU liability rules for AI liability' (2025) 1 Cambridge Forum on AI: Law and Governance 7.

[154]Gyana Swain, 'OpenAI admits AI hallucinations are mathematically inevitable, not just engineering flaws' (Computer World, 18 Sep 2025) <https://www.computerworld.com/article/4059383/openai-admits-ai-hallucinations-are-mathematically-inevitable-not-just-engineering-flaws.html?trk=feed_main-feed-card_feed-article-content> accessed 20 Jun 2026.

[155]Eliza Mik, 'Revisiting Legal Hallucinations' (10 Mar 2026), Seminar delivered at the Centre for Legal Innovation and Digital Society (CLINDS)'s 30th LegalTech Seminar, Faculty of Law, The Chinese University of Hong Kong.

[156]An electronic link embedded in a citation that allows direct access to the cited authority would prevent the careless use of AI outputs.

[157]Oliver Roberts, 'Spread of AI Hallucinations Drives Need for Sanctions Reporting' (Bloomberg Law, 20 Feb 2026) <https://news.bloomberglaw.com/legal-exchange-insights-and-commentary/spread-of-ai-hallucinations-drives-need-for-sanctions-reporting> accessed 15 Mar 2026.

[158]Anqi Shao, 'New sources of inaccuracy? A conceptual framework for studying AI hallucinations' (2025) 6 Harvard Kennedy School Misinformation Review 3.

[159]Stefano Faraoni, 'AI-Enabled Manipulative Techniques: A Contract Law Perspective' (2026) Lloyd's Maritime and Commercial Law Quarterly 77.

[160]Joseph E Stiglitz and Maxim Ventura-Bolet, 'The Impact of AI and Digital Platforms on the Information Ecosystem' (NBER Working Paper Series No. 34318, Oct 2025) 3–4 <http://www.nber.org/papers/w34318> accessed 20 Mar 2026.

[161]Alessio Azzutti, 'AI trading and the limits of EU law enforcement in deterring market manipulation' (2022) 45 Computer Law & Security Review 4.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

from their own experience based on market observation) make it difficult to identify the degree of liability (among the parties involved in designing, developing, using, and monitoring the AI system) for misconduct and harm when operating in financial transactions. An innovative proposal to implement a credible regulatory deterrence framework to achieve effective law enforcement of securities law has been suggested in the doctrinal debate with a view to addressing unlawful behaviours and conduct in the trading context.[162]

Manipulated content in deepfake financial scams undermines technological advancements and exploits vulnerabilities in software design to facilitate intentional fraud, misinformation, and emotional harm.[163] The Financial Stability Board (FSB) warned about the potential of GenAI to increase financial fraud and enable malicious actors to generate and spread disinformation in financial markets, which underscores the need for accountability mechanisms and content moderation.[164] Generative AI visual text is produced by a human designer or creator, adapted to representational media for altering users' beliefs, and distributed to mislead the audience.[165] It is the human who is at fault for deceiving and distorting reality through materials artificially created by the AI tool; this affects market structure and trading exchanges with negative implications for competition and transaction costs.[166] When AI tools self-learn and self-operate with disrupted data to achieve an unlawful objective, it is difficult for regulatory authorities and market users to understand how the outputs originate. Conduct of business rules can mitigate the risks of traders committing abuses, although the high-speed data used in financial products makes monitoring and supervisory activities meaningless with respect to technological errors.[167] Risk management and internal controls within financial firms are often inadequate to detect the potential harms created by malicious deepfake scams; in parallel, regulatory compliance frameworks and prudential standards adopted by financial authorities exhibit weaknesses in terms of skills and resources (including deficiencies in data science expertise and essential IT capabilities) necessary to assess the reliability of AI tools.[168] Model validation, ongoing monitoring, outcome analysis, and data quality assessment entail substantial compliance costs and require a thorough understanding of the vulnerabilities associated with AI models. This, in turn, calls for the development of policy frameworks that enhance engagement with private sector participants, AI developers, and other third-party service providers.[169]

[162] ibid 12–13. It is noted that such a regulatory framework is workable if it guarantees detection, prosecution, and sanctioning of misconduct. This proposal is complemented by a multi-layered liability framework for AI trading manipulation, which integrates administrative and criminal liability aspects relating to AI misconduct to ensure preventive detection of unlawful activities and collaboration in information disclosure and enforcement action with the regulatory authorities.

[163] David Hirshleifer and others, 'AI, Opinion Ecosystems, and Finance' (NBER Working Paper Series No 34807, Feb 2026) <http://www.nber.org/papers/w34807> accessed 20 Mar 2026.

[164] FSB, The Financial Stability Implications of Artificial Intelligence (14 Nov 2024) 2 <https://www.fsb.org/uploads/P14112024.pdf> accessed 18 Mar 2026.

[165] Michael D Murray, 'Visual Legal Rhetoric in the Age of Generative AI and Deepfakes: Renaissance or Dark Ages?' (2025) 28 SMU Science and Technology Law Review 199, 230–231.

[166] Chinmayi Sharma, 'AI's Hippocratic Oath' (2025) 102 Washington University Law Review 1101, 1159–1160, proposing a new professional licensing process for artificial intelligence engineers.

[167] Yesha Yadav, 'Oversight Failure in Securities Markets' (2019) 104 Cornell Law Review 1799, 1855, pointing out that 'the likelihood of error and disruption is amplified by ineffective oversight in fragmented markets … which can be addressed by introducing mutual contribution to a compensatory fund to pay out on liability claims when a single exchange or dark pool cannot' (1858). A system of shared liability funds for trading venues can reduce incentives for market participants to take risks that cannot be covered in case of losses and can facilitate industry self-monitoring and discipline.

[168] FSB, Enhancing Third-Party Risk Management and Oversight. A toolkit for financial institutions and financial authorities (4 Dec 2023) 33 <https://www.fsb.org/uploads/P041223-1.pdf> accessed 21 Mar 2026.

[169] OECD, Regulatory approaches to Artificial Intelligence in finance (OECD Artificial Intelligence Papers No. 24, September 2024) 35–36
<https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/09/regulatory-approaches-to-artificial-intelligence-in-finance_43d082c3/f1498c02-en.pdf> accessed 25 Mar 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

The shared liability model

We identify different scenarios for allocating responsibility among parties implicated in fraudulent AI-generated outcomes. We contend that a shared liability regime would provide a legal response to fraudulent content, as well as act as a deterrent by discouraging malicious content creators and online social media platforms from disseminating deepfakes. This approach stems from the wider discussion about the conceptual framework of shared single liability among multiple injurers.[170] Shared responsibility applies to multiple stakeholders proportionally responsible for a particular action or outcome.[171] The share of liability must be assessed on a case-by-case basis depending on the duty borne by each party involved in the illegal activity. The limitations of this proposed liability approach lie in the potential hurdles in identifying the specific harm associated with the deceptive action for each party involved in the AI supply chain. These limitations reflect the risk of misallocating the burden of responsibility among the actors engaged in the system life cycle, thereby both diluting accountability in the decision-making process and placing disproportionate compliance burdens on AI providers.

Andhov and Gardner have observed that complex algorithmic decision-making systems challenge the application of tort law as an appropriate private cause of action to ensure compensation for individual victims of financial harm.[172] Specifically, it could be difficult to demonstrate that the fault of the algorithm caused the harm experienced. In the context of the tort of negligence, risks of harm can be reasonably foreseeable, and the exact causal mechanism of the harm does not need to be reasonably foreseeable.[173] Howells and Twigg-Flesner have proposed a liability approach that allows users to seek redress without having to face the difficulties associated with identifying the correct party responsible.[174] Ben-Shahar offers a view on fault-based liability in tort law and apportionment problems by proposing safety score liability, which is commensurate with a party's habitual propensity to behave unsafely.[175] This proposal relies on measurable and predictable actions of individuals, which are identified through a machine learning algorithm that calculates scores to assess the likelihood of causing accidents that result in harm.[176] The implementation of safety score liability would be suitable for assessing the responsibility of content creators and software developers, although the algorithmic system that calculates the scores should be trained with accurate data on the parties involved, which may prove difficult in the case of anonymous malicious deepfake scammers. Some commentators have found limitations in liability law, pointing out that it is 'too blunt an instrument for allocating responsibility to meet society's needs, and that any attempt to improve that allocation via law and regulation needs to focus on the implementation of appropriate governance systems by GenAI supply chain members and also to recognise that risk prevention is an impossible target, instead focusing on mitigation of risks to a socially acceptable level'.[177]

[170] Anna Beckers and Gunther Teubner, 'Responsibility for Algorithmic Misconduct: Unity or Fragmentation of Liability Regimes?' (2023) 25 Yale Journal of Law and Technology 76, 94–95.

[171] Bart Custers, Henning Lahmann and Benjamyn I Scott, 'From liability gaps to liability overlaps: shared responsibilities and fiduciary duties in AI and other complex technologies' (2025) 40 AI & Society 4035.

[172] Alexandra Andhov and Jodi Gardner, 'Duties of Care for Algorithmic Decision-Making in the Financial Industry' (NUS Law and Fintech Conference, Singapore, 31 Jul to 1 Aug 2025).

[173] On this discussion see Law Commission, AI and the Law: A Discussion Paper (31 Jul 2025) 11 <https://lawcom.gov.uk/publication/artificial-intelligence-and-the-law-a-discussion-paper/> accessed 30 Mar 2026.

[174] Geraint Howells and Christian Twigg-Flesner, 'Interconnectivity and Liability. AI and the Internet of Things' in Larry A DiMatteo, Cristina Poncibò and Michel Cannarsa (eds), *The Cambridge Handbook of Artificial Intelligence* (Cambridge University Press 2022) 198.

[175] Omri Ben-Shahar, 'Safety Score Liability' (2025) 17 Journal of Legal Analysis 190.

[176] This regime proposes that parties are partially liable in proportion to their safety score, as has been noted, 'the higher the safety score, the lower the fraction of liability': ibid 196.

[177] Chris Reed and Keri Grieman, 'Responsibility for generative AI services – from aspiration to achievement' (2025) 7–8 <https://ssrn.com/abstract=5107667> accessed 25 Mar 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

We situate the shared liability regime within the law of causation in tort to identify responsible actors in the deepfake network among multiple causes of the same injury which arises in negligence and results from significant, foreseeable, and harmful risks.[178] This regime allows the identification of the burden of responsibility for manipulating the GenAI system across the supply chain while providing consumers with effective remedies to claim compensation where fraudulent practices compromise their decision-making process. The shared liability regime may have limitations in addressing the reasonable expectations of safety and reliability that an average consumer has regarding AI software.[179] The practical implementation of shared liability is analysed through the actions of the parties that increase the risk of faulty outcomes. The more fundamental problem is identifying certain actors at all, especially given limited enforcement resources, anonymity, and cross-border operations. These real-world constraints are essential to any feasible liability model.

Our proposal for a shared liability regime advances existing debates on the development of a more resilient AI risk framework, in which all parties are required to implement more robust verification technologies and protocols to protect consumers. In cases where victims are scammed, a shared liability regime would provide guidance on the allocation of liability among the relevant parties.

Liability for deepfake software developers

The missing parties being held liable in the deepfake scam supply chain are deepfake software developers. Such software can be used for legitimate and even beneficial purposes. This makes it difficult to impose blanket liability on all deepfake software developers. We consider the scenario of liability for software engineering in which developers hold responsibility for using methodologies which may be manipulated for harmful deepfake content (eg audio, visual images, advice). Developers of automated tools such as agentic integrated development environments (LLMs such as GitHub and Copilot X) can generate code to support human skills.[180] Software development methodologies require ongoing monitoring, adaptation, and refinement to avoid hacking and data leakage.[181] Developers can implement safeguards in software to prevent hackers and malicious users from executing cyberattacks. Although the safeguards can be evaded by a downstream user, a few proactive techniques can be adopted to monitor users performing harmful tasks, such as open-sourcing, pre-release stress-testing of models, structured model access for third-party auditors, and research APIs to facilitate external independent model evaluation.[182] These defensive strategies can be implemented by creating a group of 'red-team professionals' to provide effective safety evaluations of AI models.[183]

The shared liability regime can be implemented for developers based on the tort of negligence for failure to exercise due care in ensuring the safety and reliability of the software (generated code and automation) and in providing software updates. Developers should be responsible for code review

[178]Richard W Wright, 'Allocating Liability among Multiple Responsible Causes: A Principled Defense of Joint and Several Liability for Actual Harm and Risk Exposure' (1988) 21 UC Davis Law Review 1141. See also Omri Ben-Shahar, 'Causation and Foreseeability' in Gerrit De Geest (ed), *Encyclopedia of Law and Economics - vol 1* (2nd edn, Edward Elgar 2009) ch 3.

[179]Geraint Howells, 'Protecting Consumer Protection Values in the Fourth Industrial Revolution' (2020) 43 Journal of Consumer Policy 160; Peter Cartwright, 'Understanding and Protecting Vulnerable Financial Consumers' (2015) 38 Journal of Consumer Policy 119, 127.

[180]Marijn Janssen, 'Responsible governance of generative AI: conceptualizing GenAI as complex adaptive systems' (2025) 44 Policy and Society 45–46.

[181]Tim Howard, 'GenAI and software development: a new paradigm' (Defra Digital, 2 Jun 2025) <https://defradigital.blog.gov.uk/2025/06/02/genai-and-software-development-a-new-paradigm/> accessed 30 Mar 2026.

[182]Elizabeth Seger and others, 'Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives' (2023) Computers and Society 17.

[183]Red-team and research API solutions aim to foster transparency in the AI model while enhancing collaborative and multi-stakeholder efforts to evaluate the model components. See Chinmayi Sharma, 'Concentrated Digital Markets, Restrictive APIs, and the Fight for Internet Interoperability' (2019) 50 University of Memphis Law Review 441, 451–452.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

and debugging of software, although they often bargain with the online platforms, agreeing to distribute their software exclusively through big tech companies in exchange for reduced fees.[184] This suggested view is in line with article 50 of the EU AI Act, which places a number of obligations on providers and users of AI systems to enable the detection and tracing of AI-generated content.[185]

An interesting accountability regime exists in medical law, where software is viewed as a service rather than a product. This perspective arises from the shared fault liability for injuries resulting from the integration of AI systems into healthcare practice, which is distributed among physicians, hospitals, and product developers.[186] The EU Product Liability Directive (PLD)[187] includes software as a product for the purposes of no-fault liability, reframing it under the strict liability regime.[188] The impact is that producers of software can be held responsible for damage caused by their defective products, regardless of fault.[189] In the academic debate, it has been suggested that software has become an important element of products and is even remotely interconnected with them.[190] Therefore, the case for considering software like any other component has become stronger with the result that producers of the software itself could be liable in their own capacity whether as a supplier of software components or as a supplier of standalone software. On this view, it is observed that 'as AI is more pervasive and ubiquitous – just like other software – it might soon be found in a significant number of consumer products and services'.[191] However, the definition of software as a product remains cryptic in the PLD as it is not clear whether digital content with equivalent functions and software as a service are included in the scope of the revised product category.[192] Specifically, the Directive does not provide a clear demarcation between the concept of software as a service

[184]Fiona M Scott Morton, *Digital Platform Regulation: Making Markets Work for People* (Yale School of Management 2025) 135.

[185]Mar Negreiro, 'Scam calls in times of generative AI' (European Parliamentary Research Service PE 777.940, Oct 2025) <https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/777940/EPRS_ATA(2025)777940_EN.pdf> accessed 15 Apr 2026.

[186]Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. See also W Nicholson Price II, Sara Gerke and I Glenn Cohen, 'Liability for use of artificial intelligence in medicine' in Barry Solaiman and I Glenn Cohen (eds), *Research Handbook on Health, AI and the Law* (Edward Elgar 2024) 166.

[187]Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 Oct 2024 on liability for defective products and repealing Council Directive 85/374/EEC (Text with EEA relevance).

[188]Art 4(1) of the PLD. It is worth noting that, at the UK level, the Law Commission has embarked on a review of the legislation on liability for defective products. This review aims to update and substantially reform UK legislation, which dates back to the mid-1980s and was enacted as Part I of the Consumer Protection Act 1987, itself based on the EU Product Liability Directive adopted in 1985. Since then, a much wider range of products has emerged, particularly as a result of digitalisation. A large number of physical products now incorporate software, and many have the capability to connect to the internet or other networks. Furthermore, a wide array of digital products has developed in the form of applications and AI-based systems; however, none of these currently falls within the scope of existing product liability legislation in the UK. On this issue see Law Commission, 'Product Liability' <https://lawcom.gov.uk/project/product-liability/> accessed 20 Mar 2026.

[189]Recital 2 of the PLD indicates that 'liability without fault on the part of economic operators remains the sole means of adequately addressing the problem of fair apportionment of risk inherent in modern technological production'. Recital 13 makes clear that 'software is a product for the purpose of applying no-fault liability, irrespective of the mode of its supply or usage, and therefore irrespective of whether the software is stored on a device, accessed through a communication network or cloud technologies, or supplied through a software-as-a-service model'.

[190]Christian Twigg-Flesner and Geraint Howells, 'Adapting Consumer Law to New Technologies' in Roger Brownsword and Larry A Di Matteo (eds), *The Cambridge Handbook of the Governance of Technology* (Cambridge University Press 2026) 160.

[191]Przemysław Pałka and Agnieszka Jabłonowska, 'Consumer law and artificial intelligence' in Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (2nd edn, Edward Elgar 2025) 591.

[192]Teresa Rodriguez de las Heras Ballel, 'Civil liability and artificial intelligence: challenges, policy options and legal responses' in Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (2nd edn, Edward Elgar 2025) 482.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

(eg computer programs and AI systems) and software as a genuine digital service (eg digital offerings) when it is treated as a product in its own right. Recital 13 of the PLD provides some indication that software may be considered either as a standalone product or as integrated into other products as a component; however, it does not distinguish between 'software' and 'service', nor does it classify genuine digital services as related services (eg the supply of data), which are de facto excluded from the category of products.[193]

Software is capable of causing damage through its execution, and manufacturers should remain liable for defectiveness as a result of related services within their control.[194] As Howells and Twigg-Flesner observe, the term 'producer' is given a broad meaning under the PLD.[195] It goes beyond the manufacturer and can include even the producer of raw components, the manufacturer of components, and even importers of the products. We posit that deepfake software producers would fall within the PLD and would therefore be accountable for victims' losses.[196]

Article 12 of the PLD states that multiple economic operators can be held liable jointly and severally for the same damage; the liability is not reduced where the damage is caused both by the defectiveness of a product and by an act or omission of a third party.[197] Responsibility for the testing of high-risk AI systems is placed under article 60(9) of the EU AI Act, which requires that 'providers should be held liable for any damage caused in the course of their testing in real world conditions'. Limitations of the shared liability framework of the PLD emerge in relation to financial harms caused by faulty algorithmic procedures, which require an effective right to compensation and a redress scheme. Specifically, the shared liability approach may fail to capture responsibility for immaterial harms (eg discriminatory outputs, breaches of privacy, emotional losses), which necessitate enforcement mechanisms capable of awarding damages for defective products.[198]

Prevention measures to deter malicious actors also include government certification programmes and authorisation schemes for assessing the software before deployment on the platforms.[199] The licensing regime would test the technology used to ascertain whether the software components (eg, operating system, code, and data)[200] are aligned with safety and reliability requirements. It has been noted that 'regulators should only grant such licenses if applicants can show that the use of generative AI will provide consumers with financial guidance that is at least as accurate and well-tailored

[193] European Law Institute, Liability for Defective Software Under the New EU Product Liability Regime: 1st Supplement to the ELI Draft of a Revised Product Liability Directive (22 Jun 2026) 14–15 <https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/Liability_for_Defective_Software_Under_the_New_EU_Product_Liability_Regime.pdf> accessed 25 Jun 2026.

[194] Recital 50 of the PLD.

[195] Howells and Twigg-Flesner (n 174) 186–187.

[196] An example of software leading to significant damages in finance is the errors committed by Knight Capital Americas LLC in 2012. Its defective computer software led to the router sending four million orders onto the market to fill just 212 customer orders. The Securities and Exchange Commission held that Knight Capital failed to have adequate safeguards in place to limit the risks posed by its access to the markets; to prevent the entry of millions of erroneous orders, and to adequately assess the effectiveness of its controls. Strict liability, however, may hinder innovation and deepfake software has dual purposes. Arash Massoudi, 'Knight Capital glitch loss hits $461m' (Financial Times, 17 Oct 2012) <https://www.ft.com/content/928a1528-1859-11e2-80e9-00144feabdc0> accessed 1 Apr 2026; US Securities and Exchange Commission, 'SEC Charges Knight Capital With Violations of Market Access Rule' (28 Jul 2014) <https://www.sec.gov/newsroom/press-releases/2013-222> accessed 25 Mar 2026.

[197] Art 13 of the PLD.

[198] Sandra Wachter, 'Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26 Yale Journal of Law & Technology 671, 704.

[199] Anuragini Shirish and Shobana Komal, 'A Socio-Legal Inquiry on Deepfakes' (2024) 54 California Western International Law Journal 555–556.

[200] A software component is defined as 'a unit of composition with contractually specified interfaces and explicit context dependencies only. A software component can be deployed independently and is subject to composition by third parties'. See Clemens Szyperski and others, *Component Software: Beyond Object-Oriented Programming* (2nd edn, Addison-Wesley Pearson Education 2002) ch 1.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

as the guidance that is provided by more conventional symbolic AI'.[201] The rapid development of face-swapping AI technology raises questions about how to identify the scam.[202] The capacity for face-swapping to deceive is bolstered by voice-cloning applications. Voice-changing devices raise concerns about their use in various operations such as buying a product on the platform or creating voices and cloning them.[203] The significant challenge is that face-swapping can be used for both legitimate commercial purposes and malicious investment fraud. The question, therefore, is how the licensing system can regulate software with these dual uses. Proper safeguard measures such as regulatory requirements and compliance methods can be imposed on the software developer in order to detect the characteristics and purpose of synthetic media distribution and identify natural persons through fingerprint information.[204] It has been suggested that policy intervention should focus on the sociotechnical domain, such as the identification of content manipulations through authentication practices to verify the accuracy of the source of content and preserve the distribution chain.[205]

The Australian Scams Prevention Framework Act 2025 requires digital platforms to check that all advertisers of financial products have an Australian Financial Services Licence and take specific steps to verify new accounts.[206] Along with these preventive regulatory tools, tech companies have developed trial processes such as regulatory and digital sandboxes to mitigate the risks of harmful outcomes of AI models. The European Council defined regulatory sandboxes as follows:

> '[C]oncrete frameworks which, by providing a structured context for experimentation, enable where appropriate in a real-world environment the testing of innovative technologies, products, services or approaches … for a limited time and in a limited part of a sector or area under regulatory supervision ensuring that appropriate safeguards are in place'.[207]

Regulatory sandboxes such as a digital 'scale up box' can be implemented to monitor the quality and performance of GenAI applications.[208] These regulatory tools, launched by the FCA to test Fintech products, are designed to establish a safe environment to oversee technological software and foster innovation.[209] They aim to enhance cooperation between regulatory authorities

[201] Daniel Schwarcz, Tom Baker and Kyle Logue, 'Regulating Robo-Advisors in an Age of Generative Artificial Intelligence' (2025) 82 Washington & Lee Law Review 775, 806.

[202] Yi Yan, 'Deep Dive into Deepfakes – Safeguarding Our Digital Identity' (2023) 48 Brooklyn Journal of International Law 768.

[203] Liam James, 'Face-swapping: Why you can't trust a video call – and how scammers are taking advantage' (The Independent, 28 Feb 2025) <https://www.independent.co.uk/tech/deepfake-scam-face-swap-fraud-ai-b2706722.html> accessed 15 Apr 2026.

[204] Momina Masood and others, 'Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward' (2023) 53 Applied Intelligence 4013–4014.

[205] Ellen P Goodman, 'Synthetic Content: Default to Distrust' (2025) 75 Case Western Law Review 961. It is argued that 'defaulting to authentic content and requiring labelling for synthetic content would put the burden on those who want to authenticate content or provide provenance information for synthetic content'.

[206] Department of the Treasury, Scams Prevention Framework. Protecting Australians from scams (Commonwealth of Australia 2025) 5 <https://treasury.gov.au/publication/p2025-623966> accessed 20 Apr 2026.

[207] Council of the European Union, 'Outcome of Proceedings – Council Conclusions on Regulatory sandboxes and experimentation clauses as tools for an innovation-friendly, future-proof and resilient regulatory framework that masters disruptive challenges in the digital age' 13026/20 (16 Nov 2020), <https://data.consilium.europa.eu/doc/document/ST-13026-2020-INIT/en/pdf> accessed 20 May 2026.

[208] Walter G Johnson, 'Caught in quicksand? Compliance and legitimacy challenges in using regulatory sandboxes to manage emerging technologies' (2023) 17 Regulation & Governance 709.

[209] See FCA, 'Regulatory Sandbox' <https://www.fca.org.uk/firms/innovation/regulatory-sandbox> accessed 25 Jun 2026. The FCA launched a Supercharged Sandbox tool using NVIDIA accelerated computing and NVIDIA AI Enterprise Software to provide firms with access to better data, technical expertise and regulatory support to test AI systems. The Supercharged Sandbox aims to support innovative, early-stage AI projects in financial services. See FCA, 'Supercharged Sandbox'

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

and financial firms with a view to promoting best practices for transparency of automated systems.

An innovative regulatory approach against the risks of unexpected outcomes of technology applications is the FCA's consumer duty, which concerns a firm's conduct in relation to financial services.[210] The consumer duty rule sets the standard of care that firms should exercise towards customers in retail markets and holds manufacturers and distributors responsible for the assessment of foreseeable harms.[211] The duty applies across the distribution chain, including to the design or operation of products or services, and a firm's responsibilities depend on its role and ability to influence retail customer outcomes.[212] Manufacturers are required to monitor the characteristics of their products and whether they are fit for purpose, and distributors are under an obligation to have a clear understanding of the target market and the way products operate.[213] Software developers and digital platforms are expected to use watermarking techniques to deploy reliable content authentication and provenance mechanisms to enable users to identify AI-generated content.[214] Firms need to share information about harmful content in technological applications with other relevant parties, in particular where a distributor identifies foreseeable harm or problems with the way a product or service is operating in practice.[215] The consumer duty can impose a primary obligation for firms to report fraudulent behaviours in deepfake scams relating to GenAI products and implement internal controls to detect potentially harmful software.[216] The FCA duty rule can be implemented by firms to reduce the vulnerability of GenAI tools, although compliance requirements, such as risk management processes and transparency obligations, may not be effective in detecting

<https://fcainnovation.co.uk/wp-content/uploads/2025/06/supercharged-sandbox-participation-pack.pdf> accessed 20 Jun 2026; FCA, 'FCA allows firms to experiment with AI alongside NVIDIA' (9 Jun 2025) <https://www.fca.org.uk/news/press-releases/fca-allows-firms-experiment-ai-alongside-nvidia> accessed 10 Jun 2026.

[210]FCA, Final non-Handbook Guidance for firms on the Consumer Duty (Finalised Guidance FG22/5, Jul 2022) 35 <https://www.fca.org.uk/publication/finalised-guidance/fg22-5.pdf> accessed 18 Jun 2026. It is stated that 'where a firm is planning to alter or withdraw a product or service, they should consider whether it could lead to foreseeable harm for their customers or a specific group of customers (such as customers with characteristics of vulnerability) and take steps to mitigate the impact of the potential harm'.

[211]Iris H-Y Chiu and Wai-Yee Wan, 'Constructing a Taxonomy of Financial Consumer Protection Policy and Assessing the New Consumer Duty in the United Kingdom's Financial Sector' (2024) 7 Cardozo International & Comparative Law Review 465.

[212]The duty applies in the distribution chain, including the manufacture, provision, sale, and ongoing administration and management of a product or service to the end retail customer. Under the consumer duty, firms are required to assess, test, understand and evidence the types of products and services that they offer. A firm is not required to go beyond its role in the distribution chain nor monitor or take responsibility for a separate firm's compliance. See Robin Henry and Abbie Coleman, 'Could there be consumer duty failure claims within your distribution chain?' (Financial Times, 11 Oct 2023) <https://www.ftadviser.com/regulation/2023/10/11/could-there-be-consumer-duty-failure-claims-within-your-distribution-chain/> accessed 10 Jun 2026.

[213]Manufacturers and distributors need to identify foreseeable harms where products or services were poorly designed or were distributed widely to customers for whom they were not designed.

[214]Watermarking tools are used for content authentication, data monitoring, indicating authorship, and protecting copyright. See Tambiama Madiega, 'Generative AI and watermarking' (European Parliamentary Research Service PE 757.583, Dec 2023) <https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2023)757583> accessed 21 Jun 2026.

[215]Manufacturer firms need to inform distributors of the characteristics of a product or service, its target market, and the value it is intended to provide to customers, and, to support manufacturers reviewing a product or service, distributors need to provide relevant information to them. Firms are responsible only for their own activities and are not required to oversee the actions of other firms in the distribution chain. However, they may be held responsible in certain circumstances for the actions of other parties where they exercise material influence over customer outcomes.

See FCA, 'Consumer Duty implementation: good practice and areas for improvement' (Dec 2025) <https://www.fca.org.uk/publications/good-and-poor-practice/consumer-duty-implementation-good-practice-and-areas-improvement> accessed 15 May 2026; 'Consumer Duty information for firms' (Aug 2026) <https://www.fca.org.uk/firms/consumer-duty-information-firms> accessed 15 Sep 2026.

[216]FCA, Immediate areas for action and further plans for reviewing FCA requirements following introduction of the Consumer Duty (Feedback Statement FS25/2, Mar 2025)

<https://www.fca.org.uk/publication/feedback/fs25-2.pdf> accessed 25 May 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

(and avoiding) potentially harmful outputs.[217] Further, the consumer duty rule requires firms to conduct an assessment of the AI model and its use and operation before deploying the AI model in the market.[218]

Risk management and verification measures for deepfake software developers and online social media platforms

We argue that deepfake software companies and online social media platforms should be held liable for negligence if they do not adequately implement risk management processes and verification measures to prevent harmful misuse such as deepfake scams, although we acknowledge that these solutions may prove ineffective in detecting in time the unintended malicious actions of users. Software developers should use biometric and identity verification to prevent unauthorised use of someone's likeness by requiring consent verification before processing images or voice samples of identifiable individuals. In order to verify the integrity of content, watermarking techniques include hashing trace files with a small string of numbers that are lost if the audio or video is deepfaked.[219] Failure to employ digital forensic techniques in GenAI amounts to gross negligence, which may result in personal liability for individuals accountable for careless supervision of content authenticity.[220] Digital forensic techniques employ blockchain technology, which can verify the origins and distribution of videos by storing digital signatures in a ledger which is difficult to manipulate.[221]

The UK Cyber Security and Resilience Bill 2024,[222] which updated the Network and Information Systems (NIS) Regulations 2018,[223] established a set of provisions that impose obligations on managed service providers (IT service providers and managed security services),[224] holding responsible those firms for failure to take effective technical and organisational measures (eg contractual requirements, security checks, or continuity plans) and to manage risks posed to the security of the network,

[217]Teresa Rodríguez de Las Heras Ballell, 'Mapping Generative AI Liability Cases in the EU Legal Framework' in Mimi Zou and others (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge University Press 2025) 106.

[218]FCA, 'FCA helps firms to test AI safely' (Dec 2025) <https://www.fca.org.uk/news/press-releases/fca-helps-firms-test-ai-safely> accessed 22 May 2026; 'Our Consumer Duty focus areas' (May 2026), <https://www.fca.org.uk/publications/corporate-documents/consumer-duty-focus-areas> accessed 28 May 2026; 'Consumer Duty implementation: good practice and areas for improvement' (Dec 2025), <https://www.fca.org.uk/publications/good-and-poor-practice/consumer-duty-implementation-good-practice-and-areas-improvement> accessed 26 Jun 2026.

[219]Liam Kearns, Abu Alam and Jordan Allison, 'Synthetic Media Authentication Threats: Detection using a Combination of Neural Network and Blockchain Technology' (2025) 36 Transactions on Emerging Telecommunications Technologies e70225.

[220]David Atkinson and Jacob Morrison, 'A Legal Risk Taxonomy for Generative Artificial Intelligence' (2024) Computers and Society arXiv:2404.09479.

[221]Mika Westerlund, 'The Emergence of Deepfake Technology: A Review' (2019) 9 Technology Innovation Management Review 46.

[222]Cyber Security and Resilience (Network and Information Systems) Bill, HL Bill 49.

[223]The Network and Information Systems Regulations 2018, 2018 No 506.

[224]The Bill targets managed service providers by expanding the scope of security duties imposed on digital service providers (including online marketplaces, online search engines, and cloud computing services). The term 'managed service providers' refers to the ongoing provision of management support, active administration, or monitoring of IT systems, infrastructure, applications, or networks, including activities related to cybersecurity that involve a network connection or access to the customer's network and information systems. The Bill also introduces supply chain security requirements by embedding duties relating to supply chain security for operators of essential services within the legislative framework. Senior managers and directors are responsible under the Cyber Governance Code of Practice, which complements the Bill 2024 in governing cyber security risks: Department for Science, Innovation and Technology, *Cyber Governance Code of Practice* (2025) <https://www.gov.uk/government/publications/cyber-governance-code-of-practice/cyber-governance-code-of-practice> accessed 15 Apr 2026. See the editorial 'In cyber attacks, humans can be the weakest link' (Financial Times, 26 May 2025), <https://www.ft.com/content/4349b16a-8ec1-44d9-a295-3a51523805a8> accessed … .

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

infrastructure, and information systems.[225] The concept is similar to our proposal of imposing liability on those responsible for negligence. The Chinese Provisions on the Administration of Deep Synthesis of Internet-based Information Service 2023 imposed specific obligations and responsibilities on service providers, users, and online platforms for the management of synthetic media technologies.[226] Alongside these legislative frameworks, supervisory authorities should develop response programmes tailored to address the threats posed by deepfake manipulation. These programmes would establish procedures for identifying, assessing, and mitigating the harmful consequences of malicious deepfake content, supported by regulatory tools designed to enhance coordination between enforcement mechanisms, intelligence agencies, and other relevant stakeholders.

We submit that deepfake software developers and online social media platforms should implement risk management techniques to identify, assess, and mitigate the harmful impact of deepfake AI content. Where AI-generated content is detected but does not violate platform policies (eg parody), platforms should apply a prominent, immutable label (eg 'AI-generated' or 'Synthetic Media'), which remains visible even when the content is shared or downloaded. Furthermore, if a user receives a message containing phrases such as 'Bitcoin', 'investment opportunity', or 'send money', a pop-up alert could warn: 'Be careful. Scammers often promise guaranteed returns. Never send money to someone you have only met online.' Similarly, if a user shares a video identified as a potential deepfake scam, a warning could appear stating: 'This media has been identified as potentially synthetic. Learn more about deepfakes'.

Liability could also be imposed on software developers by implementing reasonable safeguards. For example, Microsoft launched safety benchmarks for developers to rank iterations from a range of providers with a view to building trust with cloud customers.[227] The safety metric evaluates whether a model can be used for malicious purposes. This testing programme based on rankings enables users to understand the risks posed by AI applications.[228] It is an innovative mechanism to monitor GenAI products by ensuring they cannot be used to create harmful content. The safety system evaluates the fairness, biases, and mistakes of AI products, and the risks of fraudulent hallucinated outputs in investment scams.[229]

In 2025, the Singapore Government launched testing pilot schemes, namely the Global AI Assurance Pilot,[230] the Joint Testing Report with Japan,[231] and the AI Safety Red Teaming Challenge Evaluation Report[232] to enhance the best practices of GenAI governance. Red teaming calls for scenario planning and adversarial testing of programmes in order to identify software developers'

[225]Jonathon Ellison, 'Cyber Security and Resilience Policy Statement to strengthen regulation of critical sectors' (National Cyber Security Council, 1 Apr 2025) <https://www.ncsc.gov.uk/pdfs/blog-post/cyber-security-resilience-bill-policy-statement.pdf> accessed 30 Apr 2026.

[226]The Chinese Provisions on the Administration of Deep Synthesis of Internet-based Information Service. For commentary see Yinuo Geng, 'Comparing "Deepfake" Regulatory Regimes in the United States, the European Union, and China' (2023) 7 Georgetown Law Technology Review 157, 169–170.

[227]Rafe Uddin and Cristina Criddle, 'Microsoft to rank 'safety' of AI models sold to cloud customers' (Financial Times, 7 Jun 2025) <https://www.ft.com/content/02f39b33-fa6e-4bb7-b1f4-8171b50738af> accessed 20 Jun 2026.

[228]Liang Yu and others, 'Measuring the quality of generative AI systems: Mapping metrics to quality characteristics - Snowballing literature review' (2025) 186 Information and Software Technology 107802.

[229]David Krause, 'Mitigating Risks for Financial Firms Using Generative AI Tools' (2023) <https://ssrn.com/abstract=4452600> accessed 18 May 2026.

[230]See AI Verify Foundation, 'Global AI Assurance Pilot' <https://aiverifyfoundation.sg/ai-assurance-pilot/> accessed 28 Mar 2026.

[231]See International Network of AI Safety Institutes, *International Network of AI Safety Institutes Joint Testing Exercise: Improving Methodologies for AI Model Evaluations Across Global Languages* (2025) <https://sgaisi.sg/wp-api/wp-content/uploads/2025/03/International-Network-of-AI-Safety-Institutes-Joint-Testing-Exercise-Improving-Methodologies-for-AI-Model-Evaluations-Across-Global-Languages.pdf> accessed 18 May 2026.

[232]See IMDA, 'Singapore AI Safety Red Teaming Challenge' (Evaluation Report, Feb 2025) <https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/singapore-ai-safety-red-teaming-challenge-evaluation-report.pdf> accessed 22 May 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

deceptive behaviour.[233] The Monetary Authority of Singapore prescribes explicit duties for banks and telecommunications providers under the Shared Responsibility Framework, acting in coordination with the Infocomm Media Development Authority (IMDA) when allocating responsibility for losses arising from a defined scope of phishing scams.[234] The Supreme Court of Singapore has attributed liability for AI-generated content to users, thereby complementing the Shared Responsibility Framework and the Protection from Scams Act 2025.[235] In parallel, English courts have held that there is a professional duty to use AI tools 'with an appropriate degree of oversight, and within a regulatory framework that ensures compliance with well-established professional and ethical standards'.[236]

Under the OSA 2023, senior managers of online social media platforms may incur criminal liability if they fail to implement proportionate systems and processes designed to prevent the publication and hosting of fraudulent advertising. Within the shared liability model, we posit that, where the scale of harm justifies it (as in the Silk Road operation), content creators should also be held liable for deepfake scams. We acknowledge the challenges associated with uncovering the identities of anonymous content creators, which is why we have focused on the liability of software developers and online social media platforms. We further contend that deepfake software companies and online social media platforms should be held liable for negligence where they fail to adequately implement risk management processes and authentication methods designed to prevent harmful misuse, such as deepfake scams.

We suggest that online social media platforms should adopt protocol design and governance policy guidelines to mitigate the harmful risks posed by AI-generated content while, in parallel, assessing user expectations as to whether the purpose of such content is to mislead or distort beliefs. Corporate governance boards (including senior management, directors, executives, and compliance officers) of software companies and online platforms should be trained and fully informed about the reputational risks that the use of defective GenAI models may pose to business operations.[237] Poor monitoring of expected outputs and a limited understanding of AI decision-making processes can lead to compliance failures and flawed corporate reporting, thereby affecting the expectations of investors and market stakeholders.

Conclusion

Advances in generative computing software provide firms with greater opportunities to create smarter services for customers. However, they also provide scammers, fraudsters, and creators of deepfakes with more powerful tools.[238] GenAI can disseminate inaccurate information and unintended 'hallucinations' that are not factually reliable, and it is often difficult to trace how a particular output has been generated.[239] Malicious deepfakes operate at different stages of the AI supply chain through digital platforms, thereby complicating the identification of fraudulent activity.

[233] Orly Lobel, 'The AI Regulatory Pyramid: A Taxonomy & Analysis of the Emerging Toolbox in the Global Race for the Regulation and Governance of Artificial Intelligence' (2025) 57 Loyola of Los Angeles Law Review 890–891.

[234] IMDA, 'MAS and IMDA announce implementation of Shared Responsibility Framework from 16 December 2024' (24 Oct 2024) <https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/implementation-of-shared-responsibility-framework> accessed 24 Jun 2026.

[235] Protection from Scams Act 2025 (No 1 of 2025).

[236] *R. (on the application of Ayinde) v Haringey LBC* [2025] EWHC 1383 (Admin) 5.

[237] Patrick J O'Malley, 'Generative AI Systems and Corporate Governance, Compliance and Liability. Rethinking Director and Officer Roles in Light of a New World of Technological, Legal and Ethical Challenges' in Mimi Zou and others (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge University Press 2025) 369, 382.

[238] Stephen Bush, 'Are we human or are we spammer?' (Financial Times, 24 Jun 2025) <https://www.ft.com/content/07ed552a-e681-472a-8eba-c4b3a7938027> accessed 25 Jun 2026.

[239] Cristina Criddle, 'How to get the best out of AI' (Financial Times, 10 Jun 2025) <https://www.ft.com/content/d2f1fa02-025c-41ff-814f-00f22ed5c6d3> accessed 15 Jun 2026.

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049

The time is ripe for regulatory intervention. In this article, we propose a shared liability regime to attribute responsibility to different actors involved in the creation and dissemination of fraudulent AI-generated content. This regime would enable a more balanced approach to compensation, whereby the parties involved would bear losses proportionately, depending on whether and to what extent they contributed to the scam. The FCA's consumer duty can complement the shared liability regime by requiring firms to avoid foreseeable harm to customers arising from the behaviour of AI models. However, the UK regulatory framework does not provide adequate responses to the risks of automated manipulation of information in synthetic media, as the impersonation of individuals by scammers exacerbates unfair practices against users. This suggests that UK legislation should consider specific liability approaches that account for the timing and nature of deepfake scams, which may undermine the safety and reliability of outcomes, while also imposing responsibility on online social media platforms to compensate victims. The normative experience of Singapore in tackling online scams and sharing the costs of fraud may offer a useful policy model for the UK Government in establishing mandatory rules for compensating losses and allocating liability among multiple parties involved in malicious deepfake activities.

Acknowledgements. The authors are grateful to the reviewers for their constructive and insightful comments on earlier drafts.

Competing interests. The author(s) declare none

Cite this article: Alison Lui and Andrea Miglionico, 'AI-Generated Deepfake Financial Scams: A Missing Liability Regime for Consumer Protection Frameworks' (2026) Asian Journal of Comparative Law 1–32. https://doi.org/10.1017/asjcl.2026.10049

Downloaded from https://www.cambridge.org/core. IP address: 86.147.190.36, on 20 Oct 2026 at 07:09:13, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/asjcl.2026.10049