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

Alison Lui* and Andrea Miglionico**

The authors declare no conflicts of interest regarding this manuscript.

Competing interests: The author(s) declare none

Abstract

Generative artificial intelligence (GenAI) outputs such as deepfakes can be useful in 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 information provided to social media users. This article argues that a liability regime for deepfakes is missing in the consumer protection frameworks. It posits that regulatory interventions do not explicitly target GenAI software developers and online social media platforms, which are required to implement appropriate risk management safeguards to prevent unlawful activities. We contend that a shared liability regime for deepfakes between multiple actors involved could offer suitable protection for victims of online financial frauds and would target the beginning of the deepfake supply chain. The shared liability regime is complemented with the UK Financial Conduct Authority's consumer duty rule, which acts as a preventive monitoring action and enforcement mechanism to avoid foreseeable harm to customers in AI applications.

1. Introduction

The rapid proliferation of GenAI has resulted in sophisticated technologies that are readily available and easy for 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; and scammers use online social media platforms available to them and constantly adapt to evade enforcement.[2] The scams are using deepfakes of celebrities in videos to trusted professionals, which is causing a growing concern for the

* Liverpool John Moores University, Faculty of Law; A.Lui@ljmu.ac.uk.

** University of Reading, School of Law; a.miglionico@reading.ac.uk.

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

[1] GenAI is a technology based on the use of machine learning and multi-layered neural networks which perform through a large dataset. Ardi Janjeva, Alexander Harris, Sarah Mercer, Alexander Kasprzyk and Anna Gausen, 'The Rapid Rise of Generative AI: Assessing risks to safety and security' (December 2023), Research Report, Alan Turing Institute, Centre for Emerging Technology and Security, https://cetas.turing.ac.uk/sites/default/files/2023-12/cetas_research_report_-_the_rapid_rise_of_generative_ai_-_2023.pdf.

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

finance sector.[3] There is consensus that deepfakes can produce outputs that are not explicitly programmed and are occasionally inaccurate with the intent to cause distress to the recipient.[4] A survey of 2024 by Regula reveals that globally, half of all businesses have experienced fraud involving audio and video deepfakes.[5] Additionally, 66% of leaders believe deepfakes pose a serious threat to their business; on average, businesses across industries have lost nearly $450,000 to deepfakes.[6] This poses multifaceted challenges for public authorities, particularly around legal and ethical implications such as consent, misleading content and cybersecurity.[7]

The consequences for individual customers may be irreparable if deepfakes distort relevant information in providing financial advice or assistance crucial for the decision-making process.[8] Deepfakes are working like a charm for scammers because many social media users simply do not know what GenAI is capable of when it comes to making convincing impersonation videos.[9] Users watch deepfake videos and believe they are real, because many do not have the information or knowledge to question it. As reported, 'ordinary people use a product or service in the real world may diverge wildly from the designers' intentions in the lab'.[10] Deepfakes are increasingly being used on video calls to impersonate senior staff at corporate organisations, persuading other staff members to process payments which turn out to be fraudulent.[11] As the technology progresses, deepfakes have the potential to make romance scams even more convincing, and could potentially be used to manipulate images of friends and family relatives making requests for money.[12] This poses a social problem and could

[3] AI-driven scams are becoming more sophisticated, intensifying harmful risks for consumers. Scam warnings are issued to be aware of potential fraud involving people being offered loans for an upfront fee by an individual posing as a representative of financial firms. Offering any financial products on behalf of intermediaries is fraudulent, and victims of a scam are required to contact their bank immediately and report it to action fraud. See Joshua Franklin, Stephen Gandel and Akila Quinio, 'Who should foot the bill for cyber scams?' *Financial Times* (London, 11 December 2024), https://www.ft.com/content/577dff0d-b7a0-4bc3-a1b1-3ca828e5ba18.

[4] Lauren E. Willis, 'Deception by Design' (2020) 34(1) *Harvard Journal of Law & Technology* 115; Bobby Chesney and Danielle Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107(6) *California Law Review* 1753.

[5] See https://regulaforensics.com/news/deepfake-fraud-doubles-down/.

[6] Henry Patishman, 'The Impact of Deepfake Fraud: Risks, Solutions, and Global Trends' (15 November 2024), https://regulaforensics.com/blog/impact-of-deepfakes-on-idv-regula-survey/.

[7] Adrienne de Ruiter, 'The Distinct Wrong of Deepfakes' (2021) 34(4) *Philosophy & Technology* 1311.

[8] Michal Lavi and Hadar Yoana Jabotinsky, 'Seeing is Believing? Deepfakes in Financial Markets' (2026) 44(1) *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* (London, 2 May 2025), https://www.ft.com/content/fcbdc88f-bbfd-4338-915a-9ef7970b2123.

[10] John Thornhill, 'We are the new gremlins in the AI machine' *Financial Times* (London, 26 June 2025), https://www.ft.com/content/aaa57d4b-fee6-4109-87ac-9222d706fe07.

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

[12] Synthetic technologies embed sophisticated algorithmic systems such as generative adversarial networks, which operate through a generator to produce synthetic media, and a discriminator to train the data and develop realistic

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 evade detection. This pushes firms to develop new ways to make it harder for fraudsters to deceive others including using facial recognition technology.[14] There is a general assumption that business-to-consumer uses of digital platforms[15] can be sufficiently regulated through existing contracts, tort (e.g., fault-based 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 of 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

information for users. See Baoping Liu, Bo Liu, Tianqing Zhu and Ming Ding, 'A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models' (2025) *International Journal of Intelligent Systems*, https://doi.org/10.1155/int/9987535.

[13] Ebba Lundberg and Peter Mozelius, 'The potential effects of deepfakes on news media and entertainment' (2025) 40(4) *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.

[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.

[16] Kathryn E. Spier and Rory Van Loo, 'Foundations for Platform Liability' (2024) Harvard Public Law Working Paper 24-16, https://ssrn.com/abstract=5015344; Gökçe Kurtulan-Güner, 'Platform Liability: Quo Vadis?' (2020) 9(3) *Journal of European Consumer and Market Law* 275; James Grimmelmann and Pengfei Zhang, 'An Economic Model of Online Intermediary Liability' (2023) 38(3) *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(2) *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(6) *Modern Law Review* 1390.

[18] Jane C. Ginsburg and Graeme W. Austin, 'Regulating Deepfakes at Home and Abroad' (2025) 48(3) *Columbia Journal of Law & the Arts* 297; Felipe Romero Moreno, 'Generative AI and deepfakes: a human rights approach to tackling harmful content' (2024) 38(3) *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(2) *Hastings Law Journal* 293; Yi Yan, 'Deep Dive into Deepfakes - Safeguarding Our Digital Identity' (2023) 48(2) *Brooklyn Journal of International Law* 767.

[19] It is worth noting that AI-generated content is not protected by copyright in most countries, unless substantial human input is involved, and even so, there is no guarantee that it attracts copyright protection. This questions who owns the rights to this AI-generated material, which mostly depends on 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* (Singapore: World Scientific Publishing 2025) 197.

collection and model training stages, which impose legislative frameworks to focus on strengthening licensing, transparency and enforcement regulatory measures.[20] Investigations are very costly and time-consuming.[21] Criminal penalties would not be suitable to deter deepfake content creators, as they require proof beyond a reasonable doubt of the intention to deceive consumers.[22] Similarly, social media platforms release their own policies for removing or banning deepfakes, which use '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 an articulated set of penalties on companies that fail to protect the data of citizens.[23] The European Artificial Intelligence Act (EU AI Act)[24] and Digital Services Act (DSA) 2022[25] show limitations to accommodate liability for AI-generated harms.[26] The EU legislation has not provided a clear definition of malicious deepfakes, making meaningless any responsibility of the deepfake supply chain. Under the UK's Online Safety Act (OSA) 2023,[27] technology companies must set performance targets to remove illegal material quickly when they become aware of it and test algorithms to make illegal content harder to disseminate.[28] 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'[29] and 'operate the

[20] House of Lords - Communications and Digital Committee, 'AI, copyright and the creative industries' (6 March 2026) 4th Report of Session 2024–26, HL Paper 267, https://publications.parliament.uk/pa/ld5901/ldselect/ldcomm/267/267.pdf.

[21] Dennis Crouch, 'Using Intellectual Property to Regulate Artificial Intelligence' (2024) 89(3) *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(3) *North Carolina Journal of Law & Technology* 447. See also Barry Scannell, 'Who owns the copyright for AI work?' *Financial Times* (London, 24 August 2025), https://www.ft.com/content/74b1841f-bf57-4934-a06a-3611d61e4319.

[22] Jacquelyn Sherman, 'A Feast of Fraud: How International Hesitations to Regulate Deepfakes Are Creating a Buffet for Financial Criminals' (2025) 56(1-2) *George Washington International Law Review* 91.

[23] Regulation (EU) 2016/679. See also Yi Yan, 'Deep Dive into Deepfakes - Safeguarding Our Digital Identity' (2023) 48(2) *Brooklyn Journal of International Law* 767.

[24] Regulation (EU) 2024/1689.

[25] Regulation (EU) 2022/2065.

[26] Guido Noto La Diega and Leonardo C.T. 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(1) *International Journal of Law and Information Technology*, https://doi.org/10.1093/ijlit/eaae021; Martina J. Block, 'A Critical Evaluation of Deepfake Regulation through the AI Act in the European Union' (2024) 13(4) *Journal of European Consumer and Market Law* 184.

[27] See https://www.legislation.gov.uk/ukpga/2023/50/contents.

[28] Beatriz Kira, 'When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act' (2024) 54 *Computer Law & Security Review*, https://doi.org/10.1016/j.clsr.2024.106024.

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

service using proportionate systems and processes designed'[30] so that users cannot view fraudulent content and that platforms remove such content as soon as possible. Nevertheless, the OSA 2023 does not impose direct liability on digital platforms towards victims of fraudulent scams.

A shared liability regime for GenAI scams would target the beginning of the deepfake supply chain. It would prescribe 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 on AI recommendations and screening potential malicious deepfakes in financial products.[31] Singapore Courts have adopted a comprehensive binding Guide on GenAI use to verify the accuracy and appropriateness of all AI-generated outputs.[32] 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 of deepfake scams when the supply chain for the scam is highly complicated?

The UK Financial Conduct Authority (FCA)'s consumer duty rule, which established regulatory standards against firms' unfair practices, can provide a solution to the exploitation of financial scams in the use of GenAI tools.[33] The duty mandates financial firms to evaluate their products, services and processes in view to protect customers in business relationships and remedy foreseeable harms.[34] 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 consumers from achieving reliable outcomes. However, the duty rule is primarily a financial regulation principle and does not explicitly apply across all sectors. It might not be suitable to mitigate detrimental generated outputs of deepfake scams.

---

[30] OSA 2023, s 27(3).
[31] Jennifer Calver, Peter Church, Jonathan Ford and Kim Rust, 'AI in financial services - the legal and regulatory landscape' in Jelena Madir (ed), *FinTech: Law and Regulation* (3rd edn., Cheltenham: Edward Elgar 2024) ch.16.
[32] See 'Guide on the Use of Generative Artificial Intelligence Tools by Court Users' Registrar's Circular No. 9 2024,
https://www.judiciary.gov.sg/docs/default-source/circulars/2024/registrar's_circular_no_9_2024_state_courts.pdf?sfvrsn=d038ec05_1.
[33] FCA, 'A new Consumer Duty. Feedback to CP21/36 and final rules' (July 2022), Policy Statement PS22/9, p.40, https://www.fca.org.uk/publication/policy/ps22-9.pdf.
[34] Sandra Booysen, 'Protecting Consumers from Payment Fraud: An International Perspective' paper presented at the 2nd International Conference the Responsible Consumer in the Digital Age: 'Evolving Perspectives on Consumer Protection, Sustainability, and AI – the Nordics and Beyond', Copenhagen, 28 April 2025.

We argue that the shared liability regime for deepfakes would combine with the consumer duty rule as a regulatory conduct of business for 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, leads to the development of a preventive enforcement mechanism for the consumer protection framework. We contend that a shared liability regime between malicious deepfake content creators, software developers and online social media platforms would be effective as a form of deterrence to hold these actors liable for fraudulent scams.

The article proceeds as follows. We begin with an introduction of GenAI techniques, illustrating how they elaborate content and produce outputs that are not explicitly programmed. Section 2 examines the regulatory framework of deepfakes in the UK, which is the focus of the article. However, we also discuss the regulatory frameworks of deepfakes in the EU and Singapore as a comparison. In the UK, we place 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, also adopts a content-based approach similar to the OSA 2023. Further, comparison with Singapore is instructive because it has a lower threshold for scams and malicious cyber activities than 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' 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 providers faster remedies to scam victims than in the UK.

Section 3 depicts the liability contours of deepfake financial scams, analysing the challenges to identifying the responsible party for GenAI fraud. This section addresses the liability attribution to deepfake content creators and the interaction with online social media platforms: it investigates the applicable regulatory standard to hold liable the fraudsters and to ensure effective consumer protection through compensation schemes for scam losses.[35] Platform responsibility in investment scams is also considered along with potential challenges in imposing liability on scam victims. This leads to the elaboration of a framework which

[35] In Australia, the Scams Prevention Framework Act 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. See https://www.legislation.gov.au/C2025A00015/asmade/text.

establishes obligations on the various actors in the deepfake supply chain. Section 4 advances a policy proposal for a shared liability regime to prevent multiple actors from disseminating malicious deepfakes in financial investments. We identify potential scenarios which allocate responsibility to content creators, software developers, 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 which created the deepfake software enabling deepfake scams. Section 5 concludes and outlines final remarks.

2. 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.[36] While it creates new duties for internet service providers,[37] it has several limitations regarding 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-financially related offences under Schedule 7 such as pornography and sexual abuse compared to fraud.[38]

Further, different types of harms call for different legal treatment. With intentional misuse of generative AI tools such as deepfake voice cloning or synthetic identity creation 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

[36] OSA 2023, s 4(2), 4(5) and 4(6).
[37] The OSA 2023 imposes extensive duties on internet service providers in relation to online communication, transparency reporting, content moderation and verification of user's ages. Department of Science, Information and Technology, 'Guidance Online Safety Act: Explainer' (Department of Science, Information and Technology, 24 April 2025), https://www.gov.uk/government/publications/online-safety-act-protection-of-children-codes-of-practice-explanatory-memorandum.
[38] ibid.

obligations under the OSA 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.[39] 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 misuse by users and governing malfunction, whilst unintended harms concern rectifying design defects in AI outputs.

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.[40] The OSA currently adopts a content-based regulatory approach, where it imposes duties on regulated services regarding the content uploaded on platforms. The OSA requires platforms to have appropriate systems in place to meet the requirements of the OSA and tackles the aspect of removing illegal content. However, this neither directly tackle human vulnerabilities nor hold the responsible actor(s) liable to a criminal standard. Criminal liability would 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,[41] the OSA 2023 criminalises fraudulent advertising, which can be either a Category 1 or Category 2A regulated service, depending on the threshold conditions regarding the type of service and number of users.[42] Under a Category 1 service, internet service providers must comply with more duties than those 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 service. Companies

[39] Alex Casey, 'Generative AI's Duty to Deal Dilemma' (2026) 36(1) *Albany Law Journal of Science & Technology*, http://dx.doi.org/10.2139/ssrn.5345913.
[40] Victoria Nash and Lisa Felton, 'Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulation' (2024) 16(4) *Policy and Internet* 818.
[41] A customer from NatWest bank in the UK lost £150,000 in a deepfake video scam where Martin Lewis, a money savings expert, was impersonated. Tali Ramsey, 'Scam alert: deepfake videos are on the rise' (17 July 2024) *Which?*, https://www.which.co.uk/news/article/scam-alert-deepfake-videos-are-on-the-rise-aIlx30B9keja.
[42] 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.

must also remove it 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 quartile 4 of 2021[43] while Instagram removed 43.8 million pieces of content in the same period.[44] 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, the most damage they can cause is when they are operating in conjunction with other technological tools such as cryptocurrency wallets, QR codes and manipulation of platform engagement metrics.[45] 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 them to impersonate the official Tesla YouTube channel before launching a livestream.

The scammers directed the investors offline through a QR code to invest. Moving investors offline and onto their mobile phones has the benefit of less content moderation and scam detection.[46] Therefore, effective regulation requires thorough processes by user-to-user services to identifying the risks when deepfakes are working alongside generative AI, QR codes, search engines.[47] The highly convincing deepfake video is only one element that contributed to the fraud. Anonymity and opacity from moving cryptocurrency in 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

[43] TikTok, TikTok Community Guidelines Enforcement, Q3 2021.
[44] Ofcom, 'Just one in six young people flag harmful content online' (16 March 2023), https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/one-in-six-young-people-flag-harmful-content-online?language=en.
[45] Shaurya Malwa, 'Weaponized Trading Bots Drain $1M From Crypto Users via AI-Generated YouTube Scam' (7 August 2025), https://www.coindesk.com/tech/2025/08/07/weaponized-trading-bots-drain-usd1m-from-crypto-users-via-ai-generated-youtube-scam.
[46] Max Rizzuto, 'Crypto-scam hosts pop-up livestream featuring a deepfaked Elon Musk' (DFR Lab, 25 September 2024), https://dfrlab.org/2024/09/25/musk-deepfake-crypto-scam/.
[47] ibid.

often operate in a landscape of other technological tools and platforms, which makes it difficult to attribute the precise attribution of liability to individual actors involved.

The *modus operandi* of deepfake enabled financial scams is varied though. Deepfake scammers can directly target senior company leaders such as the widely reported case of deepfake financial scam of 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.[48] 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, not organisational system vulnerabilities.[49]

2.1 The Legal Framework in the EU

Recital 50(4) of the EU AI Act[50] sets out disclosure requirements for "deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake".[51] Labelling AI generated or manipulated content improves transparency and enables users to know, especially when several studies support that customers find it difficult to establish whether AI generated media is real or not.[52] Similar to 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 et al. show that labelling can reduce an individual's chance of relying on and engaging with misleading AI generated images.[53]

[48] Leng Cheng and Ho-Him Chan, 'Arup lost $25mn in Hong Kong deepfake video conference scam' *Financial Times* (London, 17 May 2024), https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea.
[49] Raúl Carrillo, 'Platform Money' (2024) 41(3) *Yale Journal on Regulation* 894, 940.
[50] Regulation (EU) 2024/1689.
[51] We use the term 'deployers' to indicate professional users of AI systems. Under Article 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".
[52] Kimberly T. Mai, Sergi D. Bray, Toby Davies and Lewis D. Griffin, 'Warning: Humans Cannot Reliably Detect Speech Deepfakes' (2023) 18(8) *PLoS ONE*, https://doi.org/10.1371/journal.pone.0285333. See also Hany Farid, 'Creating, Using, Misusing, and Detecting Deep Fakes' (2022) 1(4) *Journal of Online Trust and Safety*, https://www.tsjournal.org/index.php/jots/article/view/56.
[53] Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky and David G. Rand, 'Labeling AI-generated media online' (2025) 4(6) *PNAS Nexus*, https://doi.org/10.1093/pnasnexus/pgaf170.

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.[54] However, empirical studies demonstrated that authorship labels can be effective to enhance transparency, although they are unlikely to affect the persuasiveness of the labelled content, requiring alternative techniques to address risks posed by AI-generated information.[55] Wittenberg et al. opine that labels of 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.[56] Ternovski et al.'s experiments reveal that general labels warning the public about videos led to the public distrusting *all* videos are fake, even if they were real.[57] 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[58] also mandate online platforms to conduct thorough risk analyses, curb the dissemination of illegal content and label deepfakes and false information. But there are two weaknesses. First, 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 what is "illegal content" when the definition in Article 3(h) of the EU DSA 2022 is broad and vague. Whilst the content of child abuse 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 Steel et al. report that many stakeholders, such as non-governmental organisations are of the view that the users "are not

[54] 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/pgae403.
[55] Isabel O. Gallegos, Chen Shani, Weiyan Shi, Federico Bianchi, Izzy Gainsburg, Dan Jurafsky and Robb Willer, 'Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects' (July 2025), Policy Brief HAI Policy & Society, Stanford University, https://hai.stanford.edu/assets/files/hai-policy-brief-labeling-ai-generated-content.pdf.
[56] Wittenberg et al. (n 53).
[57] John Ternovski, Joshua Kalla, and Peter Aronow, 'The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments' (2022) 1(2) *Journal of Online Trust and Safety*, https://tsjournal.org/index.php/jots/article/view/28/15.
[58] Regulation (EU) 2022/2065.

necessarily capable of accurately identifying illegal content online".[59] The users can only make assumptions that the content is likely to be illegal.

Second, the EU DSA 2022 adopts a content-based approach, similar to the OSA in the UK. Content moderators compare the content on platforms against databases storing illegal content such as the Image Watch List by the Internet Watch Foundation.[60] 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.[61] Further, auto-detection software such as watermarking and labelling methods used by online platforms present weaknesses.[62] In particular, they are not very effective with shorter texts, content translated from other languages, or factual replies and they implicitly treat synthetic media as false and deceptive.[63] Whilst there are databases for illegal child abuse (IWF Watch List and Interpol),[64] there do not appear to be similar ones for fraud. Even for established databases for illegal content, Stockwell et al. suggest that it is important that these repositories use standardised metrics, such as harmonised labelling and classification methods.[65] Standardisation of repositories would be helpful given the global nature of the internet and its related problems.

2.2 The Regulatory Approach in Singapore

Singapore adopted a piece of legislation specifically targeting deepfakes in political elections. The Elections (Integrity of Online Advertising) Bill of 2024[66] bans the publication, boosting,

---

[59] Alexandre De Steele, Elise Defreyne, Hervé Jacquemin, Michèle Ledger and Alejandra Michel, 'Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform' (Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies, June 2020).
[60] See https://www.iwf.org.uk/.
[61] Sam Stockwell, Georgia Wake, Tooska Dargahi, Oluwaseun Ajao, Annabel Latham, Ahmed Danladi Abdullahi and Dan Sexton, 'Privacy-preserving Moderation of Illegal Online Content' (April 2025), The Alan Turing Institute, 14-15,
https://cetas.turing.ac.uk/sites/default/files/2025-04/cetas_research_report_-_privacy-preserving_moderation_of_illegal_online_content_0.pdf.
[62] Google uses a watermarking tool called SynthID Text. It helps developers to spot AI generated content: Geneva Internet Platform, Digwatch, 'Google unveils open-source watermark for AI text' (24 October 2024), https://dig.watch/updates/google-unveils-open-source-watermark-for-ai-text.
[63] 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 October 2021, https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/.
[64] See 'IWF Annual Data & Insights Report 2024', https://www.iwf.org.uk/annual-data-insights-report-2024/.
[65] Stockwell et al. (n 61).
[66] See https://sso.agc.gov.sg/Bills-Supp/29-2024/Published/20240909?DocDate=20240909.

sharing and reposting of deepfake content depicting election candidates. This adds to the protection given in the Online Safety (Miscellaneous Amendments) Act 2022[67] and Online Criminal Harms Act 2023 (the OCH Act).[68] The Online Safety (Miscellaneous Amendments) Act 2022 amends the Broadcasting Act 1994[69] by imposing legal duties on social media platforms. There are two main duties. First, social media platforms with significant impact must comply with Codes of Practice. Secondly, the Infocomm Media Development Authority (IMDA) can demand social media platforms to disable access or stop the 'egregious content' being transmitted to users.[70] '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 Online Safety Act and the Singapore 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.[71] This is a possible explanation of the additional role the government plays in dealing with harmful content.

Regarding scams and malicious cyber activities, the government only needs to suspect that there is online harm to issue a direction. For other activities, reasonable suspicion is required. This contrasts with the '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 purely suspecting there is a scam, or malicious cyber activities 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, digital literate and

[67] See https://sso.agc.gov.sg/Acts-Supp/38-2022/Published/20221221?DocDate=20221221#top.
[68] See https://sso.agc.gov.sg/Acts-Supp/24-2023/Published/20230807.
[69] See https://sso.agc.gov.sg/Act/BA1994.
[70] Part 10A of the Online Safety (Miscellaneous Amendments) Act 2022.
[71] Infocomm Media Development Authority, 'Online Safety Assessment Report 2024. Designated Social Media Services' (Singapore, 17 February 2025), p.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.

compliant".[72] Further, scam victims in Singapore suffer some of the highest losses in the world.[73]

Interestingly, a new government agency, named the Online Safety Commission, would provide faster remedies for victims.[74] Under the Online Safety (Relief and Accountability) Bill, scam victims can request perpetrators' information to commence legal proceedings.[75] They can also sue online platforms and page administrators for deepfakes, cyberbullying and sexual harassment. Online platforms must remove offensive content alerted 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 18 years old can use the apps.[76] These additional protections for scam victims along with the Broadcasting Act and Protection from Harassment Act 2014[77] indicate that the Singapore legislative framework adopted a strict regulatory approach to online scams.

3. The Challenges of Liability Attribution for Malicious Deepfake Scams

Establishing liability with GenAI deepfakes poses multiple challenges. Investigations are very costly and time-consuming to identify malicious content creators.[78] 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.[79] 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.[80] Once the victims enter the promotion codes, they are

---

[72] Owen Walker, "'Rich and naive': why Singapore is engulfed in a 'scamdemic" *Financial Times* (London, 26 May 2025), https://www.ft.com/content/3299cf7e-67bd-4654-8aa9-55fc24a66b63.
[73] Infocomm Media Development Authority (n 71).
[74] See 'Singapore's Online Safety Commission: A New Era in Combatting Online Harms' (7 March 2025), https://cyberindemnity.org/2025/03/singapores-online-safety-commission-a-new-era-in-combatting-online-harms/.
[75] See https://sso.agc.gov.sg/Bills-Supp/18-2025/Published/20251015?DocDate=20251015.
[76] Osmond Chia, "New online harms support centre to operate from 2026, will offer victims faster recourse" *The Straits Times* (8 March 2025), https://www.straitstimes.com/singapore/politics/new-online-harms-support-centre-to-begin-operation-in-2026.
[77] See https://sso.agc.gov.sg/Act/PHA2014.
[78] Bart van der Sloot and Yvette Wagensveld, 'Deepfakes: regulatory challenges for the synthetic society' (2022) 46 *Computer Law & Security Review*, 12-13, https://doi.org/10.1016/j.clsr.2022.105716.
[79] Daniel Thomas, 'WPP boss targeted by deepfake scammers using voice clone' *Financial Times* (London, 10 May 2024), https://www.ft.com/content/308c42af-2bf8-47e4-a360-517d5391b0b0. See also Cheng Leng and Chan Ho-him, 'Arup lost $25mn in Hong Kong deepfake video conference scam' *Financial Times* (Hong Kong, 17 May 2024), https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea.
[80] Matthew D. Weiner, 'Destined to Deceive: The Need to Regulate Deepfakes with a Foreseeable Harm Standard' (2024) 122(4) *Michigan Law Review* 771, 800. It is argued that 'the intent of an individual who creates a deepfake

asked to pay Bitcoin and share their personal details, and they never got anything in return. This is particularly problematic, given that the number of generative AI enabled scams has increased by 78% in 2023-24 and then by 456% in 2024-25.[81] In the Elon Musk cryptocurrency scam of 2024, Rizzuto used a blockchain explorer and was able to trace the addresses of the scammers' wallets.[82] 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 results in shielded liability to scam victims.

3.1 Liability on 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 to mislead 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'.[83] In the UK, there is a gap in legislation to hold malicious content creators liable. As mentioned in section 2, deepfake scams are intended harms which involve monitoring misuse by users and governing 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 potential cause of action. 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 them liable for failing to implement verification or safeguard measures such as disclaimers as to how much content has been altered.[84] The challenge though, is that many of such content creators are anonymous because of their nefarious activities. There are significant cost and jurisdictional barriers in uncovering the identity of these anonymous content creators.

---

is not debatable; certainly, after synthetically modifying an image or video, the creator did not genuinely believe that the content they shared was accurate'.

[81] See 'AI-enabled Fraud: How Scammers Are Exploiting Generative AI' (7 May 2025), TRM Blog, https://www.trmlabs.com/resources/blog/ai-enabled-fraud-how-scammers-are-exploiting-generative-ai.

[82] Rizzuto (n 46).

[83] Yisroel Mirsky and Wenke Lee, 'The Creation and Detection of Deepfakes: A Survey' (2020) 1(1) *ACM Computing Surveys*, https://doi.org/10.1145/3425780.

[84] Felipe Romero-Moreno, 'Deepfake detection in generative AI: A legal framework proposal to protect human rights' (2025) 58 *Computer Law & Security Review*, https://doi.org/10.1016/j.clsr.2025.106162.

Law enforcement techniques in the form of metadata analysis,[85] network traffic analysis,[86] financial trails[87] and behavioural analysis[88] can all assist to uncover the identity of the content creator. Courts can also issue search warrants for electronic devices to enter and search premises for evidence of serious arrestable offences under section 8 of the Police and Criminal Evidence Act 1984.[89] Metadata analysis of Ross Ulbricht's Gmail address and 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, of which 95% of the drugs were illegal.[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.

3.2 The Interaction between Content Creators and Online Social Media Platforms

Online social media platforms adopt GenAI technologies into their product offerings to enhance content creation workflows and empower content creators. For example, ChatGPT and Copilot are AI-powered software trained using large language models. They are designed

---

[85] 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.
[86] Pascal Tippe and Adrian Tippe, 'Onion Services in the Wild: A Study of Deanonymization Attacks' (2024) 4 *Proceedings on Privacy Enhancing Technologies* 291.
[87] Sarah Meiklejohn et al, 'A fistful of bitcoins: characterizing payments among men with no names' (2016) 59(4) *Communications of the ACM* 86.
[88] Efstathios Stamatatos, 'A Survey of Modern Authorship Attribution Methods' (2009) 60(3) *Journal of the American Society for Information Science and Technology* 538.
[89] See https://www.legislation.gov.uk/ukpga/1984/60/contents.
[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.
[91] Preet Bharara and Serrin Turner, 'Government Sentencing Submission United States v. Ross William Ulbricht, 14 Cr. 68 (KBF)' (U.S. Department of Justice, 26 May 2015), https://i.brayden.id.au/gov.uscourts.nysd.422824.256.0.pdf.
[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' (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.

to assist developers 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 as they are strategic tools for creator platforms for improving user engagement, enhancing content quality, and supporting creators with limited 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 make the platform more profitable. When both creators operate on the online social media platform and choose to use AI, the content quality improves, facilitating economic return 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 online 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 been of 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 to publish misleading advertisements enabled by scammers in view of substantial revenues. This agent relationship identifies a potential collusive behaviour between content creators and online social media platforms, which can inflate 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

[94] Noam Kolt, 'Algorithmic Black Swans' (2024) 101(4) *Washington University Law Review* 1177.
[95] Di Yuan, Manmohan Aseri, Vibhanshu Abhishek and Kartik Hosanagar, 'Generative AI Adoption by Creator Platforms' (2025), The Wharton School Research Paper, p.4, https://ssrn.com/abstract=5107730.
[96] Anna Yamaoka-Enkerlin, 'Disrupting Disinformation: Deepfakes and the Law' (2020) 22(3) *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(3) *International Journal of Research in Marketing* 411, 419.
[98] Nic Fildes, 'Meta sued by Australian regulator for allegedly 'misleading' crypto ads' *Financial Times* (Sydney, 18 March 2022), https://www.ft.com/content/132c7877-bd74-4957-925b-414f12ec6aa4.
[99] Jeannie Marie Paterson, 'Banks, Authorised Push Payment Scams and Models for Compensation' (2025) 99(2) *Australian Law Journal* 178, 182.
[100] Nash and Felton (n 40) 820.

management.[101] These platforms provide analytics tools that enable creators to measure their activities although it is questioned how to control content creation, advertising strategies and user response. 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] Creators know how to profit from their content: users find platforms as dispensers of valuable products from businesses, and they can both consume content and shop. However, this interaction can determine a pernicious cycle where creators are more prone to generate income, and platforms incentivise users' engagement to accommodate creators' desire for a more predictable remuneration of content adverts.[103]

3.3 Online Social Media Platforms' Responsibility for Deepfake Scams

Sophisticated adverts, including AI-generated deepfakes of audio and videos, persuade individuals to disclose personal data or invest in fraudulent schemes. Online social media platforms are the favourite conduit to profit from carrying the adverts and Big Tech companies seem unable to stop the number of deepfakes. According to the US Federal Trade Commission 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'.[104] Frauds originating on social media (e.g., Facebook, Instagram) indicate that platforms are the main method of contact and cryptocurrency is the most used method of payment.[105] We focus on regulated social media platforms such as Facebook and Instagram; largely unregulated platforms (such as Telegram, Signal, WhatsApp)[106] are outside the scope

[101] Noor Johnson, Matthew L Druckenmiller, Finn Danielsen and Peter L Pulsifer, 'The Use of Digital Platforms for Community-Based Monitoring' (2021) 71(5) *BioScience* 452.
[102] Miriam C Buiten, Alexandre de Streel and Martin Peitz, 'Rethinking liability rules for online hosting platforms' (2020) 28(2) *International Journal of Law and Information Technology* 139.
[103] Catalina Goanta, 'The New Social Media: Contracts, Consumers, and Chaos' (2023) 108 *Iowa Law Review Online* 118.
[104] Emma Fletcher, 'Social media a gold mine for scammers in 2021' (January 2022), https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/01/social-media-gold-mine-scammers-2021.
[105] See 'FTC Action Ends Ecommerce Empire Builders Online Business Opportunity Scam' (9 May 2025), https://www.ftc.gov/news-events/news/press-releases/2025/05/ftc-action-ends-ecommerce-empire-builders-online-business-opportunity-scam; 'FTC Takes Action to Stop Sprawling 'Growth Cave' Business Opportunity and Credit Repair Scam' (7 March 2025), https://www.ftc.gov/news-events/news/press-releases/2025/03/ftc-takes-action-stop-sprawling-growth-cave-business-opportunity-credit-repair-scam; 'FTC Announces Crackdown on Deceptive AI Claims and Schemes' (25 September 2024), https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes.
[106] Jacob Gursky and Samuel Woolley, 'Countering Disinformation and Protecting Democratic Communication on Encrypted Messaging Applications' (June 2021), The Brookings Institution of Foreign Policy, https://www.brookings.edu/articles/countering-disinformation-and-protecting-democratic-communication-on-encrypted-messaging-applications/.

of this article. Public regulators and policymakers call for action to protect consumers as fraud causes severe harm to individuals and the stolen money goes to serious organised crime groups mostly hidden in the crypto platforms.[107]

Technology frauds urge more collaboration between the public and private sectors, and a 'systemic' solution is required as scammers adopt new methods and increase unauthorised losses which exacerbate a systemic problem.[108] Online social media platforms should bear a legal duty to check investment advertisers are authorised, and a legal duty not to provide advert 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 advert content.[109] Further, online social media platforms should be legally required to check that an advertiser is authorised by a regulator to sell financial services and block them in case there is not.[110] 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 DSA, although this is not equivalent to an obligation to control the use of the platform, but it seems more like a due diligence test which could result in a ticking-the-box exercise.[111] This solution would not solve the problem of which party should be liable for the harm done by such frauds, but the fact that sellers of financial products must usually be registered with regulators would incentivise to blocking of a particularly harmful online fraud.

We argue that a regulatory framework that places a shared responsibility for deepfake financial scams on malicious content creators and online social media platforms would provide a redress scheme for scam losses; it would impose obligations on them to compensate victims of fraud. Such a solution might be feasible in the case that a content creator intentionally trains an AI model to deceive the outcome and online platforms use the AI outputs to distort or alter

[107] Joshua Oliver, 'The lawless world of crypto scams' *Financial Times* (London, 19 September 2022), https://www.ft.com/content/5987649e-9345-4eae-a4b8-9bfb0142a2ab.
[108] Maisie Grice, 'UK fraud 'blight' prompts calls for action to protect consumers' *Financial Times* (London, 27 May 2025), https://www.ft.com/content/c877784c-f416-4c7b-943b-dfb35a01233c.
[109] See the editorial 'How to block the financial scammers on social media' *Financial Times* (London, 18 May 2025), https://www.ft.com/content/93d29681-c242-4d0f-9789-41dbbebea906.
[110] Akila Quinio and Martin Arnold, 'Social media must do more to tackle payment fraud, says UK regulator' *Financial Times* (London, 14 October 2024), https://www.ft.com/content/1e77ae86-a0c2-47ee-804b-cbe50765a9a8.
[111] Stefano Faraoni, 'Why Generative AI Is Not Cyrano de Bergerac. A Computational Manipulation Perspective on Generative AI' in Mimi Zou, Cristina Poncibò, Martin Ebers and Ryan Calo (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge: Cambridge University Press 2025) 56.

information (rather than moderate and report the harmful content) relevant for the individual's decision-making process.[112]

4. The Conundrum of a Liability Regime in GenAI

The liability contours of deepfakes financial scams involve multiple actors at multilevel stages of GenAI: malicious deepfake content creators, software developers, and online social media platforms. Regulatory challenges around the liability degree for malicious scams have raised concerns on the suitable remedies and normative interventions.[113] Most disputes allocate the responsibility to consumers for reasonable care in their decision-making process following the principle of 'caveat emptor'.[114] 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 tools embedded in GenAI and large language models (LLMs)[115] that 'hallucinate' users with fabricated or false information about historical events, legal filings, news and research articles.[116] AI hallucinations produce plausible but incorrect

---

[112] Peter Henderson, 'Challenges for Foundation Model Liability and Regulatory Regimes. An Analysis of US Law' in Mimi Zou, Cristina Poncibò, Martin Ebers and Ryan Calo (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge: Cambridge University Press 2025) 126-127.

[113] Judit Bayer, 'Legal implications of using generative AI in the media' (2024) 33(3) *Information & Communications Technology Law* 310.

[114] In *Santander UK Plc v CCP Graduate School Ltd* [2025] EWHC 667 (KB), the court held that even a bank might reasonably foresee harm to innocent third parties once alerted to the fact that one of its accounts had been used by fraudsters; it had no relationship with such a third party as would give rise to the necessary quality of proximity. This decision reflects the authority of the Supreme Court in *Philipp v Barclays Bank UK plc* [2023] UKSC 25, where a bank customer had been the victim of an authorised push payment fraud and had been deceived into instructing the bank to make a payment 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 (i.e., it was not in control of the fraudsters)".

[115] 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, Luciano Floridi, Stefan Larsson, Mariarosaria Taddeo and Steven L. Winter, 'The Artificial in "Artificial Intelligence": How Imagination Shapes AI Regulation' (2026), p.30-31, https://philarchive.org/archive/NOVTAI-2v2.

[116] Hallucinated outputs in GenAI result from the training data, the tools' design focus 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/.

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

outputs, even with perfect data, due to fundamental statistical and computational limits which cause unintentional errors in the decision-making process.[117] The datasets used to train the models can be biased, inaccurate or intentionally disrupted by content creators who provide faulty narratives to manipulate information.[118] As argued, corrupted data and information make GenAI outputs no longer so intelligent, profitable, or socially beneficial for market participants.[119] However, GenAI hallucinations not only produces false outputs but are entangled in the cognitive experience (e.g., memories, emotions, behaviour) that characterises the human-AI interaction in conversational interfaces.[120] GenAI tools hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty: AI models do not acknowledge uncertainty, they provide estimates as facts without deeper research about the validation of their statements.[121] The impact of uncertainty about the performance of GenAI is a matter of concern for legal practice and consumer claims disputes, where the large use of AI models poses risks of misperception in accessing valuable advisory assistance.[122]

---

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 holds: (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 [at para 72].

[117] It has been reported that epistemic uncertainty when information appeared rarely in training data, model limitations where tasks exceeded current architectures' representational capacity, and computational intractability where even superintelligent systems could not solve cryptographically hard problems. See Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala and Edwin Zhang, 'Why Language Models Hallucinate' (2025), https://arxiv.org/pdf/2509.04664.

[118] Naja Bentzen, 'Information manipulation in the age of generative artificial intelligence' (December 2025), European Parliamentary Research Service PE 779.259, https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/779259/EPRS_BRI(2025)779259_EN.pdf.

[119] Tom C.W. Lin, 'Artificial Intelligence, Misinformation, and Market Misconduct' (2024) 85(4) *Ohio State Law Journal* 685, 702.

[120] Lucy Osler, 'Hallucinating with AI: Distributed Delusions and "AI Psychosis"' (2026) 39(1) *Philosophy & Technology*, https://doi.org/10.1007/s13347-026-01034-3.

[121] 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 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" [at para 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" [at para 6].

[122] 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(4) *University of New South Wales Law Journal*, 1277.

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.[123] 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.[124] In the doctrinal debate, it has been noted that additional concerns result from the lack of a universally accepted definition of hallucinations.[125] 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'[126] in legal citations and contentions to authoritative sources - would mitigate the users' reliance to hallucinated cases through verification at the point of citation.[127]

Situations of unintended harms caused by faulty outputs (e.g., 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 (e.g., human-generated misinformation in influencing decision-making). Human misinformation is technically and conceptually different from AI hallucinations in terms of motivations, beliefs, and intent.[128] This results in human-initiated inaccuracies which involve deliberate deception or flawed prompts, leading to disinformation as intentional falsehoods and computational manipulation of individuals' choice.[129] As a corollary, intentional misuse of GenAI may exacerbate the risk of deteriorating the information ecosystem with lower-quality, synthetic, or misleading content, effectively generating untruthful content.[130] Sophisticated forms of market manipulation come from misuse of AI techniques such as algorithmic trading which faces 'challenges to

[123] 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.
[124] Gyana Swain, 'OpenAI admits AI hallucinations are mathematically inevitable, not just engineering flaws' (18 September 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.
[125] Eliza Mik, 'Revisiting Legal Hallucinations' (10 March 2026), Centre for Legal Innovation and Digital Society (CLINDS)'s 30th LegalTech Seminar, Faculty of Law, The Chinese University of Hong Kong.
[126] An electronic link embedded in a citation that allows direct access to the cited authority would refrain the careless use of AI outputs.
[127] Oliver Roberts, 'Spread of AI Hallucinations Drives Need for Sanctions Reporting' (20 February 2026) Bloomberg Law, https://news.bloomberglaw.com/legal-exchange-insights-and-commentary/spread-of-ai-hallucinations-drives-need-for-sanctions-reporting.
[128] Anqi Shao, 'New sources of inaccuracy? A conceptual framework for studying AI hallucinations' (2025) 6(4) *Harvard Kennedy School Misinformation Review*, 3.
[129] Stefano Faraoni, 'AI-Enabled Manipulative Techniques: A Contract Law Perspective' (2026) *Lloyd's Maritime and Commercial Law Quarterly* 77.
[130] Joseph E. Stiglitz and Maxim Ventura-Bolet, 'The Impact of AI and Digital Platforms on the Information Ecosystem' (October 2025) NBER Working Paper Series No. 34318, p.3-4, http://www.nber.org/papers/w34318.

guaranteeing accountability and assigning responsibility, as human experts do not explicitly program the AI behaviour'.[131] The complexity and interoperability of automated trading practices as well as their autonomy in the decision-making activity (self-learning from own experience from the observation of markets) make difficult to identify the liability degree (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 deterrence regulatory framework to achieve effective law enforcement of securities law has been suggested in the doctrinal debate with a view to address unlawful behaviours and conduct in the trading context.[132]

Manipulated content in deepfake financial scams undermines technological advancements and exploits vulnerabilities of software design in facilitating intentional fraud, misinformation and emotional harm.[133] The Financial Stability Board (FSB) warned about the potential of GenAI to increase financial fraud and the ability of malicious actors to generate and spread disinformation in financial markets, which urges the need for accountability mechanisms and content moderation.[134] 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.[135] It is human fault to deceive and distort the reality through materials artificially created by the AI tool; this affects market structure and trading exchanges with negative implications on competition and transaction costs.[136] 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 are originated. Conduct of business rules can mitigate the risks of traders committing abuses, although the high-speed data used in financial

[131] Alessio Azzutti, 'AI trading and the limits of EU law enforcement in deterring market manipulation' (2022) 45 *Computer Law & Security Review* 4.
[132] 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 with 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.
[133] David Hirshleifer et al., 'AI, Opinion Ecosystems, and Finance' (February 2026), NBER Working Paper Series No. 34807, http://www.nber.org/papers/w34807.
[134] FSB, 'The Financial Stability Implications of Artificial Intelligence' (14 November 2024), p.2, https://www.fsb.org/uploads/P14112024.pdf
[135] Michael D. Murray, 'Visual Legal Rhetoric in the Age of Generative AI and Deepfakes: Renaissance or Dark Ages?' (2025) 28(1) *SMU Science and Technology Law Review* 199, 230-231.
[136] Chinmayi Sharma, 'AI's Hippocratic Oath' (2025) 102(4) *Washington University Law Review* 1101, 1159-1160, where it is proposed a new professional licensing process for artificial intelligence engineers.

products make monitoring and supervisory activities meaningless with respect to technological errors.[137]

Risk management and internal controls of financial firms are inadequate to detect potential harm of malicious deepfake scams; in parallel, regulatory compliance and prudential standards of financial authorities show weaknesses in skills and resources (organisational skills deficiencies in data science and essential IT capabilities) that are necessary to assess AI tools.[138] Model validation, ongoing monitoring, performing outcomes analysis, and assessing data quality require substantive compliance costs and understanding of the vulnerabilities associated with AI models. This implicates considering policy frameworks that enhance engagement with private sector participants, AI developers and other third-party service providers.[139]

4.1 The Shared Liability Model

We identify different scenarios for allocating responsibility to 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 a deterrent to refrain malicious content creators and online social media platforms from disseminating deepfakes. This approach originates from the conceptual framework of single liability shared with multiple injurers.[140] The shared responsibility applies to multiple stakeholders proportionally responsible for a particular action or outcome.[141] The share of liability must be assessed on a case-by-case basis depending on the duty carried by each party involved in the illegal activity.

Andhov and Gardner observed that complex algorithmic decision-making systems challenge the application of tort law as an appropriate private cause of action to ensure compensation for

---

[137] Yesha Yadav, 'Oversight Failure in Securities Markets' (2019) 104(7) *Cornell Law Review* 1799, 1855. It is pointed 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 exchanges or dark pool cannot" (at 1858). A system of shared liability fund for trading venues can reduce incentives for market participants to take risks that cannot cover in case of losses, and can facilitate industry self-monitoring and discipline.

[138] FSB, 'Enhancing Third-Party Risk Management and Oversight. A toolkit for financial institutions and financial authorities' (4 December 2023), p.33, https://www.fsb.org/uploads/P041223-1.pdf.

[139] OECD, 'Regulatory approaches to Artificial Intelligence in finance' (September 2024), OECD Artificial Intelligence Papers No. 24, p.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.

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

[141] 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(5) *AI & Society* 4035.

individual victims of financial harm.[142] 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, it is observed that risks of harm can be reasonably foreseeable, and the exact causal mechanism of the harm does not need to be reasonably foreseeable.[143] Howells and Twigg-Flesner proposed a liability approach that allows the user to seek redress without having to face the difficulties associated with identifying the correct party responsible.[144] Ben-Shahar offers a suggestive view on fault-based liability in tort law and apportionment problems by proposing a safety score liability which is commensurate with a party's habitual propensity to behave unsafely.[145] This proposal is based on measurable and predictable persons' actions identified by a machine learning algorithm score calculation to determine the probability of causing accidents with occurred harms.[146] 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 found limitations of liability law pointing out that 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'.[147]

We situate the shared liability regime on causation in tort law 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.[148] This regime allows the identification of the burden of responsibility for manipulating the GenAI system across the supply chain

---

[142] Alexandra Andhov and Jodi Gardner, 'Duties of Care for Algorithmic Decision-Making in the Financial Industry', paper presented at the NUS Law and Fintech Conference, Singapore, 31 July-1 August 2025.
[143] On this discussion see Law Commission, 'AI and the Law' (31 July 2025), Discussion Paper, p.11, https://lawcom.gov.uk/publication/artificial-intelligence-and-the-law-a-discussion-paper/.
[144] 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: Cambridge University Press 2022) 198.
[145] Omri Ben-Shahar, 'Safety Score Liability' (2025) 17(1) *Journal of Legal Analysis* 190..
[146] This regime proposes that parties are partially liable in proportion to their safety score, as it has been noted, 'the higher the safety score, the lower the fraction of liability'. Ibid., 196.
[147] Chris Reed and Keri Grieman, 'Responsibility for generative AI services – from aspiration to achievement' (2025), p.7-8, https://ssrn.com/abstract=5107667.
[148] 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(4) *U.C. Davis Law Review* 1141. See also Omri Ben-Shahar, 'Causation and Foreseeability' in Gerrit De Geest (ed), *Encyclopedia of Law and Economics*, vol. 1, *Tort Law and Economics* (2nd edn., Cheltenham: Edward Elgar 2009) ch.3.

while providing consumers effective remedies to claim compensation where fraudulent practices compromise their decision-making process. However, the proportion of responsibility as articulated in the shared liability regime may find limitations to protect reasonable expectations of safety and reliability in AI software from an average consumer.[149] Practical implementation of the shared liability is examined in the party's actions to heighten 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 of a shared liability regime is to first create a more resilient AI risk framework where all parties implement more robust verification technologies and protocols to protect consumers. Secondly, when victims are scammed, the shared liability regime would provide guidance to establish which party should be liable.

4.2 Liability for Deepfake Software Developers

The missing party being held liable in the deepfake scam supply chain is deepfake software developers, which 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 creating software engineering in which developers hold responsibility for using methodologies which may be manipulated for harmful deepfake content (e.g., audio, visual images, advice). Developers of automated tools such as agentic integrated development environments (large language models such as GitHub and Copilot X) can generate code programming to support human skills.[150] Software development methodologies require ongoing monitoring, adaptation and refinement to avoid hacking and data leakage which gives access to fraudster content creators.[151] Developers can provide safeguards to the software to refrain hackers and malicious users from accommodating requests of cyberattacks. Although the safeguards can be eluded by a downstream user, a few proactive techniques can be adopted to monitor users performing harmful tasks, such as open-sourcing, stress-test models pre-

[149] Geraint Howells, 'Protecting Consumer Protection Values in the Fourth Industrial Revolution' (2020) 43(1) *Journal of Consumer Policy* 160; Peter Cartwright, 'Understanding and Protecting Vulnerable Financial Consumers' (2015) 38(2) *Journal of Consumer Policy* 119, 127.
[150] Marijn Janssen, 'Responsible governance of generative AI: conceptualizing GenAI as complex adaptive systems' (2025) 44(1) *Policy and Society* 45-46.
[151] Tim Howard, 'GenAI and software development: a new paradigm' (2 June 2025), Blog Defra digital, data, technology and security, https://defradigital.blog.gov.uk/2025/06/02/genai-and-software-development-a-new-paradigm/.

release, structured model access for third-party auditors and research Application Programming Interfaces (APIs) to facilitate external independent model evaluation.[152] These defensive strategies can be implemented by creating a group of 'red-team professionals' to provide effective safety evaluations of AI models.[153]

The shared liability regime can be implemented for developers based on the tort of negligence for failure to ensure due care in ensuring the safety and reliability of the software (generated code and automation) as well as providing software updates. Developers should be responsible for the code review and debugging of software, although they often bargain with the online platforms, agreeing to distribute exclusivity through big tech companies in exchange for reduced fees.[154] This 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.[155]

An interesting case of accountability regime is found in medical law where software has been interpreted as a service rather than a product, on account of a fault liability for the injury that occurred from the integration of AI systems into healthcare practice, which is shared between physicians, hospitals and product developers.[156] The EU Product Liability Directive (PLD)[157] includes software as a product for the purposes of a no-fault liability, reframing it under the strict liability regime.[158] This impact is that producers of software can be held responsible for

[152] Elizabeth Seger et al., 'Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives' (2023) *Computers and Society*, p.17, https://doi.org/10.48550/arXiv.2311.09227.

[153] 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(2) *University of Memphis Law Review* 441, 451-452.

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

[155] Mar Negreiro, 'Scam calls in times of generative AI' (October 2025), European Parliamentary Research Service PE 777.940, https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/777940/EPRS_ATA(2025)777940_EN.pdf.

[156] Regulation (EU) 2017/745. 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* (Cheltenham: Edward Elgar 2024) 166.

[157] Directive (EU) 2024/2853.

[158] Article 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 the UK legislation, which dates back to the mid-1980s and was enacted as Part 1 of the Consumer Protection Act 1987, based on the EU's 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 of connecting to the internet or other networks. Further, a wide array of digital products has emerged in the form of applications and those building on AI systems; none of these is currently

damage caused by their defective products, regardless of fault.[159] In the academic debate, it is argued that software have become important elements of products and were even remotely interconnected with them; therefore, the case for considering software like any other component have become stronger with the result that the producer of the software itself could be liable in their own capacity whether as supplier of software as a component or as standalone software.[160] On this line of thinking, 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'.[161] 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.[162] 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.[163] As Howells and Twigg-Flessner observe, the term 'producer' is given a broad meaning under the PLD.[164] 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 therefore can be accountable to any victims' losses.[165]

---

within the scope of the existing product liability legislation in the UK. See https://lawcom.gov.uk/project/product-liability/.

[159] 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 purposes 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".

[160] 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: Cambridge University Press 2026) 160.

[161] 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., Cheltenham: Edward Elgar 2025) 591.

[162] 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., Cheltenham: Edward Elgar 2025) 482.

[163] Recital 50 of the PLD.

[164] Howells and Twigg-Flessner (n 144) 186-187.

[165] 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 into the market to fill just 212 customer orders. The Securities 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 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* (London, 17 October 2012), https://www.ft.com/content/928a1528-1859-11e2-80e9-00144feabdc0. US Securities and Exchange Commission, 'SEC Charges Knight Capital With Violations of Market Access Rule (U.S., 28 July 2014), https://www.sec.gov/newsroom/press-releases/2013-222.

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.[166] Responsibility for 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'.

Prevention measures to refrain malicious actors also include government certification programme and authorisation schemes for assessing the software before running onto the platforms.[167] The licence regime would test the used technology to ascertain that the software components (e.g., operating system, code and data)[168] are aligned with the requirements of safety and reliability. On this view, 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 as the guidance that is provided by more conventional symbolic AI'.[169] The rapid development of face-swapping AI technology questioned how to identify the scam and how to measure the deepfake.[170] The capacity for face-swapping to deceive is bolstered by voice-cloning applications; the voice changing device raises concerns for its use in various operations such as buying a product on the platform or creating voices and cloning them.[171] The significant challenge is that face-swapping can be used for commercial purposes but can also be used for malicious investment fraud, although the question is, how can the licencing system work when the AI software can be used for both purposes? 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 the identification of natural persons through fingerprint

[166] Article 13 of the PLD.
[167] Anuragini Shirish and Shobana Komal, 'A Socio-Legal Inquiry on Deepfakes' (2024) 54(2) *California Western International Law Journal* 555-556.
[168] 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 et al., *Component Software: Beyond Object-Oriented Programming* (2nd edn., Boston: Addison-Wesley Pearson Education 2002) ch.1.
[169] Daniel Schwarcz, Tom Baker and Kyle Logue, 'Regulating Robo-Advisors in an Age of Generative Artificial Intelligence' (2025) 82(1) *Washington & Lee Law Review* 775, 806.
[170] Yi Yan, 'Deep Dive into Deepfakes—Safeguarding Our Digital Identity' (2023) 48(2) *Brooklyn Journal of International Law* 768.
[171] Liam James, 'Face-swapping: Why you can't trust a video call – and how scammers are taking advantage' *Independent* (28 February 2025), https://www.independent.co.uk/tech/deepfake-scam-face-swap-fraud-ai-b2706722.html.

information.[172] It has been noted 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.[173]

The Australian Scams Prevention Framework Act 2025 requires digital platforms to check all advertisers of financial products have an Australian Financial Services Licence and take specific steps in the verification of new accounts.[174] 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:

'concrete 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'.[175]

Regulatory sandboxes such as digital 'scale up box' can be implemented to monitor the quality and performance of GenAI applications.[176] 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.[177] They aim to enhance cooperation between regulatory authorities and financial firms with a view to promoting best practices for transparency of the automated systems.

---

[172] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed and Aun Irtaza, 'Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward' (2023) 53(4) *Applied Intelligence*, 4013-4014.

[173] Ellen P. Goodman, 'Synthetic Content: Default to Distrust' (2025) Rutgers Law School Research Paper, 27, https://ssrn.com/abstract=5236368. 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'.

[174] Australian Government The Treasury, 'Scams Prevention Framework. Protecting Australians from scams' (3 February 2025), p.5, https://treasury.gov.au/publication/p2025-623966.

[175] 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 (16 November 2020), 13026/20, https://data.consilium.europa.eu/doc/document/ST-13026-2020-INIT/en/pdf.

[176] Walter G. Johnson, 'Caught in quicksand? Compliance and legitimacy challenges in using regulatory sandboxes to manage emerging technologies' (2023) 17(3) *Regulation & Governance* 709.

[177] See https://www.fca.org.uk/firms/innovation/regulatory-sandbox. The FCA launched a Supercharged Sandbox tool using NVIDIA accelerated computing and NVIDIA AI Enterprise Software to provide firms 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 allows firms to experiment with AI alongside NVIDIA' (9 June 2025), https://www.fca.org.uk/news/press-releases/fca-allows-firms-experiment-ai-alongside-nvidia; https://fcainnovation.co.uk/wp-content/uploads/2025/06/supercharged-sandbox-participation-pack.pdf.

An innovative regulatory approach against the risks of unexpected outcomes of technology applications is the FCA's consumer duty which concerns the firm's conduct in relation to financial services.[178] The consumer duty rule sets the standard of care that firms should give to customers in retail markets and holds responsible manufacturers and distributors for the assessment of foreseeable harms.[179] The duty applies across the distribution chain such as the design or operation of products or services and the firm's responsibilities depend on its role and ability to influence retail customer outcomes.[180] Manufacturers are required to monitor the characteristics of their products and whether are fit for purpose, and distributors are under obligation to have a clear understanding of the target market and the way products operate.[181] 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.[182] Firms need to share information about harmful content of 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.[183] The consumer duty can impose a main obligation for firms to report fraudulent behaviours in deepfake scams relating to GenAI products and implement internal controls to

[178] FCA, 'Final non-Handbook Guidance for firms on the Consumer Duty' (July 2022), FG22/5, p.35, https://www.fca.org.uk/publication/finalised-guidance/fg22-5.pdf. 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'.

[179] 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(2) *Cardozo International & Comparative Law Review* 465.

[180] 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* (London, 11 October 2023), https://www.ftadviser.com/regulation/2023/10/11/could-there-be-consumer-duty-failure-claims-within-your-distribution-chain/.

[181] 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.

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

[183] 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 activities and do not need to oversee the actions of other firms in the distribution chain although they may be responsible in some situations for the actions of other parties where they have material influence over customer outcomes. See https://www.fca.org.uk/publications/good-and-poor-practice/consumer-duty-implementation-good-practice-and-areas-improvement; https://www.fca.org.uk/firms/consumer-duty-information-firms.

detect potential harmful software.[184] The FCA duty rule can be implemented by firms to reduce vulnerability of GenAI tools although compliance requirements, such as risk management processes and transparency obligations, may not be effective in detecting (and avoiding) potential harmful outputs.[185] Further, the consumer duty rule requires firms to conduct an assessment of the AI model and its use and operation before deploying outcomes to the market.[186]

4.3 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 to capture on time the unintended malicious actions of users. It is expected that 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. Watermarking techniques such as hashing are forgery methods to check the integrity of content by tracing files with a short string of numbers that is lost if the video or audio is deepfaked.[187] Failure to employ digital forensic techniques in GenAI amounts to gross negligence, which results in personal liability for individuals accountable for careless supervision of content authenticity.[188] Digital forensic techniques employ blockchain technology which can verify the

[184] FCA, 'Immediate areas for action and further plans for reviewing FCA requirements following introduction of the Consumer Duty' (March 2025), Feedback Statement FS25/2, https://www.fca.org.uk/publication/feedback/fs25-2.pdf.
[185] Teresa Rodríguez de Las Heras Ballell, 'Mapping Generative AI Liability Cases in the EU Legal Framework' in Mimi Zou, Cristina Poncibò, Martin Ebers and Ryan Calo (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge: Cambridge University Press 2025) 106.
[186] FCA, 'FCA helps firms to test AI safely' (December 2025), https://www.fca.org.uk/news/press-releases/fca-helps-firms-test-ai-safely; 'Our Consumer Duty focus areas' (December 2025), https://www.fca.org.uk/publications/corporate-documents/consumer-duty-focus-areas; 'Consumer Duty implementation: good practice and areas for improvement' (December 2025), https://www.fca.org.uk/publications/good-and-poor-practice/consumer-duty-implementation-good-practice-and-areas-improvement.
[187] Liam Kearns, Abu Alam, Jordan Allison, 'Synthetic Media Authentication Threats: Detection using a Combination of Neural Network and Blockchain Technology' (2025) 36(8) *Transactions on Emerging Telecommunications Technologies* e70225.
[188] David Atkinson and Jacob Morrison, 'A Legal Risk Taxonomy for Generative Artificial Intelligence' (2024) *Computers and Society*, https://arxiv.org/pdf/2404.09479.

origins and distribution of videos by storing digital signatures in a ledger which is difficult to manipulate.[189]

The UK Cyber Security and Resilience Bill 2024,[190] which updated the Network and Information Systems (NIS) Regulations 2018,[191] established a set of provisions that impose obligations on managed service providers (IT service providers and managed security services),[192] holding responsible those firms for failure to take effective technical and organisational measures (e.g., contractual requirements, security checks, or continuity plans) and to manage risks posed to the security of the network, infrastructure, and information systems.[193] The concept is similar to our proposal of imposing liability on those responsible for negligence. In China, Provisions on the Administration of Deep Synthesis of Internet-based Information Service 2023 introduced specific obligations and responsibilities on service providers, users and online platforms for the management of synthetic media technologies.[194] Alongside these legislative frameworks, supervisory authorities should develop response programmes tailored to address threats posed by deepfake manipulation. These programmes would provide procedures for identifying, assessing, and mitigating the harmful consequences of malicious deepfake content with regulatory tools that expedite coordination between enforcement mechanisms, intelligence agencies and 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. When AI-generated content is detected but is not violating a policy (i.e.,

---

[189] Mika Westerlund, 'The Emergence of Deepfake Technology: A Review' (2019) 9(11) *Technology Innovation Management Review* 46.

[190] See https://www.gov.uk/government/collections/cyber-security-and-resilience-bill.

[191] See https://www.legislation.gov.uk/uksi/2018/506.

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

[193] Jonathon Ellison, 'Cyber Security and Resilience Policy Statement to strengthen regulation of critical sectors' (1 April 2025), https://www.ncsc.gov.uk/pdfs/blog-post/cyber-security-resilience-bill-policy-statement.pdf.

[194] See https://www.pkulaw.com/en_law/90cff392df74a3ebbdfb.html. For commentary see Yinuo Geng, 'Comparing "Deepfake" Regulatory Regimes in the United States, the European Union, and China' (2023) 7(1) *Georgetown Law Technology Review* 157, 169-170.

a parody), platforms should apply a prominent, immutable label (e.g., 'AI-Generated' or 'Synthetic Media'), which must be visible even when the content is shared or downloaded. Further, if a user receives a message with 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.' If a user shares a video that has been detected as a potential deepfake scam, a warning could appear: 'This media has been identified as potentially synthetic. Learn more about deepfakes.'

Liability could also be imposed on software developers through implementing reasonable safeguards. For example, Microsoft launched safety benchmarks for developers to rank iterations from a range of providers in view to build trust with cloud customers.[195] 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.[196] It is an innovative mechanism to monitor GenAI products by ensuring they cannot be used to create harmful content. The safety system evaluates the level of fairness, biases and mistakes of AI products, and the risk of fraudulent hallucinated outputs in investment scams.[197]

The Singapore Government launched testing pilot schemes namely the Global AI Assurance Pilot,[198] the Joint Testing Report with Japan,[199] and the AI Safety Red Teaming Challenge Evaluation Report[200] to enhance the best practices of GenAI governance. Red teaming requires programmes to be tested in an adversarial manner and scenario planning, with the aim of detecting malicious and manipulative actions of software developers.[201] The Supreme Court of Singapore attributed liability for AI-generated content to users which complements the Shared Responsibility Framework 2024 and the Protection from Scams Act 2025.[202] In parallel, the

[195] Rafe Uddin and Cristina Criddle, 'Microsoft to rank 'safety' of AI models sold to cloud customers' *Financial Times* (London, 7 June 2025), https://www.ft.com/content/02f39b33-fa6e-4bb7-b1f4-8171b50738af.
[196] Liang Yu, Emil Alégroth, Panagiota Chatzipetrou and Tony Gorschek, 'Measuring the quality of generative AI systems: Mapping metrics to quality characteristics — Snowballing literature review' (2025) 186 *Information and Software Technology*, https://doi.org/10.1016/j.infsof.2025.107802.
[197] David Krause, 'Mitigating Risks for Financial Firms Using Generative AI Tools' (2023), https://ssrn.com/abstract=4452600.
[198] See https://aiverifyfoundation.sg/ai-assurance-pilot/.
[199] See 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.
[200]https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/singapore-ai-safety-red-teaming-challenge-evaluation-report.pdf.
[201] 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(4) *Loyola of Los Angeles Law Review* 859, 890-891.
[202] See https://sso.agc.gov.sg/Acts-Supp/1-2025/Published/20250217?DocDate=20250217.

English courts 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".[203]

Under the OSA 2023, senior managers of online social media platforms are criminally liable if they fail to implement proportionate systems and processes to prevent the publication and hosting of fraudulent advertising. Within the shared liability model, we posit that where the scale of the harm justifies it, such as the Silk Road operation, then content creators should be held liable for deepfake scams. We acknowledge the challenges associated with uncovering the anonymity of content creators, which is the reason we have focused on the liability of software developers and online social media platforms. We contend 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 authentication methods to prevent harmful misuse such as deepfake scams.

We suggest that online social media platforms should adopt protocol design and governance policy guidelines to prevent harmful risks of AI content while, in parallel, assessing the user expectations about whether the purpose of content is to mislead information and distort beliefs. Corporate governance boards (senior management, directors, executive and compliance officers) of software companies and online platforms should be trained and fully informed about the reputational risk that the use of defective GenAI models can determine for the business operations.[204] Poor monitoring of the expected outputs and limited understanding of the AI decision-making process lead to compliance issues and faulty corporate reporting, which affect the expectations of investors and market stakeholders.

5. Conclusion

Advances in generative computing software provide firms with greater opportunities to create smarter services for customers, but they also give more powerful tools to scammers, fraudsters and deepfakes.[205] GenAI content can disseminate poor information and unintended

[203] *R. (on the application of Ayinde) v Haringey LBC* [2025] EWHC 1383 (Admin) [at para 5].
[204] 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, Cristina Poncibò, Martin Ebers and Ryan Calo (eds), *The Cambridge Handbook of Generative AI and the Law* (Cambridge: Cambridge University Press 2025) 369 and 382.
[205] Stephen Bush, 'Are we human or are we spammer?' *Financial Times* (London, 24 June 2025), https://www.ft.com/content/07ed552a-e681-472a-8eba-c4b3a7938027.

'hallucinations' which are not factually accurate, and it is difficult to detect how it reached a particular decision.[206] Malicious deepfakes operate at different stages of the AI supply chain through digital platforms, which makes meaningless the identification of frauds.

The time is ripe for regulatory intervention. We proposed a shared liability regime to attribute responsibility to actors involved in developing and elaborating fraudulent AI-generated content. This regime would enable a more balanced approach to compensation where the parties involved will bear losses proportionally, depending on whether and to what extent participants contribute to the scam. The FCA's consumer duty rule can complement the shared liability regime in requiring firms to avoid foreseeable harm to customers in the AI model's behaviour. However, the UK regulatory framework does not offer adequate responses to the risks of automated manipulation of information in synthetic media, as the impersonation of scammers exacerbates unfair practices against users. This suggests that the UK legislation should consider specific liability approaches where the timing and nature of the deepfake scam deceive the safety and reliability of outcomes while, at the same time, making online social media platforms responsible for compensation to victims. The normative experience of Singapore in tackling online scams and sharing the costs of frauds would provide a policy guideline for the UK Government to establish mandatory rules for compensating losses and allocating liability among multiple parties of malicious deepfakes.

[206] Cristina Criddle, 'How to get the best out of AI' *Financial Times* (London, 10 June 2025), https://www.ft.com/content/d2f1fa02-025c-41ff-814f-00f22ed5c6d3.