THE UNIVERSITY OF MELBOURNE

Centre for Artificial Intelligence and Digital Ethics

CAIDE AI Policy Forums

FORUM #5 ISSUES PAPER: Demystifying Generative AI, Synthetic Content and Fraud

About the Author

Andrew Lim is a Research Associate with the Centre for Artificial Intelligence and Digital Ethics (CAIDE). He has just completed his Bachelor of Science in Physics and Diploma of Languages in Latin at the University of Melbourne. With prior experience across student advocacy and journalism, including guest lecturing on generative AI in a 'post-truth' age to young journalists at Boston College in the USA, Andrew is deeply passionate about the growing nexus between emerging technologies, public communications, and good governance in a globalised world: never more apparent than in questions of institutional trust, ground truth, and effective AI policy. Andrew would like to acknowledge the contributions of Professor Jeannie Marie Paterson to the editing of this paper.

What is CAIDE?

The Centre for Artificial Intelligence and Digital Ethics (CAIDE) is a cross-disciplinary research centre at the University of Melbourne. CAIDE facilitates cross-disciplinary research, teaching and leadership on the ethical, technical, regulatory and legal issues relating to AI and digital technologies. CAIDE is directed by Professor Jeannie Marie Paterson from Melbourne Law School. For more information about CAIDE, see our website: https://www.unimelb.edu.au/caide.

Acknowledgements and Sponsorships

The AI Policy forums are supported by:

- The Centre for AI and Digital Ethics, funded by the Faculty of Engineering and Information Technology and Melbourne Law School at the University of Melbourne
- The Ninian Stephen Law Program powered by the Menzies Foundation, as part of the project *New Legal Thinking for Emerging Technologies*
- Microsoft, Atlassian and the Tech Council of Australia as part of the project *Demystifying Generative AI*.

CAIDE AI Policy Forums

This issues paper was prepared as reading for the CAIDE AI Policy Forum #5, held at the University of Melbourne in 2024. It should be read in conjunction with the discussion paper on GenAI, deepfakes and fraud, available here. The aim of the AI Policy Forums is to provide an opportunity for discussing policy and law issues raised by the emergence of AI that go beyond the headlines. This event considered the impact of AI, particularly Generative AI and deepfake technology, on trust, truth and fraud.

Contact CAIDE

Email: uom-caide@unimelb.edu.au.
Website: https://www.unimelb.edu.au/caide.
LinkedIn: https://au.linkedin.com/company/caide-unimelb.

Demystifying Generative AI, Synthetic Content and Fraud

Background

The rise of tools able to quickly and easily generate believable synthetic content have raised serious concerns about their potential for societal and individual harm. Their use has prompted some to call for regulatory action to ensure responsible usage of this technology.

The Tech

The term 'deepfake' has often been used as a catch-all for these types of synthetic content. A deepfake is strictly "a digital photo, video or sound file of a real person that has been edited to create an extremely realistic but false depiction of them doing or saying something they did not actually do or say",[1] though the concept has been extended in popular media to describe entirely synthetic content generated by AI.[2]

Deepfakes can have entirely legitimate usages, especially reducing cost in marketing sectors – companies like Zalando, Cadbury (Mondelez) and Megafon have already used deepfakes of brand ambassadors and influencers to allow localised advertising without the cost of hiring the individuals to shoot hours of footage in different locations.[3] However, it may also be used for intimate image abuse, political misinformation and extortion. In this paper, we focus primarily on its use for scams and other financial fraud.

Financial Harms

Financial scams using generative AI have emerged at a variety of scales. In February 2024, multinational consulting engineering firm Arup lost HK$200 million after an employee was invited onto a video conference call with scammers posing as senior company officers using generative-AI-enabled deepfake technology.[4] Advertising group WPP was targeted in a similar scam using a voice clone of the group's CEO, though this was ultimately unsuccessful.[5] On a smaller scale, push-payment imposter scams where the scammer poses as a child or grandchild of the victim and requests emergency money have been enhanced by the ability to generate convincing deepfake audio.[6] So too have widespread cryptocurrency scams using deepfake-generated celebrity endorsements spread online.[7]

More recently, pull payment scams involving scammers posing as customers have become a major concern for banks reliant on voice identification security (especially after, in 2023, a journalist accessed his Lloyds Bank account using free, publicly available technology from tech company ElevenLabs[8]) though only two major Australian banks, Bank Australia[9] and ANZ[10], currently use the technology. Overseas banks, including HSBC, Citibank, Lloyds, Wells Fargo, TD and Chase Bank, are all currently more vulnerable to these types of scams due to their widespread adoption of voice verification.[11]

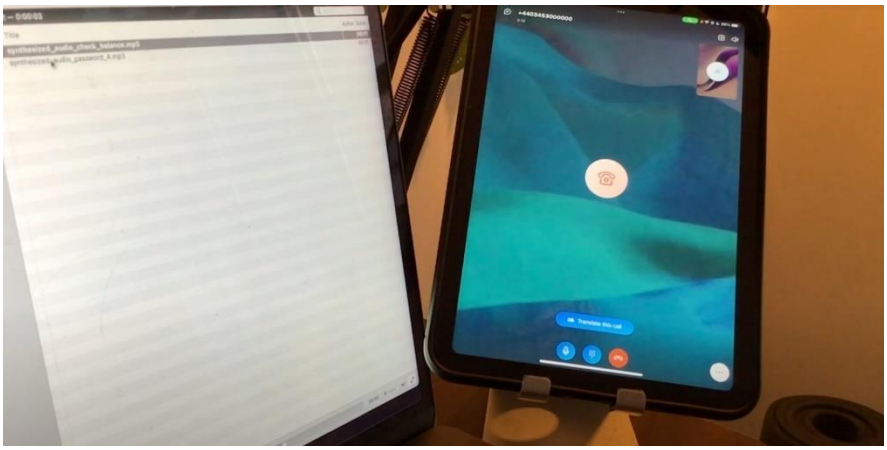

Figure 1: Journalist Joseph Cox accesses his voice-verified bank account using an AI-generated voice made with publicly available audio 'deepfake' technology. VICE

Proposals for Law Reform

With these potential harms, so too have emerged legislative, regulatory and technological proposals for solutions. Legal approaches have both taken the form of both catch-all deepfake legislation[12] and specific measures designed to deal with different uses of deepfakes for financial fraud. They have also ranged from outright bans and disclaimer requirements to more technologically focussed solutions like digital fingerprinting and watermark requirements. Regulators like the FCC in the United States have also utilised existing powers to crack down on actions where possible (especially with telephone consumer law provisions against AI deepfake robocalls).[13] And a plethora of technology-based

approaches purporting to detect both video and audio deepfakes have also purported to provide an effective defence against deepfakes financial fraud for both consumers and banks.[14]

The truth is that no one piece of legislation or technological advancement will, however developed, provide an easy solution. Rather, each one presents its own unique set of trade-offs between competing desires – for easy-to-navigate regulatory frameworks; for reliable and safe financial institutions; and much more. Any integrated policy solution must consider both normative legal concerns and practical technical realities, reflecting the breadth of all these issues to craft an integrated policy approach.

The Policy Issues

1: Technological Solutions – is this really a legal issue?

Technological solutions have often been the first port of call for banks combatting financial scams generated by AI. Deepfake technology has been of particular concern, with the Australian Federal Police highlighting it as one of the two most common scam tactics of the 2023-24 financial year,[15] though more general applications for AI, in particular allowing the efficient generation at scale of spear phishing content (where a scam is personalised to publicly available information about the victim) have been raised too.[16] Many banks have therefore turned to technology-based solutions as they attempt to combat both push-payment (where a consumer is tricked into paying money into a fraudster's account) and, increasingly, pull-payment (where a bank is tricked by a fraudster into believing they are another customer) scams.[17]

In Australia, the Commonwealth Bank has utilised generative AI technology to keep up with the ease of scam technology, automating the process of writing code to protect against individual scam types. Likewise, at ANZ, more traditional machine learning algorithms have wholly automated about 35% of incident response processes.[18] Yet, if somewhat effective, these methods run the risk of simply augmenting existing scam prevention measures, rather than dealing directly with the threat posed by deepfakes. More tailored processes do exist, especially in the form of providers of technology purporting to both prevent and detect deepfake scams in progress – though these are specifically aimed at pull payment scams: allowing banks to see through deepfake-generated customer voices.[19] More aimed at general consumer use are tools like McAfee's new Deepfake Detector, which purports to alert users when videos they are watching are likely AI-generated.[20]

While not present in Australia, other jurisdictions have proposed approaches that would mandate digital fingerprinting and watermarks, essentially preventing scammers from using most publicly available tools and in theory reducing the ease with which scammers can generate deepfakes en masse.[21] It is worth noting that not all financial deepfake fraud measures are technologically-specific: in particular, calls to adopt measures (similar to those in the UK[22]) where banks refund money lost in push payment fraud,[23] while particularly pertinent in the context of deepfake-aided fraud, are not reliant on any technological developments or lack thereof.

The Treasury Laws Amendment (Scams Prevention Framework) Bill 2024 (Cth) attempts to address both concerns. It provides for sector-specific mandatory codes of practice for multiple sectors involved in the scam process (to begin, banks, telecommunication providers, digital communications platforms, search engine adverts and direct messaging platforms) and associated penalties. This would in theory allow regulators to mandate technological measures (e.g. fingerprinting and watermarks) as necessary on a sector-by-sector basis. It also provides for a single external dispute resolution consumer scheme run by the Australian Financial Complaints Authority (AFCA), allowing consumers to easily seek compensation from banks, telcos and social media companies, addressing some of the push-payment-specific concerns. Some concerns remain, however, especially regarding the ease of accessibility to AFCA from ordinary consumers, and the cost of implementing mandatory minimums for smaller banks.

2: Utilising Existing Legislation – how much new statute law is really needed?

All this said, there may also be a place for existing legislation, especially in Australia, to provide the protections necessary to deal with deepfake-related issues, especially concerning financial fraud. In particular, the Australian Consumer Law provides an ideal tool for consumer remedies in scams like the ever-present celebrity cryptocurrency scam[24], which would potentially fall under both the s 29(1)(g) prohibition on sponsorship misrepresentations and the general s 18 misleading and deceptive conduct prohibition.[25]

A statute-light approach has been seen by some as not only inevitable but desirable, placing AI in a lineage of other technological developments of the past few centuries that did not require technology-specific legislation for effective regulation. As a Productivity Commission report in January 2024 put it:

> "Many potential harms have been encountered with past technologies and adequately dealt with by existing regulatory frameworks in areas such as consumer protection, privacy, anti-discrimination, negligence and sector-specific and profession-specific requirements. AI is no different"[26]

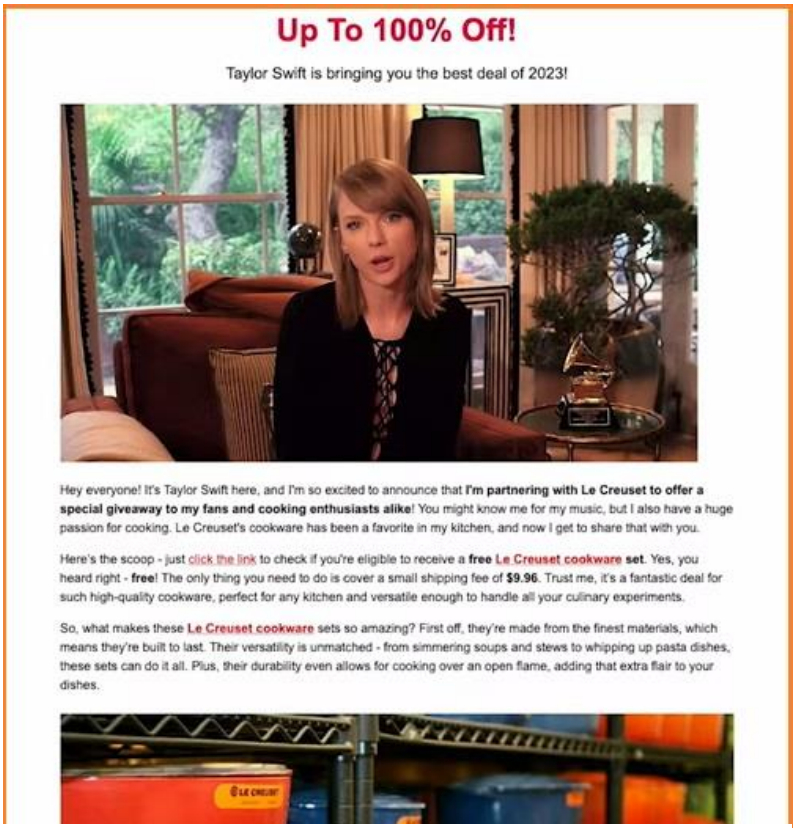

Figure 2: Viral celebrity deepfake cryptocurrency scams, like the one pictured here, are likely able to be dealt with under the Australian Consumer Law's existing provisions. Australian Computer Society

This approach was most clearly seen in the UK under the Sunak Conservative Government. Over the past couple of years, it pursued a policy of a "non-statutory basis" when it comes to AI regulation with the express aim of boosting technological innovation, noting that "many regulators… [can already] implement [AI] principles within their remits" and citing the Competition and Markets Authority, Advertising Standards Authority and Office of Communications as particular examples.[27] This approach was not without controversy – in particular, the consultation process on this policy raised concerns that "principles would be ineffective without statutory backing" and that many AI related issues bordered on fundamental "systemic risks…to democracy and the rule of law".[28] It is telling that this did not preclude the UK's action on push payment scams, but rather ensured it was done with a focus on bank obligations rather than technological requirements.

Indeed, this non-binding approach is likely to change in coming months – in July, the King's Speech of the incoming Starmer Labour Government suggested a new set of binding statutory measures will be introduced in the current parliament, to "place requirements on those working to develop the most powerful artificial intelligence models".[29] This does raise significant questions about how long non-statutory measures can last and how broad they can be.

3: Beyond Scams – an integrated 'risk-based approach'

In Australia, the approach to AI regulation has been generally quite broad: while in some cases, individual regulatory bodies have moved swiftly to regulate (including in circumstances involving, for instance, deepfake generation tools)[30], this has not necessarily been the case more widely. Australia's current approach to AI regulation is presented in the Department of Industry, Science and Resources' September 2024 *Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings*.[31] They outline a risk-based approach that differentiates low-, medium- and high-risk use cases, emphasising in high-risk scenarios the implementation of *ex ante* measures to prevent misuse rather than create liability (and induce expensive litigation) through a series of mandatory guardrails.[32]

Of special note to the deepfake/synthetic content discussion is Guardrail 6, the duty to "inform end users regarding AI-enabled decisions, interactions with AI and AI-generated content", which notably includes the requirement that "Organisations must apply best efforts to ensure AI-generated outputs, including synthetic text, image, audio or video content, can be detected as artificially generated or manipulated."[33] Some early proposals that could fit this requirement include the C2PA, a technical standard developed by several leading media and technology organisation, to establish provenance (if not veracity) from trusted institutions: for example showing a photo came from a government agency and was published by the news outlet without modification (if not declaring its inherent truthfulness).

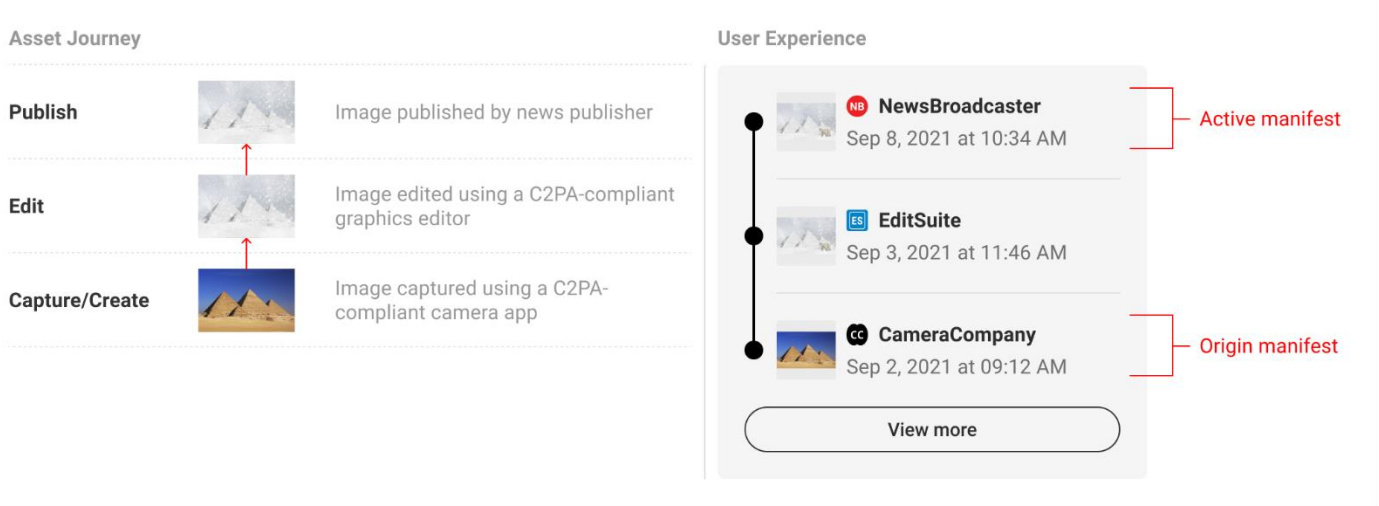

Figure 3: The C2PA technical standard is one technical solution to transparency issues – giving users a timeline of modifications to an image/video/audio attributed to trusted entities and programs, confirming its provenance (though not its truthfulness). C2PA

Notably, the *Proposals Paper* does not dictate who bears responsibility for regulating these guardrails. Instead, it suggests three pathways: one a legislation-light approach similar to that in Section 2; the other two differing on whether to establish a specialised AI regulator or simply to delegate new powers to existing regulatory agencies and frameworks.

We have already seen attempts towards the former overseas, with Canada's draft Artificial Intelligence and Data Bill[34] establishing an AI and data regulator to monitor and enforce compliance with its requirements. By contrast, under the latter, individual regulatory agencies would be empowered by amendments to their existing regulatory frameworks referring to a broad piece of framework legislation. AFCA and ACCC would retain their existing remits – but such decisions would be cross-referenced to a national risk-based approach to AI in general with commonly understood and communicated categories.

What is common to all proposals is a need for an integrated approach to AI governance. This is especially fraught around synthetic content. Its different risks are not isolated, either – financial impacts around fraud or misconduct are not isolated from societal ones around misinformation. For example, in 2023, an AI-generated image showed an explosion near the Pentagon, causing climbing Treasury bond and gold prices, and dips in the US stock market, until the image was shown to be false.[35] More on remedies to these societal harms is provided in *Demystifying Generative AI, Synthetic Content and the Truth*, another in this CAIDE Policy Paper series.

In the meantime, piecemeal legislation addressing individual synthetic content concerns will continue to emerge. These include not just the Scam Prevention Framework, but also other attempts to mitigate online harms, whether by dealing with misinformation and disinformation[36], placing age limits on social media[37] or reforming the *Online Safety Act* 2021 (Cth).[38] As these measures multiply and grow in scope, they will overlap,[39] growing this regulatory Gordian knot until a clearer framework is agreed.

Figure 1: Table of Proposed Australian Measures on GenAI and Scams (Updated 29 November 2024)

Measure	Status	Key Measures
Treasury Laws Amendment (Scams Prevention Framework) Bill 2024 (Cth)	Before House	Creates new ACCC-enforced penalties for non-compliance under *Competition and Consumer Act* 2010 (Cth) Enables designation of sector-specific mandatory codes of practice for sectors in scam ecosystem (including but not limited to banks, telecom, social media, search engine advertising and direct messaging) Mandates accessible internal dispute resolution for sectors and empowers AFCA to provide a single comprehensive external dispute resolution scheme for consumers.
Online Safety Act 2021 (Cth) Basic Online Safety Expectations (BOSE)	BOSE amended in June 2024	Requires reports on child/adult user breakdown, includes 'best interests of the child' test, and emphasises user safety criteria for generative AI
Online Safety Amendment (Social Media Minimum Age) Bill 2024 (Cth)	Passed; Awaiting Royal Assent	Sets a minimum age of 16 years for access to social media platforms, and introduces an obligation on providers to take 'reasonable steps' to use age assurance to prevent underage users from accessing social media, with failure to do so resulting in fines up to $49.5 million. Excludes messaging apps, education/health services and online gaming services from the definition of social media platforms.
Digital Platform Levy	Recommended by Select Committee	Levy exacted on digital platforms to extract funding for public interest journalism aimed at strengthening institutional protections against mis/dis-information.
Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth)	Withdrawn due to Senate opposition Further post-election attempts possible	Requires providers to report on misinformation risk on platforms. Empowers ACMA to create misinformation and disinformation standards, codes and dispute resolution procedures, with civil penalties up to 5% of annual turnover for social media companies. Includes explicit carve-outs for mainstream media/news, academic, artistic, satirical, humorous, scientific and religious speech

[1] eSafety Commissioner, 'Deepfake trends and challenges – position statement' (19 August 2024) <https://www.esafety.gov.au/industry/tech-trends-and-challenges/deepfakes>.
[2] Hannah Smith and Katherine Manstead, *Weaponised Deep Fakes – National Security and Democracy* (ASPI Policy Brief, 29 April 2020) <https://www.aspi.org.au/report/weaponised-deep-fakes>.
[3] Herbert Smith Freehills, 'Deepfakes in advertising – who's behind the camera?', *Digital TMT and Sourcing Notes* (28 February 2024) <https://www.herbertsmithfreehills.com/notes/tmt/2024-02/deepfakes-in-advertising-whos-behind-the-camera>.
[4] Dan Milmo, 'UK engineering firm Arup falls victim to £20m deepfake scam', *The Guardian* (17 May 2024) <https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video>.
[5] Nick Robins-Early, 'CEO of world's biggest ad firm targeted by deepfake scam', *The Guardian* (10 May 2024) <https://www.theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam>.
[6] Federal Communications Commission (US), 'Deep-Fake Audio and Video Links Make Robocalls and Scam Texts Harder to Spot' (Consumer Guide, 8 June 2024) <https://www.fcc.gov/consumers/guides/deep-fake-audio-and-video-links-make-robocalls-and-scam-texts-harder-spot>
[7] Stuart A Thompson, 'How 'Deepfake Elon Musk' Became the Internet's Biggest Scammer', *The New York Times* (14 August 2024). <https://www.nytimes.com/interactive/2024/08/14/technology/elon-musk-ai-deepfake-scam.html>
[8] Joseph Cox, 'How I Broke Into a Bank Account With an AI-Generated Voice', *VICE* (23 February 2024) <https://www.vice.com/en/article/how-i-broke-into-a-bank-account-with-an-ai-generated-voice/>
[9] 'VoiceID', *Bank Australia* <https://www.bankaust.com.au/support/voiceid>.
[10] 'ANZ Voice ID | ANZ', *ANZ Bank* <https://www.anz.com.au/ways-to-bank/mobile-banking-apps/voice-id/>.
[11] Elizabeth Cartier et al, 'Deep Fakes and Misinformation in the Finance Sector - Strategies to Prevent and Deter' (Consultancy Report, Columbia University School of International and Public Affairs, Spring 2023). <https://www.sipa.columbia.edu/deep-fakes-and-misinformation-finance-sector-strategies-prevent-and-deter>
[12] See, e.g., in the United States, the Nurture Originals, Foster Art, and Keep Entertainment Safe (No FAKES) Act, S 4875, 118th Congress (2024) and the No Artificial Intelligence Fake Replicas And Unauthorized Duplications (No AI FRAUD) Act, HR 6943, 118th Congress (2024).
[13] See Federal Communications Commission (US), 'FCC Makes AI-Generated Voices in Robocalls Illegal' (News Release, 8 February 2024) <https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal> and Federal Communications Commission (US), 'FCC-State Robocall Investigation Partnerships' (11 March 2024) <https://www.fcc.gov/fcc-state-robocall-investigation-partnerships>.

[14] See Cartier et al (n 19) and Lenovo Storyhub, 'McAfee Launches World's First Automatic and AI-powered Deepfake Detector Exclusively on Select New Lenovo AI PCs' (Press Release, 21 August 2024) <https://news.lenovo.com/pressroom/press-releases/mcafee-first-automatic-ai-powered-deepfake-detector/>.
[15] Australian Federal Police, 'AFP alerts Australians to common cryptocurrency investment scam tactics for Scams Awareness Week' (Media Release, 28 August 2024) <https://www.afp.gov.au/news-centre/media-release/afp-alerts-australians-common-cryptocurrency-investment-scam-tactics>
[16] 'Annual Cyber Threat Report 2023-2024' (Australian Signals Directorate, 20 November 2024) 33-34. <https://www.cyber.gov.au/about-us/view-all-content/reports-and-statistics/annual-cyber-threat-report-2023-2024>
[17] See, e.g., Cartier et al (n 22).
[18] Christopher Niesche, 'Banks turn to Gen AI to protect customers from scams', *Australian Financial Review* (22 July 2024) <https://www-afr-com.eu1.proxy.openathens.net/technology/banks-turn-to-gen-ai-to-protect-customers-from-scams-20240711-p5jsxf>.
[19] Cartier et al (n 22) 14-17.
[20] McAfee, 'McAfee® Deepfake Detector flags AI-generated audio within seconds', *McAfee* <https://www.mcafee.com/ai/deepfake-detector/>
[21] See, from the US, the Protecting Consumers from Deceptive AI Act, HR 7766, 118th Congress (2024).
[22] Payment Systems Regulator (UK), 'PSR continues to take bold action on APP fraud as it publishes final reimbursement details ahead of 2024 implementation' (Media Release, 19 December 2023) <https://www.psr.org.uk/news-and-updates/latest-news/news/psr-continues-to-take-bold-action-on-app-fraud-as-it-publishes-final-reimbursement-details-ahead-of-2024-implementation/>.
[23] Cait Kelly, 'Australian banks should reimburse scam victims, ACCC and consumer advocates say', *The Guardian* (2 February 2023) <https://www.theguardian.com/money/2023/feb/01/australian-banks-should-reimburse-scam-victims-accc-and-consumer-advocates-say>.
[24] See Thompson (n 15).
[25] Ted Tallas and Maggie Kearney, 'Diving into the Deep End: Regulating Deepfakes Online' (2019) 38(3) *Communications Law Bulletin* 11.
[26] 'Making the most of the AI opportunity: The challenges of regulating AI' (Research Paper No 2, Productivity Commission (Cth), 2024) 1. <https://www.pc.gov.au/research/completed/making-the-most-of-the-ai-opportunity/ai-paper2-regulating.pdf>
[27] 'Consultation Outcome – A pro-innovation approach to AI regulation: government response' (Command Paper No CP1019, Department for Science, Innovation & Technology (UK), 6 February 2024) [16] <https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response>.
[28] Ibid [19]-[21].
[29] United Kingdom, *Parliamentary Debates*, House of Lords, 17 July 2024, vol 839, col 7.
[30] See, e.g., the eSafety Commissioner's introduction of reporting requirements for high-risk generative AI services in *Online Safety (Designated Internet Services—Class 1A and Class 1B Material) Industry Standard 2024* (Cth), due to come into effect Dec 2024.
[31] Department of Industry, Science and Resources (Cth), *Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings* (September 2024). <https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf>
[32] Ibid 16-17.
[33] Ibid 39.
[34] Innovation, Science and Economic Development Canada, 'Artificial Intelligence and Data Act' (27 September 2023) <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act>
[35] Philip Marcelo, 'FACT FOCUS: Fake image of Pentagon explosion briefly sends jitters through stock market,' *AP News* (24 May 2023) <https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4>.
[36] Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth).
[37] Prime Minister of Australia, 'Albanese Government set to introduce minimum age for social media access' (Media Release, 10 September 2024) <https://www.pm.gov.au/media/albanese-government-set-introduce-minimum-age-social-media-access>.
[38] Department of Infrastructure, Transport, Regional Development, Communications and the Arts, 'Government to introduce legislation to combat seriously harmful misinformation and disinformation' (Media Release, 12 September 2024) <https://minister.infrastructure.gov.au/rowland/media-release/online-safety-expectations-boost-transparency-and-accountability-digital-platforms>.
[39] Already, early work at dealing with "unlawful and harmful material" is under the purview of the eSafety Commissioner, while proposed "serious harm" disinformation categories will come under ACMA: ibid, Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth).

THE UNIVERSITY OF
MELBOURNE

Centre for AI and Digital Ethics

Level 8
Melbourne Connect
700 Swanston Street
Carlton 3053

unimelb.edu.au/caide