13 June 2024

Associate Professor Andrew Meares
Deputy Director, School of Cybernetics
ANU College of Engineering, Computing & Cybernetics
The Australian National University
[redacted]

Senator Tony Sheldon
Chair of the Senate Select Committee on adopting artificial intelligence (AI)
Australian Senate Parliament House
Canberra ACT 2600

CC: Dr Sean Turner, Committee Secretary

Dear Chair,

Response: question-on-notice

On Monday 20 May 2024 representatives from the ANU School of Cybernetics appeared before the Senate Select Committee on adopting Artificial Intelligence (AI).

We received the following question-on-notice from Senator Shoebridge:

> Could you respond to this now or take it on notice. We've had the real-time experiment in South Korea, where they could see the threat of deepfakes coming and they prohibited the production of deepfake video images and voice in their election period. They have a 90-day election period. They put that legislation through in December. They resourced a body to police it, and it identified 388 deepfakes. That obviously meant that it didn't stop all deepfakes, because they found 388 of them. But the reporting seems to suggest that, having done that early, having put in place communications with social media and media outlets, those deepfakes had a much more limited impact on the election than otherwise would have happened. Are any of you in a position to respond to what happened in South Korea, whether now or on notice? That might be a very useful real-time model for us to look at.

Our response considers the amendment to the South Korean Official Election Act 2023. We situate this development in relation to the contested media and institutional arrangements of contemporary South Korea. We conclude this section by addressing two political "deepfakes" examples from South Korea.

We then offer a cybernetic approach to “deepfakes”. One that acknowledges technical systems are comprised of people, ecology, and technology. These systems have backstories and hidden stories that can provide a richer interpretation of how we encounter, imagine, accommodate and regulate these systems in the present.

Republic of Korea Public Official Election Act 2023

On 5 December 2023 the National Assembly of the Republic of Korea, through a parliamentary special committee on political reform, passed a revision to the Public Official Election Act.

The revision is available from:
https://nec.go.kr/site/nec/ex/bbs/View.do?cbIdx=1130&bcIdx=196646
and included as Appendix 1.

An exert translated from Korean to English reads:

> Introduction of new regulation provisions to AI-related new technology, such as deepfake videos:
>
> - (From 90 days prior to an election) Any AI-generated deepfake content (AI-generated images, sounds and/or videos that are hard to distinguish from real ones) for the purpose of aiding electoral processes are completely prohibited. If violated, it will result in up to seven years in prison or fines of up to AU$55,000 (50 million won).
>
> - (Out of the period of 90 days prior to an election) It is mandatory to clearly indicate that any AI-assisted content is virtual information as guided by the National Election Commission. If violated, it will result in fines of up to AU$11,000 (10 million won). (Rectification in the Public Official Election Act 2023)

The amendment took effect on 11 January 2024 ahead of the 10 April 2024 general elections (Korea Times 2023. Available at: https://www.koreatimes.co.kr/www/nation/2024/06/113_364513.html)
Reporting of South Korean election “deepfake” detection appeared on 19 February 2024 in ***The Korea Times***:

The National Election Commission (NEC) said on Monday [19 February] that it had identified 129 instances of election-related content utilizing deepfake

technology from Jan. 29 to Feb. 16. Each of these cases was determined to be in violation of the Public Official Election Act (Hyo-jin, 2024. Available at: https://www.koreatimes.co.kr/www/nation/2024/04/113_369059.html)
Following the 10 April election, Dain Oh of ***Readable***, reported that the National Election Committee claimed 388 cases that had violated "the Public Official Election Act — Article 82-8, with deletion requests" (Oh, 2024. Available at: https://thereadable.co/security-in-numbers-388-deepfakes-appeared-in-south-korean-elections/)

Without access to source documentation or clarity on "deefakes" detection and classification methods it is difficult to evaluate the National Electoral Commission claims, as reported.

South Korean context

The urgency to amend the Public Official Election Act 2023 ahead of the 2024 elections should be considered alongside the local challenges to the Republic of Korea electoral processes.

Turkish-American sociologist Zeynep Tufekci who has worked extensively on social media and the COVID-19 pandemic notes "Misinformation is not something that can be overcome solely by spelling out facts the right way. Defeating it requires earning and keeping the public's trust" (Tufekci 2024). A consideration of the South Korean "deepfake" situation emphasises the importance of maintaining public trust in and by the Australian political, media and legal systems, and our electoral integrity processes.

Since the impeachment of Park Geun-hye in 2017 an ongoing period of political turmoil has unfolded leading to "a wide variety of false and misleading information spread rapidly through online and social media platforms" (Yoo et al 2022 p3).

In their assessment of South Korean fake news networks in the 2020 election, Sheehy et al, included as Appendix 2, have pointed to the constrained media landscape: "South Korean news consumers express low levels of trust and approval of the news media overall, this distrust and dissatisfaction expressed by South Koreans constitutes a significant problem" (Sheehy et al 2024, p 4). The drift from brand news sources to online news aggregation and the very high engagement of news online video via youtube.com have contributed to this media environment (Newman et al 2019, p142).

The 2019 Reuters Institute Digital News Report report included “trust” in news media surveys from Australia and South Korea (Newman et al 2019). The report assessed South Korea as ranked last on trust out of the 38 markets surveyed with just 22% of respondents having “overall trust” in news media (Newman et al 2019, p142). Australia was ranked 18 out of 38 with 44% of respondents having “overall trust in new media” (Newman et al 2019, p131).

The 2023 Reuters Institute Digital News Report noted that in South Korea “trust in the news has been stuck at a low level, with just 28% of respondents saying that they “trust most news most of the time’” (Newman 2023, p 143). The Reuters Institute ranked South Korea 41 out of 46 countries surveyed. The report authors noted that South Korean “Respondents gave particularly low marks to the performance of the Korean media in representing the socially underprivileged, monitoring the government and public figures, and keeping an eye on corporate activities ((Newman 2023, p 142). The 2019 report correlated the low trust metric with media accountability - “The reasons are clear with just a fifth (21%) agreeing that the news media are doing a good job in monitoring powerful people and businesses (Newman 2019, p 141). In the 2023 Reuters Institute survey, 43% of Australian respondents expressed trust in media, ranking Australia 14 out of the 46 countries surveyed (Newman 2023, p 127).

A recent report of the 2024 South Korean election prepared by the International Institute for Democracy and Electoral Assistance (IDEA) noted South Korean politics has “traditionally been characterized by deep polarization, with politics being highly antagonistic and personality driven” (Spinelli 2024. Available at: https://www.idea.int/news/2024-south-korean-national-assembly-election-efficiency-amid-political-polarisation). The report described the 2024 election campaign as “significantly tainted by hostile discourse, slanderous attacks between rival parties, and harsh rhetoric. Policy and reform discussion was notably absent from the discourse” (Spinelli 2024).

This context is informative when considering the recent amendments to the Public Official Election Act in South Korea in late 2023. We believe it is important to draw upon a broader appreciation of the social and cultural histories, institutions and relational processes rather than a focus primarily towards alarm arising from computational methods applied to digital media production alone. Looking at two examples from South Korea is helpful in this regard.

In 2022 Republic of Korea presidential candidates, including the People Power Party's candidate Yoon Suk Yeol [elected President in May 2022], commissioned and deployed an "official deepfake". Under the headline "Deepfake democracy: South Korean candidate goes virtual for votes" on France 24.com, the Agence France-Presse wrote: "In a crowded campaign office in Seoul, young, trendy staffers are using deepfake technology to try to achieve the near-impossible: make a middle-aged, establishment South Korean presidential candidate cool" (France 24, 2022. Available at: https://www.france24.com/en/live-news/20220214-deepfake-democracy-south-korean-candidate-goes-virtual-for-votes).

In late 2023, a 46 second video, titled "Artificially crafted confession speech of President Yoon" was posted to Instagram, TikTok and Facebook. ***Korea JoongAng Daily*** reported "That the video was originally thought to be a deepfake, but further investigation revealed it to be a manipulated compilation of Yoon's statements from a televised debate in February 2022 when he was running for president" (Korea JoongAng Daily 2024). On 23 February 2024 the Korea Communications Standards Commission blocked the streaming of the video on social media. Presidential spokesperson Kim Soo-kyung claimed: "Though some media outlets called the video political satire or justified its use because it was labelled 'artificially crafted' this runs counter to media ethics to combat disinformation" (Korea JoongAng Daily 2024). The spokesperson explained "Even if the viral video of President Yoon is labelled as fake, it should still be eradicated, as edited versions of the video without the label are widely being spread across social media" (Korea JoongAng Daily 2024). This intervention alarmed Yoon Chang-hyeon head of the National Union of Media Workers who claimed the action sought to gag critical voices and that the regulator was acting as a "state censorship organ ... more concerned about who is the subject of satire" (Chan-kyong 2024).

These examples show how "deepfakes" operate as a "conceptually ambiguous buzzword" (Birrer and Just 2024, p 5). What was positioned as "deepfake democracy" in 2022 was cause for concern and contested intervention in 2024. Birrer and Just in their paper "What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape", included as Appendix 3, argue "'deepfakes' have so far caused far less turmoil than less sophisticated forms of visual disinformation and decontextualized images" (Birrer and Just 2024, p5).

These South Korean examples are informative when considering “deepfake” regulation of political speech in an Australian setting. ANU Visiting Fellow Andrew Ray wrote on the topic of “political deepfakes” in 2021 under the title “Disinformation, deepfakes and democracies: The need for legislative reform” (Ray 2021, p983). This paper is included as Appendix 4. Ray suggested two amendments to the electoral act seeking to balance “significant free speech concerns, as well as questions about where liability should fall” (Ray 2021, p983). Ray addresses the limitations of private (intellectual property and tort law), public remedies (Electoral Act), and Constitutional complications under the operation of the implied freedom of political communication pointing to the “unintended chilling effects around freedom of expression, harming rather than protecting democratic institutions (Ray 2021, pp. 983-1005).

Our overview of “deepfakes” in South Korea encourages a broader consideration of the interrelationships of institutions and histories in shaping the social and cultural aspects of how we use and talk about technology.

A cybernetic approach

At the School of Cybernetics our research often commences with an enquiry into the social and cultural origins of technical systems. We are attentive to the people and places who imagine and build them. We seek to reveal the entwined attitudes, techniques, and outputs that reveal interdependencies that evolve over time and geographies.

“Deepfakes” origins

The positioning of “Deepfakes” can be seen as consistent with what psychologist Amy Orben describes as the “Sisyphean cycle of technological panics” (Orben, 2020, p. 1). Orben’s analysis starts with a “panic creation” inspired by a “new technology”. This is followed by political concern and public expectation about the harms of the new technology and turning to researchers with specialised knowledge. But research can often be slow due to lack the theoretical and methodological frameworks to efficiently produce evidence quickly and effectively to inform policy interventions. The cycle evolves as the current “new technology – panic creation” subsides with the introduction of a new “panic creation” inspired by a “new technology” (Orben 2020, p. 4). Orben suggests the cycle can be broken by combining research about “older technologies [with] more current research considering recent technological developments” (Orben 2020, p11). The School of Cybernetics research agenda embraces framing questions about our futures informed by our present and pasts. Through a cybernetic approach, our School applies critical thinking and critical doing to imagine, build and maintain safe, responsible and sustainable systems.

The School of Cybernetics has written about the pasts, present and futures of: The National Library of Australia and artificial intelligence in ***Custodians and Midwives: the Library of the Future*** (Bell et al 2021); the metaverse in ***Blueprints and Backbones*** (Bell et al 2023); and Australian innovation with cybernetics, art, and technology in ***Australian Cybernetic: A point through time*** (Meares et al 2024). In addition to a Master and PhD program the School delivers a program of one-day workshops including a course titled ***Decoding AI.***

Tracing the emergence, ambiguity, adaptation, and adoption of "deepfakes" is informative as we contemplate the shifting futures of the term, the techniques, the motivations, the implications, and potential mitigation and regulation options.

"Deepfakes" first appeared in November 2017 as an anonymous username to register and post on the American online aggregation, rating and forum site Reddit (Cole 2017, Birrer & Just 2024, p. 2). Registered Reddit users, known as "Redditors", can upload content and web links as posts on Reddit. The posts are organized by subject into user-created boards called "communities" or "subreddits", allowing comments and rankings to enhance engagement (Reddit 2024).

Reddit user "Deepfakes" created a subreddit labelled "r/deepfakes" to share videos that included non-consensual pornographic videos edited to include the faces of women celebrities (Cole 2017). The inclusion of "fake" in the username "Deepfakes" and the subreddit "r/deepfakes" provided disclosure of and drew attention to the visual deception. What distinguished this moment was the amplification of aspects of computational production techniques used to create the visual deception (Cole 2017).

From inception, the power and associated attention of a "deepfake" video was not located in the images produced being undetectable from 'the real' but in the necessary admission of how the images were produced and edited using computational methods (Taylor 2021, p. 2, Weikmann & Lecheler 2023, p. 3703). If a "deepfake" video is not conveyed through labelling, visual, audio or editing clues or a declaration or accusation it is just a video.

On 12 December 2017 the online magazine ***Motherload*** published an article by journalist Samantha Cole titled “AI-assisted fake porn is here and we are all fucked” (Cole 2017). As the first article the use of the term this positioning is consistent with Orben’s “new technology - panic creation” (Orben 2020, p. 4). Cole interviewed “deepfakes” and wrote:

> “According to deepfakes — who declined to give his identity to me to avoid public scrutiny — the software is based on multiple open-source libraries, like Keras [an open-source library that provides a Python interface for machine learning] with TensorFlow [open-source software library for machine learning] backend. To compile the celebrities’ faces, deepfakes said he used Google image search, stock photos, and YouTube videos” (Cole 2017).

The introduction of then novel and accessible media and machine learning libraries and editing methods for digital media production that combined computing capability, computational methods, editing software, and digitised source materials have contributed to the anxiety and notoriety of the colloquial term “Deepfakes” (Cole 2017, Hao 2018, Citron & Chesney 2019, Birrer & Just 2024).

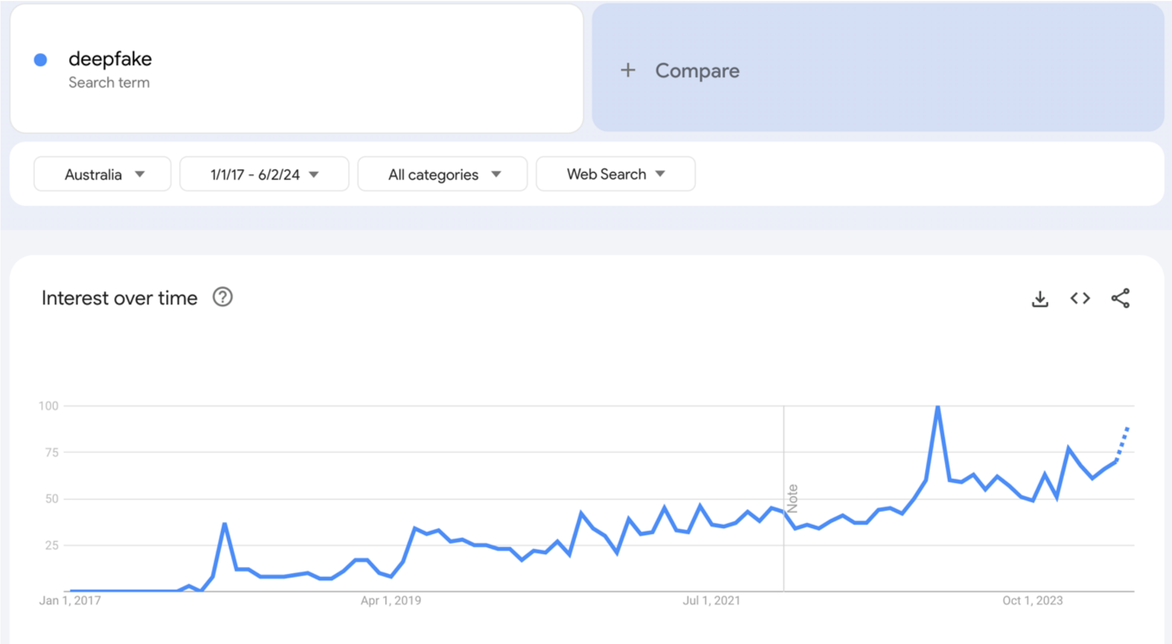
In a follow-up ***Motherboard*** article in January 2018 titled “We are truly fucked: Everyone is making AI-generated fake porn now”, Cole claimed, “the word ‘deepfake’ itself is now a noun for the kinds of neural-network generated fake videos their namesake pioneered” (Cole 2018a).

In February 2018 Reddit banned the subreddit “r/deepfake” by adjusting the Reddit Content Policy “regarding involuntary pornography and sexual or suggestive content involving minors” (Cole 2018b, Hern 2018). Reddit followed sites such as Discord, Gfycat, Pornhub and Twitter, which banned “involuntary pornography” earlier in 2018. We note the introduction of the ***Criminal Code Amendment (Deepfake Sexual Material) Bill 2024*** by Attorney-General The Hon. Mark Dreyfus KC MP on 5 June 2024 that seeks to impose “criminal offences to ban the sharing of non-consensual deepfake sexually explicit material” (Dreyfus 2024).

The portmanteau of “deep” and “fake” builds upon the “Deep Learning” machine learning techniques, which may include the use of “generative adversarial networks” or GANs (Goodfellow et al 2014) popularised in a 2015 computer-science paper titled "Deep Learning" by Yann Lecun, Yoshua Bengio, and

Geoffrey Hinton (Lecun et al. 2015). “Fake” rose to prominence at this time through the frequent use of “fake news” by US Presidential candidate Donald J Trump during the 2016 Presidential election period (Farhall et al. 2019).

The increased prominence the word “deepfake” can be observed through a Google trends graph (Figure 1. and Figure 2.) that provide insight to popular search queries that are entered into Google Search across regions and languages.



deepfake
Search term

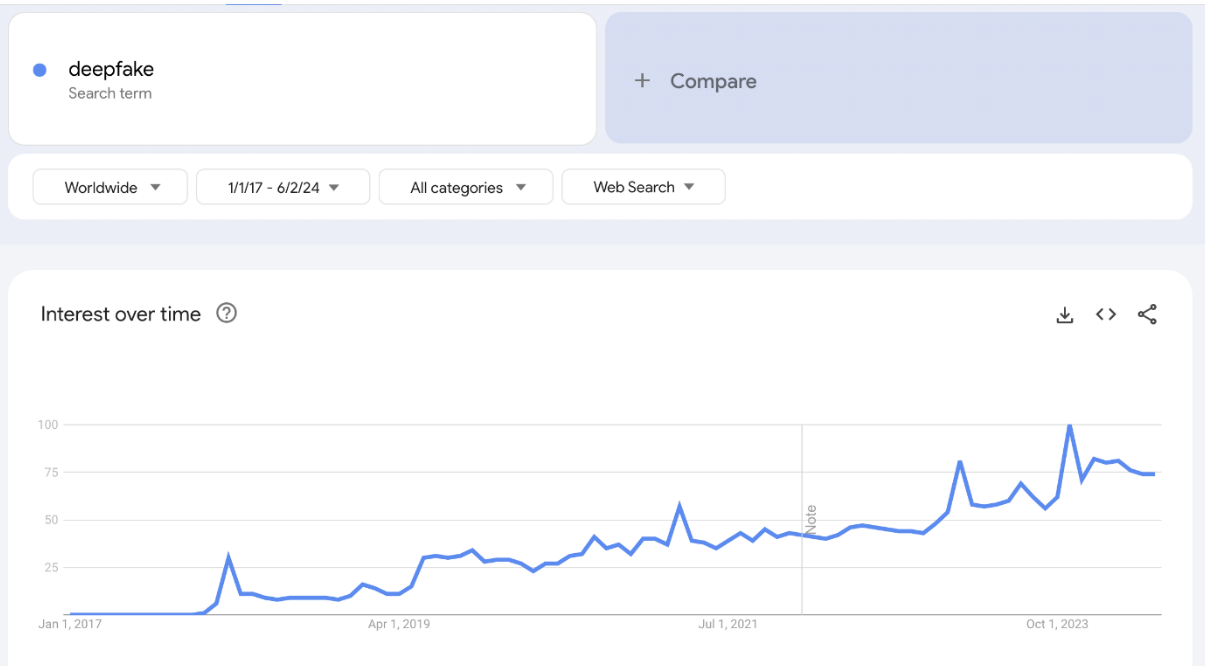
Australia

1/1/17 - 6/2/24

All categories

Web Search

Figure 1. Graph of Google Trends search term “deepfake” for the period 1 Jan 2017 to 2 June 2024 for the region of “Australia”. “A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular” (Google 2024). LINK



deepfake
Search term

Worldwide
1/1/17 - 6/2/24
All categories
Web Search

Figure 2. Graph of Google Trends search term “deepfake” for the period 1 Jan 2017 to 2 June 2024 for the region of “Worldwide” “A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular” (Google 2024). LINK

The popularity of the term was recognised by the Macquarie Dictionary committee for the Word of the Year 2018 which awarded an honourable mention to “deepfake” citing:

> The deepfake is a real sign of our times. Incredible advances in artificial intelligence have enabled the creation of remarkably accurate likenesses, and social media has enabled the lightning speed with which these videos are spread. Suddenly the deepfake is something that we all should be taking into account when we try to sort out the fake from the real in the news we are fed. (Macquarie Dictionary 2019)

The ambiguity of the new word “deepfake”, which was introduced as an anonymous social media username, has developed into an emotive description inspiring media, scholarship, social, cultural and political debate. This ability for adaption and broad interpretation and application of the term has contributed to its adoption and its ascribed power, threat and promise.

Researcher Alena Birrer and Professor Natascha Just argue that “much of the current debate is driven by anecdotal and speculative alarmism than by well-founded evidence and reasonable predictions” (Birrer and Just 2024, p2). The literature review Birrer and Just conducted asserts that there is “no universally accepted definition” but rather “deepfakes” acts as “conceptually ambiguous buzzword akin to ‘fake news’” (Birrer and Just 2024, p5). They acknowledge agreement on “the use of technology as a key characteristic of deepfakes, but the lack of clarity on the specific technology required blurs the conceptual boundaries between deepfakes and less sophisticated audiovisual manipulations know as ‘cheap fakes or ‘shallow fakes’” (Birrer and Just 2024, p5). They cite 2023 research that asserts that “deepfakes have so far caused far less turmoil than less sophisticated forms of visual disinformation and decontextualised images” (Birrer and Just 2024, p8).

Birrer and Just argue “deepfake technology may not introduce entirely new and unique regulatory problems at present, it can amplify existing problems” (Birrer and Just 2024, p1) they state:

> … deepfakes do not introduce fundamentally new and unique regulatory challenges. Instead, they add to the repertoire of tools available for spreading harmful or illegal content such as disinformation and non-consensual pornography. Consequently, the primary challenge lies in the effective oversight and enforcement of existing rules, along with careful considerations of required adjustments. This also necessitates consideration of potential unintended consequences when crafting countermeasures. (Birrer and Just 2024 p9)

Knowing aspects of the “deepfakes” origin story and the contemporary interpretations, evolving techniques, and imaginary potentials reminds us that disclosure of the computational production techniques, was always and remains essential, for a video to be classified as a “deepfake” video. Rather than avoiding detection, we argue that a “deepfake” video seeks to be exposed and amplified by being labelled a “deepfake”.

“Deepfakes” as mimesis

Mimesis can be understood as imitation or reproduction of the supposed words or image of someone else or imitation of the real world. The concept has informed identity, philosophy, art and literary criticism for thousands of years, or longer. The emotive narrative technique of replacing faces and voices, expressed as mimesis, is evident in the long histories of mythology, folklore, speculative fiction, theatre and cinema. While our storytelling techniques and methods may change over time the stories we tell each other can persist. By seeing “deepfakes” as mimesis we can make clear the important expectations of an audience in meaning making (Potolsky 2006, p4).

In 2017 in the first media report on “deepfakes”, Reddit user “Deepfakes” told the ***Motherboard*** online magazine “I just found a clever way to do face-swap” (Cole 2017). Two years prior to the “r/deepfakes” 2017 moment, in 2015, Laan Labs released “Face Swap Live” which provided an application for a mobile device that enabled the device camera and video display features to transpose live video of one user's face, as shown on the device display, with another user’s face at the same time (Lann Labs 2024). Automating a form of theatrical masking can be traced all the way back to early cinema and animation production through the work of Eadweard Muybridge’s zoopraxiscope (1879) and Max Fleischer who developed the rotoscope in 1914 (See: https://en.wikipedia.org/wiki/Rotoscoping).

The array of data, computational methods, techniques, and capacity, as well as reference texts and mixed genres that are required for “deepfakes” production and distribution infrastructures also demand audience participation. As a form of image and audio production “Deepfakes” contribute to the ways we make meanings and they also challenge our understandings of how meaning gets made.

Associate Professor of Filmmaking at the University of Reading Dominic Lees recognises the important role of the audience in association with “deepfakes”:

> … despite the honest full disclosure [of a “deepfake”], deception is intrinsic to this and all deepfakes. The pleasures for the audience include the enjoyment of feeling deceived, and appreciation of the technological skill involved in achieving this deception. There is a parallel with our appreciation of the skills of a secular magician conducting cards or conjuring tricks, despite our knowledge that what we are witnessing is trickery (Lees 2024, p110).

Lees’s reference to magicians recalls the stagecraft of the early film makers such as George Méliès, David Devant, Gaston Velle, and Alexander and Adelaide Hermann who translated illusion skills and helped develop new editing techniques the nascent film medium encouraged and that we now take for granted as visual grammar

Examples include:
Un homme de têtes [The four troublesome heads] (1898) - Georges Méliès
https://youtu.be/N71Ylc-EcJU?feature=shared

The Haunted Curiosity Shop (1901) - W. R. Booth
https://youtu.be/mfDearYwBF8?feature=shared

Uncle Josh at the Moving Picture Show (1902) - Edwin S. Porter
https://youtu.be/J1x_EslAylc?feature=shared

A Prize Fight or Glove Fight Between John Bull & President Kruger (1900) - John Sloane Barnes
https://player.bfi.org.uk/free/film/watch-a-prize-fight-or-glove-fight-between-john-bull-and-president-kruger-1900-online

These early films were the cutting- edge tech of their day. Through exploration and iteration the meaning-making possibilities of the medium were negotiated, accepted and resisted. Thinking about mimesis in association with "deepfakes" encourages us, as with cinema, to also include an audience who are interested not only in the content of deception but how that content is made and a deception attempted and achieved.

Politicians and cinema

The ongoing process of political "deepfakes" can be informed by a brief consideration of how politicians accommodated early cinema. One of the first Australian feature films ever made, ***The Story of the Kelly Gang*** by Charles Tait, 1906, also became one of the first feature films to be banned when state authorities became concerned the film "glorified criminal activities" (Milner 2019).

In a description of scenes knowable to us in 2024 the Australian Labor Party in 1909 embraced election campaigning by "cinematograph" (***The North Western Advocate and the Emu Bay Times***, 21 Sepetember 1909, p3. Available at: http://nla.gov.au/nla.news-article64869373).

> LABOR AND THE CINEMATOGRAPH.
>
> MELBOURNE. Monday. — A private demonstration of cinematograph pictures taken in connection with the Labor Party election campaign was, held to-day. The pictures include presentations of the Labor caucus at work; Labor members entering and leaving Parliament

> House; Mr. Fisher and other leaders delivering public addresses; and other political and Parliamentary scenes. In addition are scenes of men and women at work in various industries; also, pictures of slum life compared with the life of the wealthy. These will be shown throughout the Commonwealth.

By 1911 ***The Beverley Times*** in Western Australia was reporting on "Marvellous things the cinematograph may accomplish" noting that "the science of cinematogtaphy is only in its infancy" (***The Beverley Times***, 23 December 1911 , p2. Available at: http://nla.gov.au/nla.news-article90724741). The article covers the recent use of cinematograph film to identify rioters in France noting that "There is no escaping the truth of the cinemagraph film" (***The Beverley Times***, 23 December 1911, p2.) But it is the wonder of sound and image genres connected with worldwide distribution that sees a future "when all these scattered inventions are combined, we may expect an amazing apparatus" (***The Beverley Times***, 23 December 1911, p2). The article explains:

> Mr Gaumont has lately Invented a machine which he calls the "chronograph." This is a perfect combination of the moving picture and the talking-machine. This adds considerably to the realistic effect of a picture by giving the exact sounds which accompanied the movements in real life while the photograph was being taken…
> But the cinematograph is likely to be most useful in political campaigns and especially at general election times. We shall no longer have to be content with reading reports of important speeches made by the party leaders. Every elector, even those in the most remote towns and villages, will have the opportunity of hearing and judging for himself those vital utterances as well as if he had heard the original speeches. (***The Beverley Times***, 23 December 1911, p2)

By 1929 Australian Prime Minister James Scullin was described as "Australia's first talkie star" as he inaugurated the Australasian edition of the Fox-Movietone News "Delivering his message from the screens of the State, Regent and Theatre Royal, the Prime Minister, Mr Jamnes Scullin, looms as this week's main attraction" (***The Evening News***, 1 November 1929, p13. Available at: http://nla.gov.au/nla.news-article119010591

From banning ***The Story of the Kelly Gang*** to the Prime Minsiter becoming "Australia's first talkie star" a consideration of politicians and cinema offers lessons for our “deepfakes” moment on how “new technologies” are adapted and adopted.

Conclusion

Our response considered the amendment to the South Korean Official Election Act 2023 and connected this initiative to the contested media and institutional arrangements of contemporary South Korea. We offered two political “deepfakes” examples from South Korea that demonstrate how “deepfakes” operate as a “conceptually ambiguous buzzword” (Birrer and Just 2024, p 5). We believe the interrelationships of institutions and histories are important in shaping the social and cultural aspects of how we use and talk about technology. We take a cybernetic approach to “deepfakes”. Understanding the origins helps us ask informed questions about our present. We explored the concept of mimesis and “deepfakes” to include the overlooked aspects of audience meaning-making beyond a focus just on computational editing techniques. Our social and cultural enquiry traced the technology to early cinema and briefly looked at some of the first examples of Australian politicians using new technologies. An approach to “deepfakes” that includes social, cultural and technological relationships, a cybernetic approach, allows us to see an ongoing and relational process of navigating methods and outputs of media production, distribution, ambiguity, adaptation, adoption, and regulation as we seek to understand, ameliorate, and accommodate “deepfakes”.

Kind regards,

[redacted]

Andrew Meares
Associate Professor
Australian National University
Deputy Director
School of Cybernetics

Supported by: Professor Katherine Daniell, Thomas Biedermann, and Ellen O’Brien.

References

Baker, Simon. (2014). Politic and film 1903-1935. ***Screenonline***. Available at: http://www.screenonline.org.uk/film/id/1196906/index.html

Bell, Genevieve., Zafiroglu, Alex., Assaad, Zena., Bradley, Charlotte., Cooper, Ned., O'Brien, Ellen., Reid, Kathy., and Ruster, Lorenn. ***Custodians and Midwives: the Library of the Future.*** School of Cybernetics, The Australian National University. Available at: https://cybernetics.anu.edu.au/projects/custodians-and-midwives/

Bell, Genevieve., Gould, Maia., O'Brien Ellen., & Paulk, Charlie. (2022). ***Backbones & Blueprints: cybernetic approaches to the metaverse.*** School of Cybernetics, The Australian National University. Available at: https://cybernetics.anu.edu.au/projects/the-metaverse-building-systems-by-design/

Birrer, Alena., & Just, Natascha. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. ***New Media Society***, 0(0). https://doi.org/10.1177/14614448241253138

Chan-kyong, Park. (2024, 26 February). South Korea's Yoon accused of using 'fake news' crackdown to gag dissent ahead of polls. ***South China Morning Post.*** Available at: https://www.scmp.com/week-asia/politics/article/3253245/south-koreas-yoon-accused-using-fake-news-crackdown-gag-dissent-ahead-polls

Citron, Danielle. K. & Chesney, Robert. (2019). "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security", in 107 ***California Law Review*** 1753. Available at: https://scholarship.law.bu.edu/faculty_scholarship/640

Cole, Samantha. (2017, 12 December). AI-Assisted Fake Porn Is Here and We're All Fucked, ***Motherboard***. Available at: https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Cole, Samantha. (2018a, 25 January). We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. ***Motherboard***. Available at: https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley

Cole, Samantha. (2018b, 8 February). Reddit Just Shut Down the Deepfakes Subreddit, ***Motherboard***. Available at: https://www.vice.com/en/article/neqb98/reddit-shuts-down-deepfakes

Dreyfus, Mark. (2024). New criminal laws to combat sexually explicit deepfakes. Media Release. ***Australian Governmnet Attorney-General's Department***.

Available at: https://ministers.ag.gov.au/media-centre/new-criminal-laws-combat-sexually-explicit-deepfakes-05-06-2024

Farhall, Kate., Carson, Andrea., Wright, Scott., Gibbons, Andrew., & Lukamto, William. (2019). Political Elites' Use of Fake News Discourse Across Communications Platforms. ***International Journal Of Communication***, 13, 23. Available at: https://ijoc.org/index.php/ijoc/article/view/10677/2787

France 24. (2022, 14 February). Deepfake democracy: South Korean candidate goes virtual for votes, ***France 24***. Available at: https://www.france24.com/en/live-news/20220214-deepfake-democracy-south-korean-candidate-goes-virtual-for-votes).

Goodfellow, Ian., Pouget-Abadie, Jean., Mirza, Mehdi., Xu, Bing., Warde-Farley, David., Ozair, Sherjil., Courville, Aaron., Bengio, Yoshua. (2014). Generative adversarial nets. In ***Advances in neural information processing systems*** (pp. 2672–2680). Available at: https://doi.org/10.48550/arXiv.1406.2661

Hao, Karen. (2018). “Inside the world of AI that forges beautiful art and terrifying deepfakes”. ***MIT Technology Review.*** Available at: https://www.technologyreview.com/s/612501/inside-the-world-of-aithat-forges-beautiful-art-and-terrifying-deepfakes/

Hern, Alex. (2018, 8 February). Reddit bans 'deepfakes' face-swap porn community, ***The Guardian***. Available at: https://www.theguardian.com/technology/2018/feb/08/reddit-bans-deepfakes-face-swap-porn-community#:~:text=Social%20news%20site%20Reddit%20has,first%20AI%2Dcreated%20video%20clips.

Hyder, Liz. (2023). Magicians and film-makers, masters of illusions. ***Historia***. 19 June 2023. Available at: https://www.historiamag.com/magicians-and-film-makers-masters-of-illusions/).

Lann Labs. (2024, 11 June). ***Featured Apps*** https://labs.laan.com/#apps

Lecun, Yann., Bengio, Yoshua., Hinton, Geoffrey. (2015). Deep Learning. ***Nature***, 521 (7553), pp.436-444. Available at: https://hal.science/hal-04206682/document

Lees, Dominic. (2024). Deepfakes in documentary film production: images of deception in the representation of the real. ***Studies in Documentary Film***, 18(2), pp. 108–129. Available at: https://doi.org/10.1080/17503280.2023.2284680

Macquarie Dictionary (2019). ***The Committee's Choice of the Word of the Year is…*** Available at: https://www.macquariedictionary.com.au/the-committees-choice-word-of-the-year-2018-is/

Meares, Andrew., McLennan, Amy., & Pegram, Caroline. (2022). ***Australian Cybernetics; A point through time***. School of Cybernetics, The Australian National University. Available at: https://cybernetics.anu.edu.au/futures/australian-cybernetic/#:~:text=Australian%20Cybernetic%20was%20imagined%20by,the%20early%201960s%20to%20today.

Milner, Johnny. (2019). From the silent era to the 21st Century.***Film and Sound Archive***. Available at: https://www.nfsa.gov.au/latest/ned-kelly-australian-cinema-1906-2019).

Newman, Nic., Fletcher, Richard., Kalogeropoulos, Antonis & Nielsen, Kleis Rasmus. (2019). ***Reuters Institute Digital News Report 2019***. Available at: https://reutersinstitute.politics.ox.ac.uk/our-research/digital-news-report-2019

Newman, Nic., Fletcher, Richard., Eddy, Kirsten., Robertson, Craig T., & Nielsen, Kleis Rasmus. (2023). ***Reuters Institute Digital News Report 2023***. Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

Orben, Amy. (2020). The Sisyphean Cycle of Technology Panics. ***Perspectives on Psychological Science***, 15(5), 1143-1157. Available at: https://doi.org/10.1177/1745691620919372

Reddit (2024, 11 June) ***Reddit User Agreement***. Available at: https://www.redditinc.com/policies/user-agreement

Sheehy, Benedict., Choi, Sujin., Khan, Md. Irfanuzzaman., Arnold, Bruce. Baer., Sang, Yoonmo., & Lee, Jae-Jin. (2024). Truths and Tales: Understanding Online Fake News Networks in South Korea. ***Journal of Asian and African Studies***, 0(0). Available at: https://doi.org/10.1177/00219096231224672

Taylor, Bryan, C. (2020). Defending the state from digital Deceit: the reflexive securitization of deepfake, ***Critical Studies in Media Communication***, 38(1), 1-17. Available at: https://doi.org/10.1080/15295036.2020.1833058

The Beverly Times (1911, 23 December). Possibilities in pictures Marvelous things the cinematopgraph may accomplish, ***The Beverly Times***, p2. Available at: http://nla.gov.au/nla.news-article206505243

The Evening News (1929, 1 November). News of the screen and stars, ***The Evening News***, p13. Available at: http://nla.gov.au/nla.news-article119010591

The North Western Advocate and the Emu Bay Times, (1909, 21 Sepetember 1909), Labor and the cinematograph. ***The North Western Advocate and the Emu Bay Times***, p3. Available at: http://nla.gov.au/nla.news-article64869373).

Tufekci, Zeynep. (2024, 8 June). An object lesson from Covid on how to destroy public trust. ***The New York Times***. Available at: https://www.nytimes.com/2024/06/08/opinion/covid-fauci-hearings-health.html

Yoo, Joseph., Kim, Daekyung., & Kim, Wi-Geun. (2022). Fake news on you, Not me: The Third-Person Effects of Fake News in South Korea. ***Communication Research Reports***, ***39***(3), 115–125. Available at: https://doi.org/10.1080/08824096.2022.2054790

Weikmann, Teresa., & Lecheler, Sophie. (2023). Visual disinformation in a digital age: A literature synthesis and research agenda. ***New media & Society,*** Vol. 25(12), 3696-3713. Available at: https://doi.org/10.1177/14614448221141648

'23. 12. 20. 본회의 의결 정치관계법 일부개정법률 주요내용

I 공직선거법

1 인공지능 기반 딥페이크 영상 등 새로운 기술에 대한 규제 조치 도입(§82의8①·②, 250④, 255⑤, 261③4 신설 등)

현 행	개 정
<신 설>	► (선거일 전 90일부터 선거일까지) 선거운동용 딥페이크 영상등(인공지능 기술 등을 이용하여 만든 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등) 전면 금지(7년 이하 징역 또는 1천만원 이상 5천만원 이하 벌금) ► (선거일 전 90일부터 선거일까지의 기간이 아닌 때) 딥페이크영상등을 가상의 정보라는 사실을 중앙선거관리위원회규칙으로 정하는 바에 따라 표시(위반 시 1천만원 이하 과태료 부과)하고, 표시의무 위반 허위사실공표는 가중처벌

※ 공포 후 1개월이 경과한 날부터 시행

2 정보통신망 위법게시물 게시자에 대한 삭제요청 근거 마련(§82의4③·④)

현 행	개 정
► 각급 선거관리위원회 또는 후보자는 「공직선거법」을 위반한 영상 등의 정보 관련 인터넷홈페이지 관리·운영자 또는 정보통신서비스 제공자에게 삭제요청 등 가능	► 각급 선거관리위원회 또는 후보자는 「공직선거법」을 위반한 영상 등의 정보를 '게시한 자'에게도 삭제요청 등 가능

※ 공포한 날부터 시행

3 예비후보자의 선거운동 방법 확대(§60의3①5)

현행	개정
► 예비후보자 표지물을 착용하는 행위 가능	► 예비후보자 표지물을 착용하는 행위를 착용하거나 소지하여 내보이는 행위로 확대

※ 공포한 날부터 시행

Ⅱ 정당법

1 비례대표국회의원선거 후보자 추천 민주화(§36의2 신설)

현행	개정
<신 설>	► 정당은 당헌·당규 등에서 정한 민주적 절차에 따라 공직선거의 후보자를 추천(선언적 규정)

※ 공포한 날부터 시행

Ⅲ 정치자금법

1 여성추천보조금 지급구간 조정(§26②)

구 분	현 행		개 정	
	여성추천비율	보조금 총액	여성추천비율	보조금 총액
보조금 지급 구간			40%이상	40%
	30%이상	50%	30%이상 ~ 40%미만	30%
	20%이상 ~ 30%미만	30%	20%이상 ~ 30%미만	20%
	10%이상 ~ 20%미만	20%	10%이상 ~ 20%미만	10%

※ 법 시행 이후 여성추천보조금을 배분·지급하는 경우부터 적용

2 여성정치발전을 위한 경상보조금의 용도 구체화(§28② 후단 신설)

현 행	개 정
<신 설>	► 1. 여성정책 관련 정책개발비 2. 여성 공직선거 후보자 지원 선거관계경비 3. 여성정치인 발굴 및 교육 관련 경비 4. 양성평등의식 제고 등을 위한 당원 교육 관련 경비 5. 여성 국회의원·지방의회의원 정치활동 지원 관련 경비 6. 그 밖에 여성정치발전에 필요한 활동비, 인건비 등의 경비로서 중앙선거관리위원회규칙으로 정하는 경비

※ 공포 후 3개월이 경과한 날부터 시행

JAAS

Original Research Article

Journal of Asian and African Studies
1–28

DOI: 10.1177/00219096231224672
journals.sagepub.com/home/jas

Sage

Truths and Tales: Understanding Online Fake News Networks in South Korea

Benedict Sheehy
Canberra Law School, University of Canberra, Australia

Sujin Choi
Department of Media, Kyung Hee University, Republic of Korea

Md Irfanuzzaman Khan
Canberra Business School, University of Canberra, Australia

Bruce Baer Arnold
Canberra Law School, University of Canberra, Australia

Yoonmo Sang
Department of Media Communication, Sungshin Women's University, Republic of Korea

Jae-Jin Lee
Department of Media and Communication, Hanyang University, Republic of Korea

Abstract

This study investigates the features of fake news networks and how they spread during the 2020 South Korean election. Using actor–network theory (ANT), we assessed the network's central players and how they are connected. Results reveal the characteristics of the videoclips and channel networks responsible for the propagation of fake news. Analysis of the videoclip network reveals a high number of detected fake news videos and a high density of connections among users. Assessment of news videoclips on both actual and fake news networks reveals that the real news network is more concentrated. However, the scale of the network may play a role in these variations. Statistics for network centralization reveal that users are spread out over the network, pointing to its decentralized character. A closer look at the real and fake news networks inside videos and channels reveals similar trends. We find that the density of the real news videoclip network is higher than that of the fake news network, whereas the fake news channel networks are denser than their real news counterparts, which may indicate greater activity and interconnectedness in their transmission. We also found that fake news videoclips had more likes than real news videoclips,

Corresponding author:
Md Irfanuzzaman Khan, Canberra Business School, Faculty of Business, Government and Law, University of Canberra, Bruce, ACT 2617, Australia.
Email: [redacted]

whereas real news videoclips had more dislikes than fake news videoclips. These findings strongly suggest that fake news videoclips are more accepted when people watch them on YouTube. In addition, we used semantic networks and automated content analysis to uncover common language patterns in fake news, which helps us better understand the structure and dynamics of the networks involved in the dissemination of fake news. The findings reported here provide important insights on how fake news spread via social networks during the South Korean election of 2020. The results of this study have important implications for the campaign against fake news and ensuring factual coverage.

Keywords

Fake news, South Korean election, social network analysis, YouTube, actor–network theory

Introduction

Fake news, a phenomenon amplified exponentially by new network technologies, has emerged as a formidable societal challenge. This phenomenon extends its reach across critical domains, from the integrity of democratic elections to the management of public health crises, as witnessed during the recent COVID-19 pandemic (Rudgard, 2020). For example, during COVID-19, misperceptions were successfully created through simple content alterations and the addition of popular anti-COVID-19 hashtags such as #COVIDIOT and #covidhoax to otherwise valid Twitter content, thus encouraging the hesitant and skeptical minority to be open to commenting, retweeting, like, and sharing rumors about vaccines' efficacy (Sharevski et al., 2022). The ubiquity of social media platforms that make possible the creation of such networks has assisted individuals and groups in spreading fake news, enabling the spread of their disinformation and misinformation at unprecedented speed, reaching more network participants and remaining longer in the public domain (Rhodes, 2022). In the political landscape, digital platforms have become conduits for the dissemination of false information and propaganda during electoral processes, exerting influence over voter behaviors and posing a threat to democratic principles (Azis Prasetyo and Aisyah, 2018; Igwebuike and Chimuanya, 2021). The 2016 US presidential election serves as a stark illustration, where an inundation of fake news eclipsed authentic narratives, leading to a widespread acceptance of erroneous information (Budak, 2019). This exposure to fake news predisposes individuals to adopt various political misperceptions (Ognyanova et al., 2020), ultimately shaping their subsequent behavior, including voting decisions (Cantarella et al., 2023).

Digital networks that host fake news, misinformation, and other forms of disinformation exist in a context in which traditional news media and political institutions are viewed with a growing level of mistrust, and the "wisdom of the ordinary, non-specialist 'hacker' and/or purported secret information from supposed insiders is believed instead" (Bleakley, 2023). Recent research reveals that sources of fake news frequently attack mainstream media organizations, claiming that they are biased and incapable of doing their jobs properly. A drop in trust in the media of 5% was predicted among the people who took part in the study if they were exposed to disinformation during the month leading up to the 2018 election. In addition, a discernible correlation emerges between consumption of fake news and diminished trust in mainstream media across all levels of political ideology (Ognyanova et al., 2020). Albright aptly characterizes the phenomenon, noting the rapid dissemination of emotionally charged messages on platforms like Twitter. This calculated distortion of attention hastens the spread of misinformation, giving rise to the establishment of alternative, often unfounded, narratives (Albright, 2017).

In this context, digital platforms readily embrace subversive and discriminatory claims, as they are swiftly reinforced by peers and prove challenging for authorities to counter effectively. Extant research clearly indicates that online social networks are usually formed among like-minded people who use digital platforms to both facilitate sharing among themselves and promote and bolster belief in both true/fake news within and beyond their groups (Bleakley, 2023). While individuals tend to seek out like-minded people, the algorithms of digital platforms have amplified this tendency through filtering challenging information and providing confirmatory information, which algorithms determine from platform users' previous behaviors and choices (Nolin and Olson, 2016; Pariser, 2011). Albright (2017) emphasizes the necessity of scrutinizing the fake news ecosystem, which can be accomplished by tracing the flow of information across expansive networks of websites, profiles, and platforms.

Against this backdrop, it becomes imperative to turn our attention to South Korea, a nation experiencing a period of unprecedented political transformation. In 2017, the impeachment of a president and the ensuing political upheaval were accompanied by a surge of false and misleading information propagated through online channels (Yoo et al., 2022). This unique socio-political environment sets the stage for an in-depth exploration of how fake news influenced the political landscape during the pivotal 2020 general election in South Korea. This study aims to examine the dynamics of fake news dissemination within this distinctive context, offering valuable insights for navigating the challenges posed by digital networks in the South Korean political environment.

The substantial societal costs stemming from the misuse of digital news networks highlight the urgency of this inquiry. While previous studies have examined the spread of misinformation during the COVID-19 epidemic, the specific influence of fake news network structures in a country like South Korea remains a critical knowledge gap. Given its unique political and social landscape which is very different from its Western counterparts, this study endeavors to elucidate how fake news impacted the political terrain of South Korea during the 2020 election.

The subsequent sections of this article are structured as follows. Section "Literature review" provides a comprehensive literature review, situating the phenomenon of fake news within the broader global and South Korean contexts. Section "Method: case study analysis" delineates the methodology, encompassing data collection and analytical approaches. Section "Results" presents the results, revealing the specific manifestations and consequences of fake news in South Korea. Section "Videoclip network" focuses on discussion, drawing connections between our findings and broader theoretical frameworks and practical context. Finally, in Section "Discussion," we conclude with a comment on this study's limitations and recommendations for future research.

Literature review

Misinformation in the South Korean context

Over the past several years, South Korea has experienced unparalleled political changes. In 2017, we saw the impeachment of a president. During the same time, a slew of false and misleading information proliferated quickly through online and social media channels during this time of political unrest (Yoo et al., 2022). Unison of people who share similar ideologies is to be expected in such a strongly polarized environment (Choi et al., 2020). Even when the information is erroneous, people in these echo chambers frequently absorb false information that supports their ideologies. According to the Institute for the Study of Journalism's Digital News Report 2019, among 38 countries surveyed, South Korean news consumers have the lowest level of trust in the news media (Newman et al., 2019). Furthermore, approximately 40% of South Korean news consumers access news through YouTube, and the country ranked highly in terms of podcasts usage (Newman et al., 2019).

Given that South Korean news consumers express low levels of trust and approval of the news media overall, this distrust and dissatisfaction expressed by South Koreans constitutes a significant problem. According to a survey conducted by the South Korea Press Foundation in March 2018, 69.2% of the 1500 respondents had seen or heard about manipulated or false information in the form of news distributed on social media (Yoo et al., 2022). Furthermore, the networks have been used to disseminate fake news aimed at undermining legitimate political processes during national elections in South Korea, including news stories about the major presidential candidates (Park and Youm, 2019). For example, in 2017, following the candlelight demonstrations and the subsequent impeachment of South Korea's former president, Park Geun-hye, fake news was widely distributed among supporters. In the aftermath of Guen Hye's impeachment, South Korean political parties used fake news to mobilize their supporters in order to gain an advantage in situations involving political divisions and confrontations between the pro-impeachment, progressive young generation and the anti-impeachment, conservative senior generation. The communications asserted that US President Donald Trump had expressed opposition to impeachment and that North Korea was the mastermind behind the impeachment scheme (Go and Lee, 2020).

While various research studies have looked at how misinformation spread during the COVID-19 epidemic (Freiling et al., 2023; Zhang et al., 2023), it is still unclear how the network structure of fake news influences its spread in a country like South Korea. It is politically and socially different from Western countries. Thus, this study examines how fake news influenced the political landscape of South Korea during the 2020 election.

Analytical framework: actor–network theory

The analytical framework for this study is based on actor–network theory (ANT). ANT challenges traditional sociological approaches that prioritize human agency and social structures. Instead, ANT treats both human and nonhuman actors as having agency and influence in shaping social phenomena. It emphasizes the importance of studying how networks are formed and how actors, both human and nonhuman, join and mobilize within these networks (Sharifzadeh, 2016). According to Latour (2007), the ANT incorporates a wide variety of actants, from tangible elements to abstract concepts like declarations and ideas. Interactions and networking among ANT's actants are its primary concern. These distinguishable interactions reflect the network's inscription. An actor network might represent a social network, so it comprises not just people interacting with one another, but also interactions with nonhuman actants. This is evident in the features of social media technology and legislation that mediate the relationship between humans (Labafi, 2020). Our research aims to uncover the differences between real news and fake news networks in terms of network density, geodesic distances, and centrality measures. These findings will contribute to ANT by illustrating how different actors and entities, including individuals, channels, and video-clips, are organized and interconnected within the network. This provides empirical evidence of the heterogeneous composition and structure of actor networks involved in the dissemination of real and fake news, highlighting the role of both human and nonhuman actors.

Fake news

The surge in deliberately fabricated false stories, often referred to as "fake news," has become a pressing concern in today's information landscape (Kar et al., 2023; Lazer et al., 2018). While some scholars express reservations about the term itself, citing its potential to erode the credibility associated with "news," alternative descriptors like "misinformation," "disinformation," or "fabricated news" have been proposed (Allcott and Gentzkow, 2017; Kar et al., 2023). Nevertheless, empirical evidence suggests that public interest in the term "fake news" surpasses related terms like "misinformation," "disinformation," and "fabricated news" by a significant margin, indicating its enduring relevance (Ansar and Goswami, 2021).

The transformative role of social media platforms deserves special attention. Originally conceived as spaces for social interaction, they have evolved into influential information ecosystems. Social media now facilitate not only connections but also the seamless sharing and reception of information across borders, fundamentally altering the way individuals engage with content (Grover et al., 2022). This evolution has given rise to algorithm-driven content curation, potentially leading to the formation of echo chambers where individuals are predominantly exposed to information that aligns with their existing beliefs (Rodrigues da Cunha Palmieri, 2023). Consequently, there is a growing concern regarding the reinforcement of pre-existing opinions and the potential isolation of individuals from diverse perspectives (Diaz Ruiz and Nilsson, 2023).

While significant strides have been made in detecting fake news online, much remains to be uncovered (Chen et al., 2015; Conroy et al., 2015; Bastick, 2021). Studies examining the propagation of political fake news in South Korean online communities offer valuable insights into the dynamics at play. Choi (2014) and (Choi, Yang, and Chen, 2018) delved into political discussions, revealing a notable centralization and cliquishness in information flow. These discussions were characterized by heightened emotion, particularly anger, and participants tended to refer primarily to like-minded messages. Furthermore, those with more reciprocal relationships and higher popularity within the forum tended to maintain or create more discussion ties. It is important to acknowledge that the dynamics of online political discussions may not mirror those of fake news distribution networks. In the case of fake news, if widely disseminated by a motivated and like-minded group, its distribution network might exhibit higher levels of centralization and cliquishness compared with typical discussion networks.

In addition, the popularity of the author may hold greater sway in fake news distribution than the emotional content or social effects of the message itself (Choi, 2014; Choi et al., 2018). Likewise, a study involving 10,000 users and 555,684 tweets suggests that factors such as emotion stability, polarity stability, hashtag consolidation ratio, hashtag diversity, lexical diversity, favorites count, and friends count may influence the propagation of both misinformation and information (Kar and Aswani, 2021). Similarly, recent research by Aswani et al. (2019) delves into the management of misinformation in social media, shedding light on factors contributing to its rapid propagation. Their analysis of approximately 1.5 million tweets in cases involving misinformation highlights the role of emotions and polarity in determining content authenticity. Notably, tweets with a higher element of surprise combined with other emotions are more likely to be associated with misinformation. Furthermore, tweets featuring neutral content are less prone to virality when it comes to spreading misinformation.

Building on this foundation, the current study undertakes a comprehensive assessment of the structure and content of fake news distribution, employing network analysis and automated content analysis. This approach represents a critical stride toward a deeper understanding of the mechanisms underpinning the dissemination of fake news. However, it is imperative to highlight that while social media data analysis effectively describes the phenomenon, it often falls short in addressing the underlying cause-and-effect relationships. Thus, an empirical data analysis approach is adopted to shed light on how fake news has been disseminated, addressing South Korea-specific questions. In doing so, we recognize the need to look beyond American-centric concerns and approaches, as the United States does not offer a universal model, particularly in an Asian context. For instance, the prevalence of fake news on platforms like Kakaotalk distinguishes South Korea's information landscape from that of other nations where distribution may be more prominent on platforms like YouTube, Facebook, and Twitter.

Having established the contextual backdrop, we now turn our attention to the methodology employed in this study to dissect the dissemination and characteristics of fake news surrounding the unfounded concerns about election fraud in South Korea's 2020 general election.

Method: case study analysis

In this research, we have selected the unfounded concerns about election fraud in South Korea's 2020 general election, as the allegations of electoral fraud gained a considerable amount of attention in South Korea (Kim, 2020). Although the electoral fraud issue is likely to stem from a political conspiracy, it was widely discussed on YouTube during the period of data collection and likely to provide in-depth insights in terms of their thematic genres and fake news formation. During the data collection phase, we had to differentiate between "fake news" and "real news." In terms of this classification, in the electoral fraud case, we classified the content that *supports or agrees with* the electoral fraud argument as fake news, whereas we classified content that refutes this argument as real news. We classified content that *simply introduced* the idea of electoral fraud as real news. Then, we addressed the four research questions specific to YouTube platforms:

1. How South Korean fake news is disseminated;
2. Whether its dissemination exhibits different path characteristics from that of real news;
3. Whether this spread is driven by author effect, message effect, or social effect;
4. Which words commonly and frequently appear in fake news messages.

To address these questions, we conducted four discrete analyses: (1) a social network analysis (SNA), (2) natural language processing, (3) semantic network analysis, and (4) automated content analysis. We conducted this analysis using two distinct network models in the form of channel network and videoclip network.

Social network analysis

SNA was chosen as a pivotal analytical approach due to its proficiency in elucidating the underlying structure and dynamics of social interactions within a digital platform like YouTube. By representing users, channels, and videoclips as nodes, and their interactions as edges, SNA enables us to visualize the patterns of engagement and identify key actors or content within the network. SNA also makes possible the calculation of various network indices, including measures of centrality, clustering, and density (Altuntas et al., 2022). These metrics offer valuable insights into the prominence and influence of specific nodes or channels, as well as the overall cohesion and connectivity of the network. This approach is particularly relevant when investigating the dissemination of information, as it illuminates how users and content are interconnected (Ulibarri and Scott, 2017). First, nodes represent participants in an independent network. Edges denote connectivity between nodes. Edges can represent followings, followers, mentions, and replies in social media. Clusters are heavily interconnected elements that are rarely connected to other blocks. Network density represents the average value of a random network connection. Denser networks contain more and/or more valuable relationships, thereby augmenting the average tie value (Eom et al., 2018).

Several alternative approaches could have been considered to complement or enhance the analysis of fake news dissemination in the context of South Korea's 2020 general election. One potential alternative approach is the integration of sentiment analysis with SNA. This combined approach would enable the assessment of emotional or attitudinal aspects associated with interactions within the network, providing a more nuanced understanding of user engagement patterns and sentiment dynamics (Röchert et al., 2020). Another viable alternative approach is social media analytics, which focuses on extracting meaningful insights from large-scale social media data through techniques like text mining, sentiment analysis, and content categorization (Khan and Malik, 2022). While social media analytics provides valuable insights into content characteristics and sentiment trends, it may not capture the underlying network structures and relationships (Serrat and Serrat, 2017) that are essential in understanding the dissemination patterns of fake news. Given the focus on understanding how information is disseminated and the role of user interactions, SNA emerged as the most suitable analytical framework for unraveling the complexities of fake news distribution in the South Korean 2020 election context.

Content analysis

In addition to examining the structure of fake news dissemination, we investigated the content of fake news stories. Natural language processing was employed to identify frequently mentioned words and pairs of closely linked words. Semantic network analysis further revealed semantic relationships between words. Automated content analysis enabled the identification of linguistic features characteristic of fake news stories.

Data collection

Legislative elections were held in South Korea on 15 April 2020. The year 2020 had witnessed a surge in pandemic-driven misinformation and the impending election also heightened the spread of fake news throughout the country (Jang et al., 2023). Participants ranging from political groups to individuals and potentially external entities were alleged to have manipulated public opinion and influenced voter behavior through the dissemination of false or misleading information (Choi, 2020; Ko, 2020). In order to comprehensively analyze this phenomenon, we collected YouTube data between 17 April and 10 August 2020 for the electoral fraud case using YouTube API. We used keywords or terms such as “electoral fraud,” “Election Commission,” and “early voting manipulation” to search relevant YouTube videoclips. After this data collection via API, the researchers manually filtered out videoclips that were not relevant to each case. We also gathered information about the YouTube channel that uploaded those videoclips. We transformed the videoclips into text using Google’s Speech-to-text (STT) service.

For SNA, we built models of two distinct networks: first, a channel network; and second, a videoclip network. In the channel network, nodes represent individual YouTube channels, while edges indicate interactions between channels. These interactions include elements such as subscriptions, mentions, or replies from one channel to another. This network model offers insights into the relationships among channels, highlighting which ones are more central or influential within the network. Similarly, the videoclip network comprises nodes representing individual videoclips, with edges denoting interactions between videoclips. These interactions were established when the same users replied to multiple videoclips, indicating a connection between them. The links were directed based on the temporal sequence of replies, providing further granularity in understanding the flow of interactions. For instance, if user A replied to videoclips X and Y, then the link was formed between X and Y. Each link is valued given the number of users shared between two videoclips. The links are directed by considering the time when replies were created. If user A replied to X first and then replied to Y afterwards, we formed X Y. Based on these networks formed, we calculated network density, geodesic path lengths, degree centrality, and other network indices.[1]

Table 1. Overall network statistics of electoral fraud case.

	Videoclip network	Channel network
Node	793	213
Centralization	0.067	0.044
Out-centralization	0.054	0.032
In-centralization	0.038	0.042
Density	1.98	23.74
Minimum	0	0
Maximum	479	22,318
Node attribute		
Fake news	675	159
Real news	118	54

Additional statistical analyses will be conducted to compare the differences of these indices in terms of fake/real and author/message/social factors.

Results

Overall network statistics

Table 1 summarizes the network statistics of videoclip network and channel network regarding electoral fraud case. Videoclip network is composed of 793 nodes, which also indicates that the number of relevant videoclips is in fact 793. On the level of dyadic relationships, the number of users who moved from one videoclip to the other ranged from 0 to 479. The value of network density was 1.98, which means that the average number of users who moved from one videoclip to another was 1.98. The network centralization value was 0.054 for out-degree (arrows heading out from the node in the network diagram) and 0.038 for in-degree (arrows heading into the node in the network diagram). These values closer to zero suggest that users spread out to the overall network, rather than being concentrated on a few videoclips. Of the above-stated 793 videoclips, 675 were classified as fake news.

Channel network consists of 213 nodes, which also indicates that the number of relevant channels is 213. On the level of dyadic relationships, the number of users who moved from one channel to the other ranged from 0 to 22,318. The value of network density was 23.74, which means that the average number of users who moved from one channel to another was 23.74. The network centralization value was 0.032 for out-degree and 0.042 for in-degree. Among 213 channels, 159 were classified as fake news.

Fake news dissemination

While Table 1 exhibits the characteristics of overall network, Table 2 shows the characteristics of fake news network and real news network, respectively. Regarding real news, videoclip network is composed of 118 nodes, which also means that the number of relevant videoclips is 118. On the level of dyadic relationships the number of users who moved from one videoclip to the other ranged from 0 to 387. The value of network density was 2.98, which means that the average number of users who moved from one videoclip to another was 2.98. The network centralization value was 0.041 for out-degree and 0.029 for in-degree. These values closer to zero suggest that users spread out to the overall network, rather than being concentrated on a few videoclips.

Table 2. Videoclip network statistics of electoral fraud case.

	Real news	Fake news
Node	118	675
Centralization	0.055	0.058
Out-centralization	0.041	0.056
In-centralization	0.029	0.039
Density	2.98	2.19
Minimum	0	0
Maximum	387	479

Regarding fake news, videoclip network consists of 675 nodes, which also indicates that the number of relevant videoclips is 675. On the level of dyadic relationships, the number of users who moved from one videoclip to the other ranged from 0 to 479. The value of network density was 2.19, meaning that the average number of users who moved from one videoclip to another amounted to 2.19. The network centralization value was 0.056 for out-degree and 0.039 for in-degree. These values closer to zero suggest that users spread out to the overall network, rather than being concentrated on a few videoclips.

Overall, the videoclip network of real news had relatively higher density than that of fake news, but we cannot rule out the possibility that this difference might have stemmed from the tendency that the former has a smaller number of nodes than the latter—it is highly likely that the network with fewer nodes tends to have greater value of density than the network with more nodes. The difference in terms of network centralization between the two networks was negligible.

Table 3 summarizes the network statistics of a channel network of real news and fake news, respectively. In our dataset, none of the channels had *both* real news videoclips and fake news videoclips. Regarding real news, channel network is composed of 54 nodes, which also indicates that the number of relevant channels is 54. On the level of dyadic relationships, the number of users who moved from one channel to the other ranged from 0 to 1621. The value of network density was 12.09, which means that the average number of users who moved from one channel to another was 12.09. The network centralization value was 0.084 for out-degree and 0.064 for in-degree. These values closer to zero suggest that users spread out to the overall network, rather than being concentrated on a few channels.

Regarding fake news, channel network is composed of 159 nodes, which also indicates that the number of relevant channels is 159. On the level of dyadic relationships, the number of users who moved from one channel to the other ranged from 0 to 22,318. The value of network density was 33.16, which means that the average number of users who moved from one channel to another was 33.16. The network centralization value was 0.039 for out-degree and 0.048 for in-degree. These values closer to zero suggest that users spread out to the overall network, instead of being concentrated on a few channels.

Overall, the channel network of real news had relatively lower density than that of fake news, even though the former had fewer nodes. (For reference, it is highly likely that the network with a smaller number of nodes tends to have greater value of density than the network with a larger number of nodes.) This finding implies that the channel network of fake news is denser and more active than that of real news. The difference in terms of network centralization between the two networks was negligible.

Network visualization

We visualized networks that were discussed in the previous sections. Network visualization plays a critical role in enhancing the accessibility and interpretability of complex network data. They

Table 3. Channel network statistics of electoral fraud case.

	Real news	Fake news
Node	54	159
Centralization	0.094	0.052
Out-centralization	0.084	0.039
In-centralization	0.064	0.048
Density	12.09	33.16
Minimum	0	0
Maximum	1,621	22,318

provide a condensed overview of the complex relationships and offer a visual narrative that complements the textual analysis, providing readers with a more comprehensive understanding of the data (Unwin, 2020). In order to provide an overview of the network structures and characteristics elucidated in our research, we provided the truncated version of those whole networks. The truncated version allows us to identify the links better visually among channels or videoclips. In the truncated version, we visualized nodes whose degree centrality is above the average or ranks within the top 25%. If the number of nodes amounted to more than 100 even after we applied the aforementioned rule, then we visualized nodes whose degree centrality ranks within the top 10%.

Channel network of real news

In the case of channel networks (see Figures 1 to 3) the higher the value of degree centrality, the darker the color of the node.

Videoclip network

In the case of the videoclip network (see Figures 4 to 6), the size of each node is proportional to its degree centrality—the higher the node's degree centrality is, the larger in size it will be. The nodes' colors are different based on the channels they belong to. Channels having relatively few videoclips are all colored gray. We visualized the label of each node by extracting one or two words from the title of each videoclip. Note that we were not able to visualize the whole title of each videoclip because it was too long and difficult to visualize in a readable manner.

Geodesic distances

We calculated the geodesic distances (the length of the shortest path between two nodes) of videoclip networks. When we were not able to calculate the geodesic distance between two nodes because they were not linked to each other, then we added distance 1 to the largest value of the geodesic distance identified within the network. The value 1 of geodesic distance represents the direct connection between two videoclips. As shown in Table 4, the values of geodesic distances between two nodes are mostly 1 or 2 in the videoclip network of real news. In contrast, regarding the videoclip network of fake news, a considerable number of dyadic distances had the geodesic distance of 3. Moreover, the videoclip network of fake news had a larger maximum value of the geodesic distance than that of real news.

As shown in Table 5, we statistically compared the difference of videoclip networks between real news and fake news. We normalized in-degree centrality and out-degree centrality, taking into

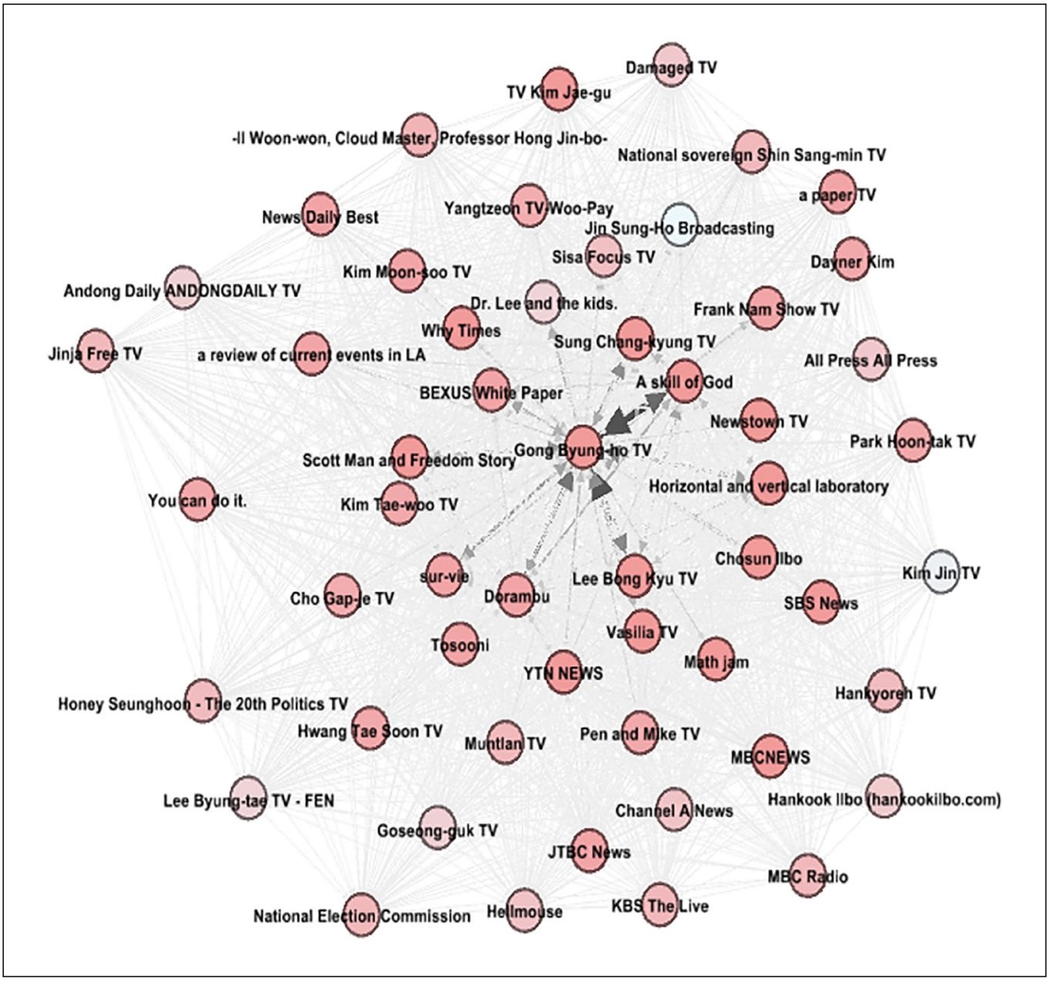


Figure 1. Overall channel network: Visualized nodes whose degree centrality ranks within the top 25%.

account the number of nodes involved in each network. Implementing Mann–Whitney U test, we found statistically significant differences in terms of these variables. The videoclip network of fake news tends to have smaller values when compared with real news. Taken together, these results suggest that the spread of fake news videoclips has different network characteristics from that of real news videoclips.

Author effect, message effect, and social effect

We compared the differences between real news videoclips and fake news videoclips in terms of author effect (i.e. YouTube channels), message effect (i.e. the content of videoclips uploaded by YouTube channels), and social effects (i.e. online social indicators such as the numbers of comments, likes, dislikes, and view that each videoclip have garnered; see Table 6). Regarding social effects, we found statistically significant differences between real news videoclips and fake news videoclips in terms of the numbers of comments, likes, and dislikes. Specifically, fake news

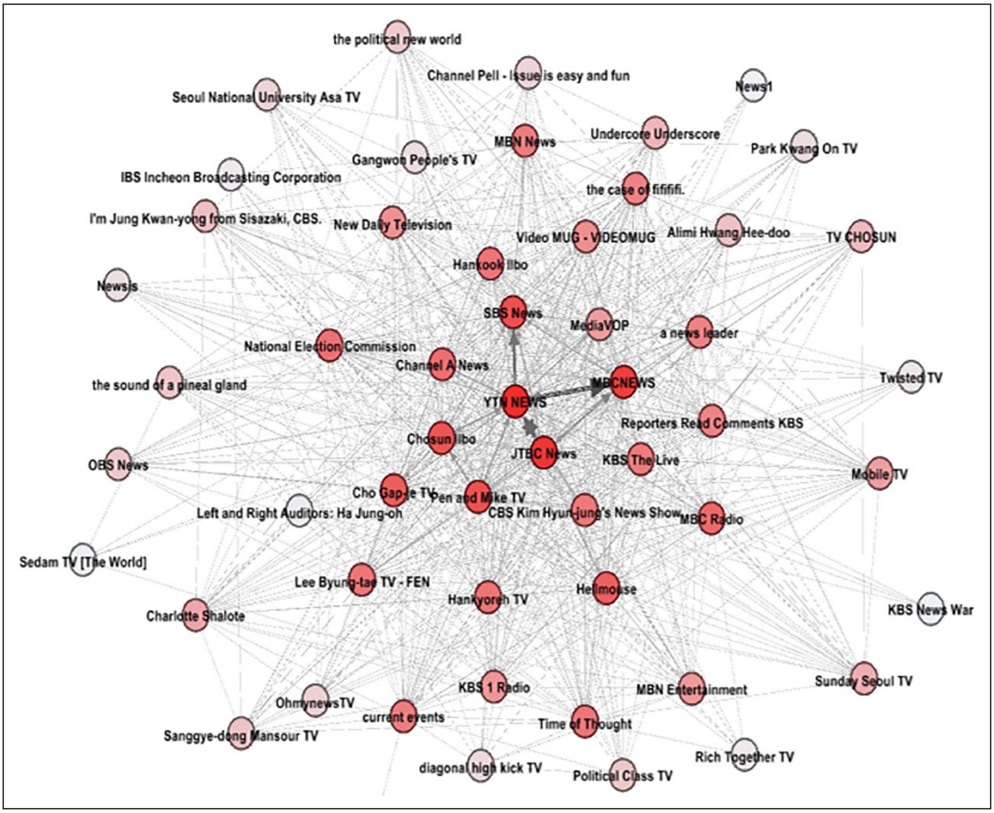


Figure 2. Channel network of real news.

videoclips had the larger number of likes than real news videoclips, whereas real news videoclips had the larger number of dislikes than fake news videoclips. These findings strongly suggest that fake news videoclips are more welcomed than real news videoclips on YouTube.

Regarding message effects, we analyzed the content of videoclips of real news and fake news. We found that the length of videoclips was statistically significantly different between the two—fake news videoclips were longer by 10 minutes and 37 seconds in average than real news videoclips. The number of main speakers in the videoclips was also markedly different between real news and fake news. Real news videoclips had an average of 0.48 more main speakers than the fake news videoclips. Furthermore, the formats of videoclips we categorized as one-sided delivery, conversational format, and field-footage insert greatly differed between the two. In the case of fake news, the observed number of videoclips that used conversational formats was smaller than statistically expected, whereas the opposite situation was found for real news. We also checked whether the videoclips uploaded and released during the period of data collection had been made private (i.e. not open to the public) afterwards (this was checked on 12 January 2021). We discovered that the number of videoclips eliminated or made private was greater in the case of fake news. Regarding the latter, the observed number of videoclips eliminated or made private was even greater than statistically expected.

Referring to author effects, we focused on the channels that created and uploaded videoclips about the electoral fraud. The number of subscribers each channel has was found to be a

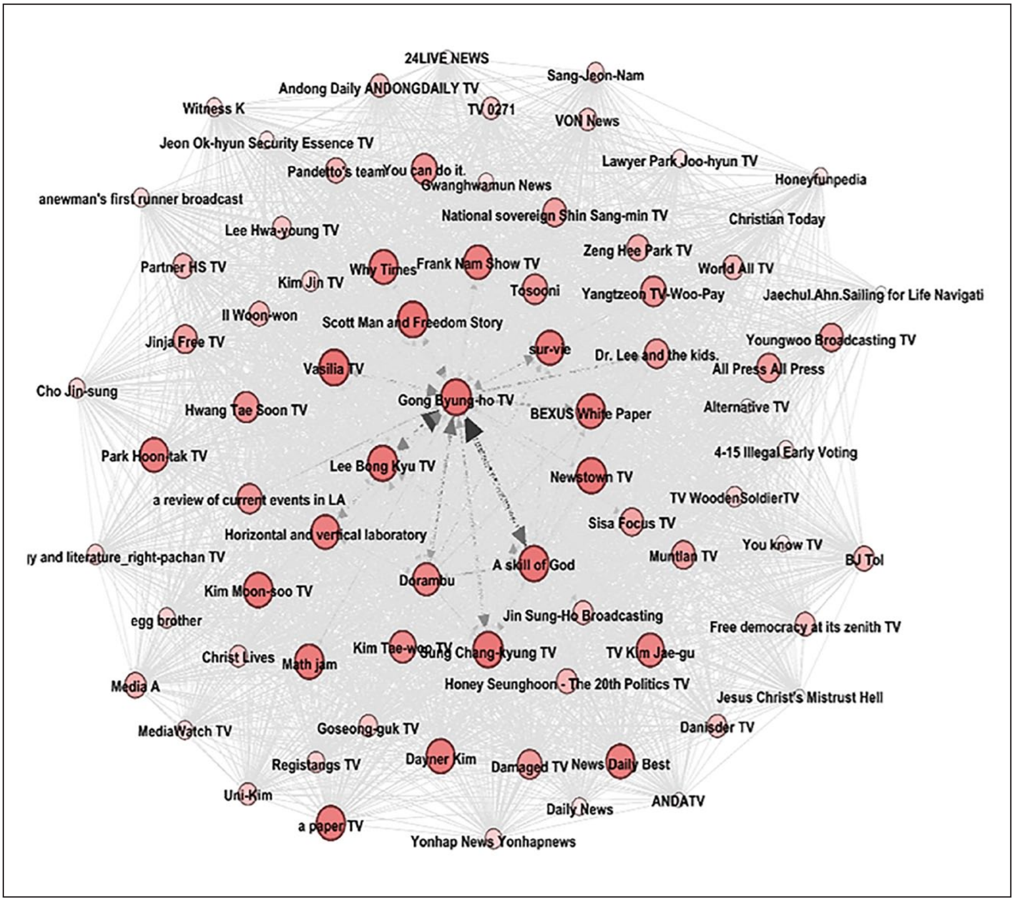


Figure 3. Channel network of fake news: Visualized nodes whose degree centrality is above the average.

statistically significant factor that differentiates channels that uploaded fake news videoclips from channels that uploaded real news videoclips. The latter had on average 203 more subscribers compared with the former. As well, we found that far larger channels were news organizations in the case of real news. However, we did not find any statistical significance regarding the variable whether the labels of channels are based on real names or not.

Fake news messaging

We implemented natural language processing using the texts extracted from videoclips and extracted only nouns from these texts. Then, we formed semantic networks composed of those nouns. The links between nouns were made when the two nouns appeared in the same videoclip. Based on the semantic network of real news and fake news, respectively, we calculated the degree centrality of each noun included in the network. Table 7 shows the top 50 words (i.e. nouns) that had higher degree centrality than others. Unlike the scenario involving real news, the terms such as “the Republic of Korea,” “Moon Jae In” (i.e. the President of South Korea), “evidence,” “freedom,” “QR code,” “the United States,” and “China” were frequently mentioned.

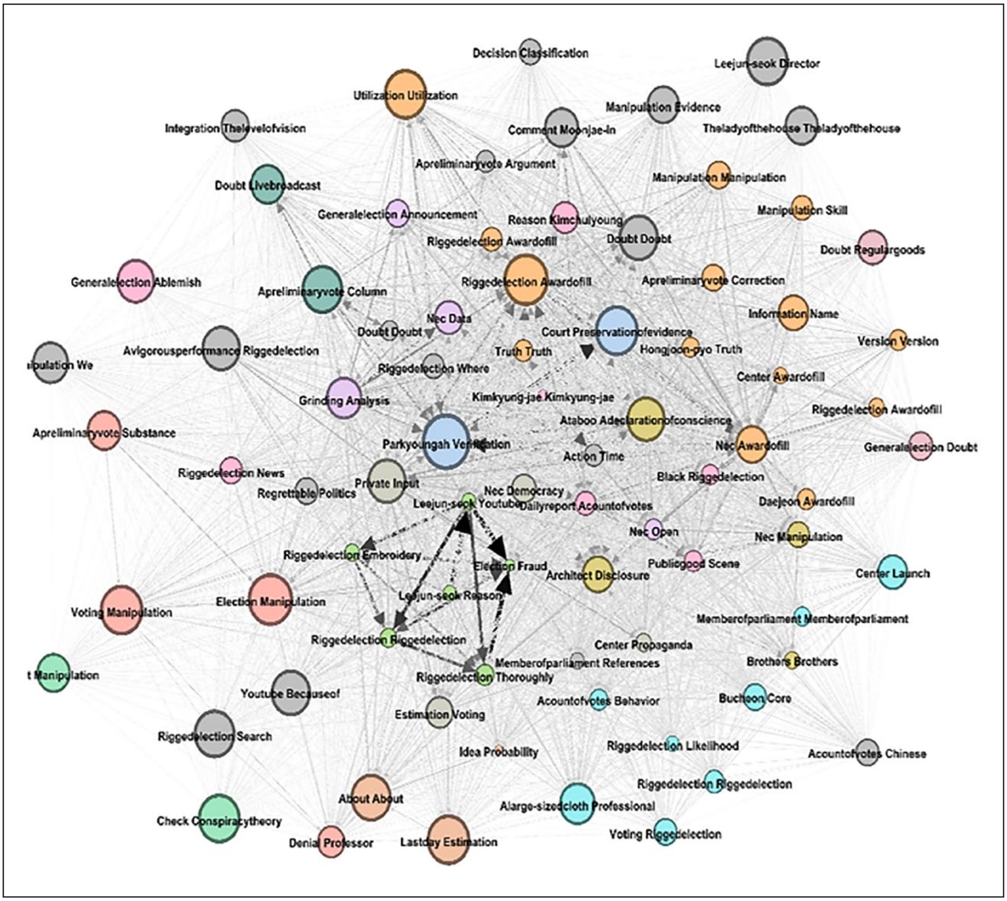


Figure 4. Overall videoclips network: Visualized nodes whose degree centrality ranks within top 10%.

We also implemented sentiment analysis with the texts extracted from real and fake news videoclips. For this analysis, we used KOSAC (Korean Sentiment Analysis Corpus) dictionary (Shin et al., 2012) and tagged positive or negative sentiments on each part-of-speech (POS) that the texts contain. We subtracted the number of negative POS from the number of positive POS and then divided these subtracted values with the number of total POS that each text contains (to eliminate the effect of text length). When the values generated from this procedure were greater than zero, we classified the text of the videoclip as having positive sentiment on the electoral fraud. For this analysis, we excluded 31 texts from our dataset, since they had the same number of positive and negative POS or they were not retrievable due to the elimination of videoclips at the time of this analysis. As shown in Table 8, the percentage of fake news videoclip texts categorized as having negative sentiment is 53.6%, whereas for real news videoclip texts it is 54.8%. We were not able to detect any statistical significance regarding the sentiment difference between real and fake news videoclip texts (see Table 8).

Finally, we conducted CONCOR (CONvergence of iterated CORrelations) analyses on the semantic networks of real and fake news videoclip texts. CONCOR allows us to cluster words

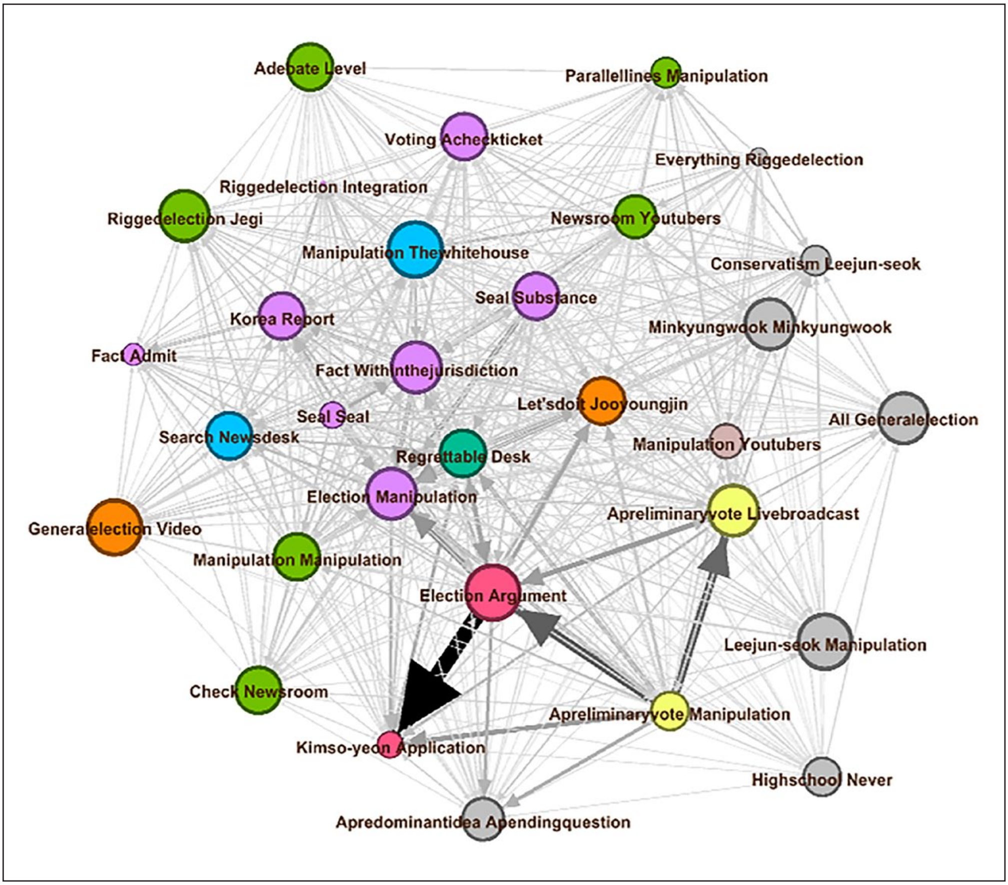


Figure 5. Videoclip network of real news: Visualized nodes whose degree centrality ranks within the top 25%.

based on the structural equivalence that words have in the semantic network and enables us to point out the implications of the network's thematic structure. For visualization purposes, we only included the top 200 words in terms of word occurrence frequency in the CONCOR analyses. The sizes of the nodes are proportional to their degree centrality. Figure 7 shows the CONCOR result of the semantic network of real news videoclip texts, while Figure 8 shows CONCOR result of the semantic network of fake news videoclip texts.

As shown in Figure 7, real news videoclips discussed issues such as the correction of the early-voting turnout rate (e.g. voter, suspicion, pre-voting manipulation), the correction of electoral fraud allegation (e.g. classifier, program, equipment), key factors related to electoral fraud issue (e.g. preliminary vote, manipulation, ballot paper), and political reactions to electoral fraud (e.g. Democratic Party of Korea, United Future Party, politics).

As shown in Figure 8, fake news videoclips addressed criticism of current Korean society (e.g. prosecution, Korea, worry), claims to preserved evidence of electoral fraud (e.g. equipment, method, processing), doubts raised about the voting classifier (e.g. QR code, server, storage), and key factors related to the electoral fraud issue (e.g. preliminary vote, election, ballot paper).

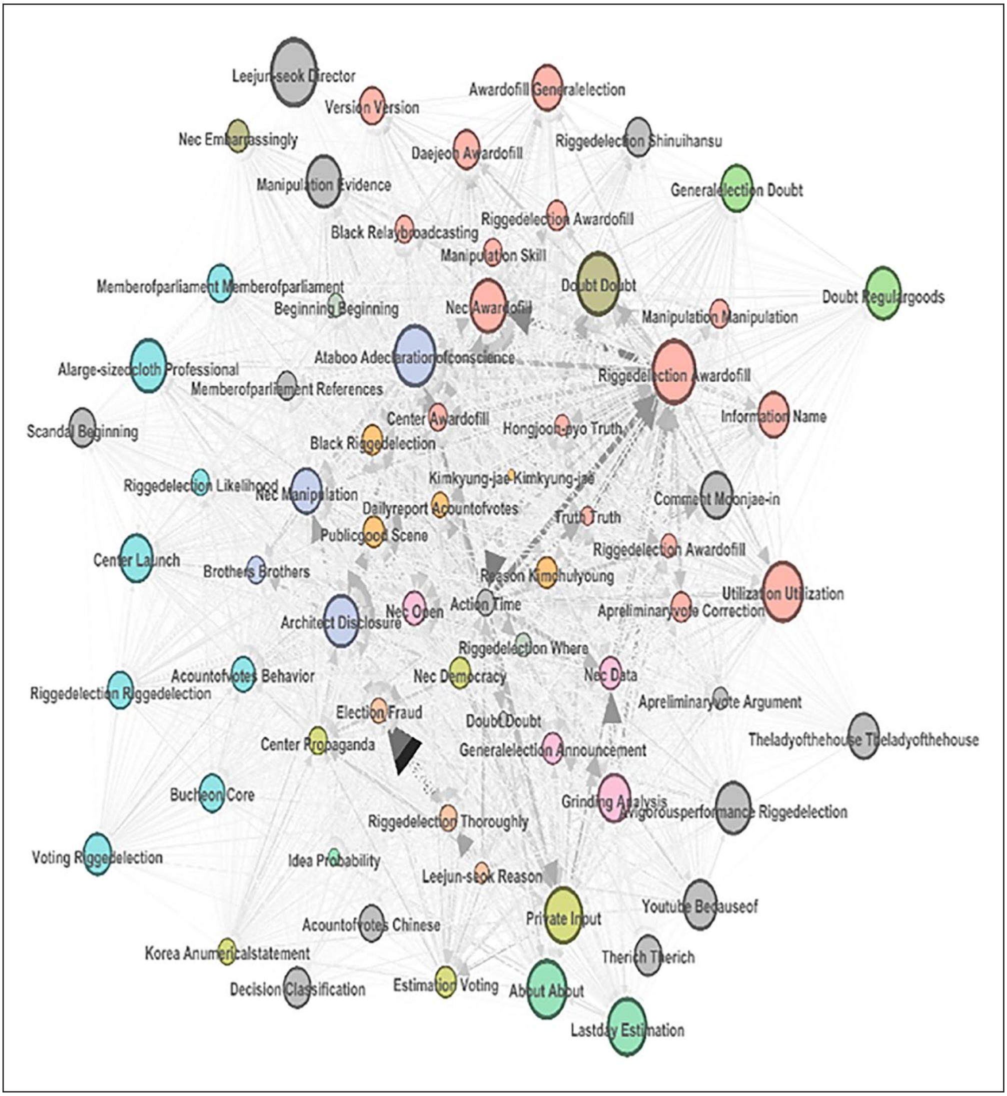


Figure 6. Videoclip network of fake news: Visualized nodes whose degree centrality ranks within the top 10%.

Discussion

Practical implications

To explain the characteristics of fake news propagation, this study utilized SNA and content analysis. The results provide important insights into the structure and characteristics of these networks, which can be useful in curbing the spread of fake news and dissemination of factual data. The study's contributions are particularly significant, when placed in the context of the 2020 South Korean election. Because of the increased interest in the election at that time, it has unfortunately

Table 4. Geodesic distances of videoclip network.

Real news			Fake news		
	Frequency	Proportion		Frequency	Proportion
1–1	5088	36.9	1–1	120,977	26.6
2–2	6845	49.6	2–2	218,086	47.9
3–3	921	6.7	3–3	79,132	17.4
4–4	20	0.1	4–4	12,520	2.8
5–5	932	6.8	5–5	923	0.2
			6–6	25	0
			7–7	23,287	5.1

Table 5. Mann–Whitney U test: results comparing videoclip networks between real news and fake news.

	Mean-rank diff. (Fake − Real)	Statistic	*p* Value
Out-degree	–89.86	30,799	<0.001
In-degree	–83.65	31,423	<0.001

Table 6. Mann–Whitney U and chi-square test results comparing real and fake news videoclips.

	Mean-rank difference (Fake − Real)	Statistic (*df*)	*p* Value
Social effect (online social indicators)			
Comment count	–98.92	29,151	<0.001
Dislike count	–165.16	21,985	<0.001
Like count	135.62	51,661	<0.001
View count	–9.59	38,862	0.68
Message effect (videoclip content)			
Length (in seconds)	93.47	49,213	<0.001
Number of main speakers	–111.65	28,611	<0.001
Information delivery format (one-sided, conversation, footage insert)[a]	–	23.16 (2)	<0.001
Video eliminated/made private or not[a]	–	7.18 (2)	0.007
Author effect (channel characteristic)			
Subscriber count	–28.49	3,145	0.003
News organizations or not[a]	–	52.12 (1)	<0.001
Label of channel based on real name or not[a]	–	0.84 (1)	0.36

[a]Chi-squared test.

Table 6-1. Information delivery format.

	One-side	Interaction	Scene
Real new[a]	72 (85)	33 (16)	11 (14)
Fake news[a]	461 (448)	70 (87)	79 (76)

[a]Observed value (expected value).

Table 6-2. Video eliminated/made private or not.

	No	Yes
Real news[a]	116 (108)	2 (9.97)
Fake news[a]	610 (618)	65 (57)

[a]Observed value (expected value).

Table 6-3. News organization or not.

	No	Yes
Real news[a]	27 (45)	27 (9)
Fake news[a]	152 (134)	7 (25)

[a]Observed value (expected value).

Table 6-4. Label of channel based on real name or not.

	No	Yes
Real news[a]	45 (42)	9 (12)
Fake news[a]	121 (124)	38 (35)

[a]Observed value (expected value).

become a prime target for the dissemination of disinformation and fake news. To evaluate the effects and potential implications of misinformation campaigns, it is essential to gain insight into the features of the networks involved in spreading fake news in future elections. Our study points to several significant practical implications that are of paramount importance in the battle against misinformation, particularly in the context of electoral fraud cases.

First, the analysis of network statistics highlights the critical need for strategic targeting of high-influence nodes in the dissemination process. Nodes with higher out-degree centrality in the videoclip network played a pivotal role in spreading information. This finding highlights the importance of identifying and focusing efforts on these influential nodes for effective interventions, such as promoting fact-checking initiatives or implementing content moderation measures, to combat the spread of fake news (Budak, Agrawal, and El Abbadi, 2011; Pham et al., 2020; Zhen et al., 2023). According to Zhen et al. (2023), a viable strategy to address this issue may involve specifying influential users within a community who have a track record of disseminating false information. The authors suggest that it would be beneficial for social media platforms and governments to monitor and potentially label the messages of these accounts as potentially misleading, or even censor them outright. Consequently, this targeted approach may prove considerably more precise and efficient compared with indiscriminate interventions (Zhen et al., 2023).

The identification of a substantial volume of misleading content emphasizes the urgency for a robust fact-checking infrastructure (Gradoń et al., 2021). The denser and more active channel network of fake news, coupled with the higher prevalence of direct links, signifies a concerted effort in disseminating false information. This highlights not only the urgency in implementing technical measures (e.g. extended algorithm) to counteract the spread of misinformation (Kumar and Geethakumari, 2014; Pham et al., 2020) but also the need for honest messaging to debunk the spread of misinformation (Schnackenberg and Tomlinson, 2016).

Table 7. Top 50 words (sorted in descending order in terms of degree centrality).

	Real news	Fake news
1	Voting	People
2	People	Election
3	Election	Voting
4	Preliminary vote	Fraud election
5	A ballot paper	A ballot paper
6	Manipulation	Preliminary vote
7	Problem	The public
8	Democratic Party of Korea	Problem
9	Check	Central election commission
10	Result	Thanks
11	Representative	Manipulation
12	Central election commission	Republic of Korea
13	The public	Check
14	YouTube	Situation
15	Explanation	Representative
16	Situation	Relationship
17	Count of votes	Time
18	Conservatism	Moon Jae-in
19	Candidate	Broadcasting
20	Denial	Result
21	Jegi	Area
22	General election	Democratic Party of Korea
23	Process	President
24	Officer of the crown	Denial
25	Fraud election	Evidence
26	Area	General election
27	Doubt	Count of votes
28	Broadcasting	Freedom
29	Member of parliament	Beginning
30	President	Explanation
31	Integration	QR code
32	Politics	Count
33	United Future Party	Officer of the crown
34	Count	Candidate
35	Time	Doubt
36	Relationship	The United States
37	Editor	China
38	Seoul	Use
39	Position	Video
40	Whole	World
41	Jurisdiction	Jegi
42	Beginning	Photo
43	Survey	Reason
44	Reason	Member of parliament
45	Refined sugar	Information
46	System	Process

(Continued)

Table 7. (Continued)

	Real news	Fake news
47	Private	YouTube
48	Myself	Election commission
49	Thanks	Citizen
50	Photo	Country

Words highlighted in yellow indicate words that do not overlap between real and fake news videoclip texts.

Table 8. Chi-square test.

	χ^2 (*df*)	*p* Value
Neg/Pos news	0.016 (1)	0.90

Table 8-1. Neg/Pos video.

	Negative	Positive
Real news	63 (62)	52 (53)
Fake news	347 (348)	300 (299)

[a]Observed value (expected value).

The abundance of direct links in the network indicates that video snippets were widely used to disseminate news across various media categories. However, the higher prevalence of fake news videoclips in this group suggests active and widespread disinformation dissemination. Consequently, establishing dedicated fact-checking teams within media organizations is critical in swiftly identifying and debunking false information. This proactive approach ensures a more accurate public discourse and prevents the rapid spread of misinformation. In addition, the identification of statistically significant differences in comments, likes, and dislikes between actual and false news videos emphasizes the need for robust fact-checking infrastructure to ensure accurate information dissemination (Li and Chang, 2023).

Textual analysis of the fake news videoclip reveals words or terms such as “the Republic of Korea,” ‘Moon Jae In’ (i.e. the President of South Korea), “evidence,” “freedom,” “QR code,” “the United States,” and “China” were frequently mentioned. Fake news videoclips addressed criticism of the current Korean society (e.g. prosecution, Korea, worry), claims to preserve evidence of electoral fraud (e.g. equipment, method, processing), doubts raised about the voting classifier (e.g. QR code, server, storage), and key factors related to electoral fraud issue (e.g. preliminary vote, election, ballot paper). This underscores the importance of online media literacy programs that equip individuals with the skills to critically evaluate content and discern between credible information and misleading narratives (Lee and Ramazan, 2021).

We further compared real news and fake news videoclips in terms of author effect (i.e. YouTube channels), message effect (i.e. the content of videoclips uploaded by YouTube channels), and social effects (i.e. online social indicators like comments, likes, dislikes, and views; see Table 6). We identified statistically significant differences in comments, likes, and dislikes between actual and false news videos. Fake news videos received more likes while true news videos had more dislikes.

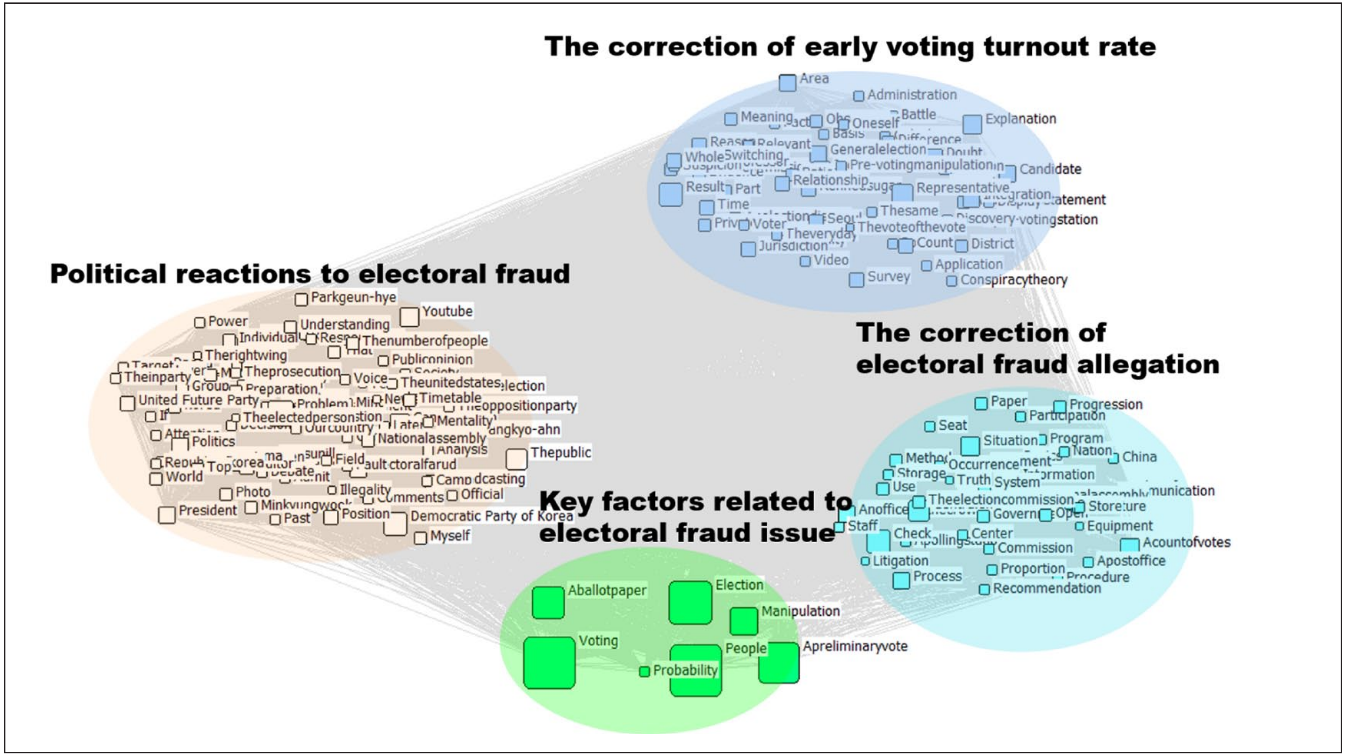


Figure 7. CONCOR result of the semantic network of real news videoclip texts.

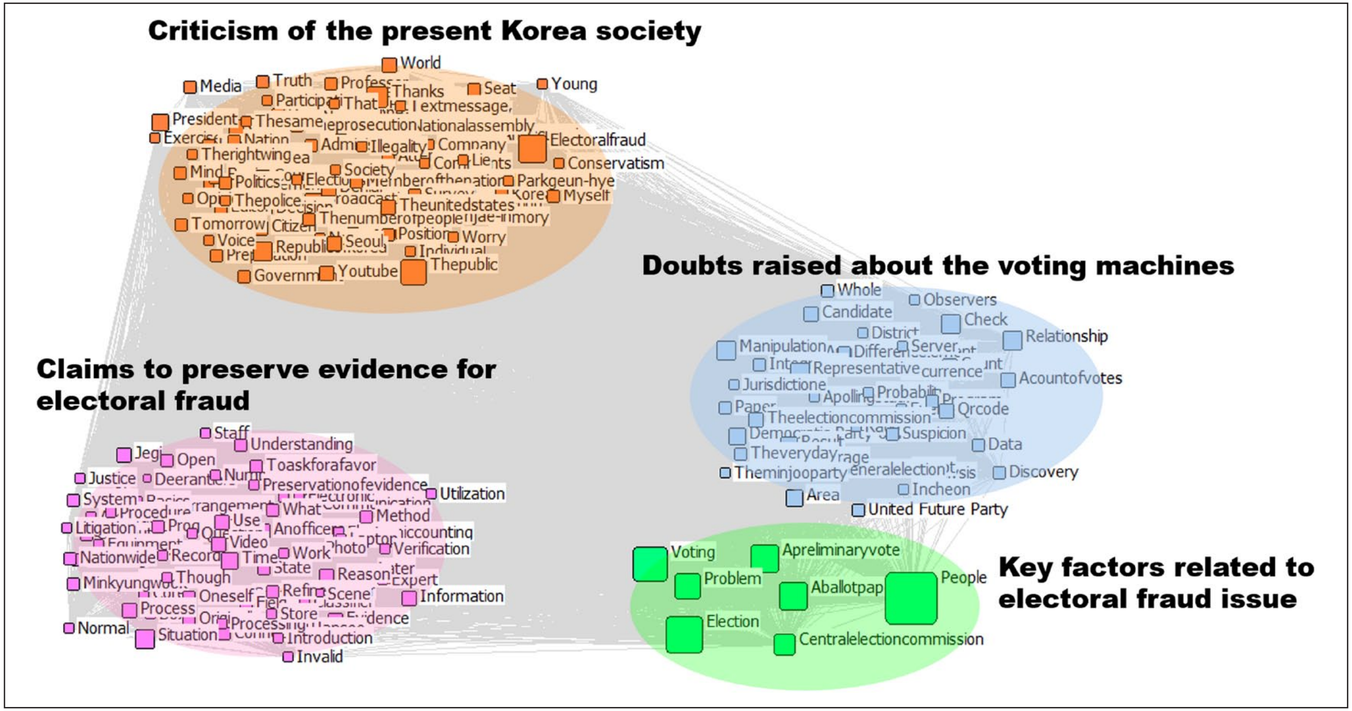


Figure 8. CONCOR result of the semantic network of fake news videoclip texts.

In terms of message impact, fake news videoclips averaged 10 minutes and 37 seconds longer than true news clips.

We also examined electoral fraud video channels for author effects. The number of subscribers was statistically significant in differentiating fake news channels from legitimate news channels. This finding (see Table 6) suggests that the number of subscribers a channel has can act as an important clue in distinguishing between fake and real news channels. Channels with a larger following are more likely to be linked with reliable and trustworthy news sources, while those with fewer subscribers may have a higher tendency to spread false information. Therefore, platforms and authorities should consider subscriber counts as one of the factors when assessing the credibility of social media channels. Channels with a larger audience may be given priority in monitoring and fact-checking efforts, as they are likely to have a wider impact on spreading information. To counteract the dissemination of fake news, it is important for authorities to consider engagement metrics, content characteristics, and channel attributes. This could involve targeted interventions aimed at addressing the specific features of fake news content and channels, along with strategies to promote media literacy and advanced detection methods (Zhang and Ghorbani, 2020). In addition, platforms and content creators should be vigilant in monitoring and addressing the spread of misinformation, taking into account these specific characteristics.

The observation that fake news networks exhibited higher density and active dissemination is a cause for concern. The result indicates that there was a concerted effort to spread fake news through videoclips and channels, aiming to manipulate public perceptions and shape election outcomes. The higher density indicates that there are more individuals involved in the propagation of fake news, which could amplify its reach and influence on public discourse. By identifying the linguistic elements, frequently used words, and content patterns prevalent in fake news articles, this study unpacks the strategies employed by those spreading misinformation during the election. This critical analysis of the content helps us to understand how false narratives were constructed and spread. However, it is important to note that the study does not delve into the specific motivations or intentions of the individuals or groups behind the dissemination of fake news.

Theoretical implications

ANT is a sociological framework that seeks to understand social phenomena by examining the relationships and interactions between both human and nonhuman actors (entities that have agency and can influence outcomes) work together to shape reality (Nawararthne and Storni, 2023). ANT is particularly interested in how these actors come together to form networks, how they negotiate their interests, and how they shape the course of events. One significant criticism of ANT pertains to its assertion of equal agency among all actors in a network (Hadden and Jasny, 2019; Whittle and Spicer, 2008). Despite this, some scholars have defended this assumption of equal agency (Law, 1992). Through SNA, our research provides empirical evidence that supports the notion that certain actors wield significant influence in the dissemination of information within a social network. The identification of statistically significant differences in the in-degree and out-degree centrality between real news and fake news networks contributes to ANT's understanding of power dynamics and influence within actor networks. The lower values observed in the videoclip network of fake news suggest that certain actors connected to fake news may have limited influence compared with those involved in real news dissemination. This highlights the complexities of power relationships and the ways in which different actors mobilize resources and become visible within the network. Recognizing these influential actors allows individuals and organizations to effectively control the rapid spread of information and, in turn, address any barriers to the flow of information (Kolli and Khajeheian, 2020).

The findings also confirm that ANT is a suitable framework to examine the spread of fake news within the context of electoral processes and affirms the role of technological platforms and algorithms in shaping the characteristics of actor networks (Weikmann and Lecheler, 2023). The interconnectedness and dissemination patterns observed in the videoclip network indicate the role of nonhuman actors, such as social media platforms (YouTube) or ineptness of recommendation algorithms, in countering the spread of fake news. This aligns with ANT's perspective on the agency of nonhuman actors and the reciprocal relationship between humans and technologies in network formation (Latour, 2007).

This study addresses a critical gap in existing ANT literature by focusing on the contemporary issue of fake news propagation within the context of electoral processes. While earlier ANT studies have shown how a claim can be construed as fact or fiction through networks of human and nonhuman interactions (Pantumsinchai, 2018), or how fake videos propagate in a fact-checking network (Weikmann and Lecheler, 2023), or the agenda-setting power of fake news in the political landscape (Vargo et al., 2018), to the best of the authors' knowledge, this is one of the first studies that investigated the propagation of real and fake news in the South Korean election through the lens of ANT. It enriches the existing body of literature by showcasing the continued relevance and adaptability of ANT as a theoretical framework in understanding complex socio-technical systems.

Limitations and future research

Our analysis of network characteristics reveals the unique characteristics of the videoclips and channel networks in South Korea, emphasizing the dynamic transmission of fake news and its distinctive network patterns. Electoral processes can be better protected from the negative effects of false news by taking advantage of the knowledge gained from a thorough understanding of these network properties. However, it is important to note that our findings are limited in terms of the depth of exploration of actor–network dynamics. The analysis primarily focuses on network characteristics and does not exclusively delve into the controversies and power dynamics (Whittle and Spicer, 2008) and existing discoursers within the network (Pantumsinchai, 2018). Research in the future could expand on these dimensions to provide a more comprehensive contribution to ANT's theoretical framework.

Also, in accordance with the inherent limitations of ANT as an analytical lens (Whittle and Spicer, 2008), this study primarily focused on network characteristics and content evaluation without explicitly examining the impact of fake news on the electoral processes. Thus, the study does not explore the extent to which this fake news influenced voter behavior or electoral outcomes. Hence, we need more research and real-world evidence like qualitative interviews or experiments, to understand how voters act when they encounter fake news.

Another key factor to consider is the role of social media platforms and their algorithms in aiding the spread of fake news during the South Korean election. Furthermore, we do not examine the role of platforms like YouTube in facilitating disinformation distribution or the efficiency of content moderation procedures in limiting its spread (Ibrahim et al., 2023). Future research should examine the systemic variables that contribute to the spread of fake news, such as platform policies, algorithms, and user engagement dynamics.

Finally, given the rapid changes occurring in media and technology, future research might look at how emerging systems like artificial intelligence (AI)-generated content contribute to the creation and propagation of fake news (Gutiérrez, 2023). It is critical to understand the potential risks and obstacles produced by these changes in order to design successful tactics to combat fake news in the future. Despite the fact that the findings provide insights regarding the characteristics of real and fake news networks and the common patterns in their content, a thorough analysis is still necessary. Future research should continue to conduct network analysis across various political and social situations to determine, first, whether the network characteristics and fake news dissemination patterns seen in the 2020 South Korean election are consistent; or second, if there are variations due to cultural, political, or social realities or circumstances.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Korea Foundation, Grant # 1024000-2769, Republic of Korea.

ORCID iD

Md Irfanuzzaman Khan https://orcid.org/0000-0003-2671-7053

Note

1. For those who are interested in the data and command files employed in this study, please refer to the dataset repository at the University of Canberra, which can be found at Sheehy et al. (2023). You will find a folder called "Fake News and Real News" which includes separate subfolders for all tables and figures used in this study.

References

Albright J (2017) Welcome to the era of fake news. *Media and Communication* 5(2): 87–89.

Allcott H and Gentzkow M (2017) Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2): 211–236.

Altuntas F, Altuntas S and Dereli T (2022) Social network analysis of tourism data: a case study of quarantine decisions in COVID-19 pandemic. *International Journal of Information Management Data Insights* 2(2): 100108.

Ansar W and Goswami S (2021) Combating the menace: a survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights* 1(2): 100052.

Aswani R, Kar AK and Ilavarasan PV (2019) Experience: managing misinformation in social media—insights for policymakers from Twitter analytics. *Journal of Data and Information Quality* 12(1): 1–18.

Azis Prasetyo R and Aisyah U (2018) Social media, radicalism, terrorism and threats for democracy process in public space. In: *Proceedings of the International Post-Graduate Conference on Media and Communication*, Surabaya, Indonesia, 13 November.

Bastick Z (2021) Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior* 116: 106633.

Bleakley P (2023) Panic, pizza and mainstreaming the alt-right: a social media analysis of Pizzagate and the rise of the QAnon conspiracy. *Current Sociology* 71(3): 509–525.

Budak C (2019) What happened? The spread of fake news publisher content during the 2016 U.S. presidential election. In: *WWW'19: The World Wide Web Conference*, San Francisco, CA, 13–17 May, pp.139–150. New York: Association for Computing Machinery.

Budak C, Agrawal D and El Abbadi A (2011) Limiting the spread of misinformation in social networks. In: *Proceedings of the 20th International Conference on World Wide Web*, Hyderabad, India, 28 March–1 April, pp.665–674. New York: Association for Computing Machinery.

Cantarella M, Fraccaroli N and Volpe R (2023) Does fake news affect voting behaviour? *Research Policy* 52(1): 104628.

Chen Y, Conroy N and Rubin V (2015) News in an online world: the need for an "automatic crap detector." *Proceedings of the Association for Information Science and Technology* 52(1): 1–4.

Choi D, Chun S, Oh H, et al. (2020) Rumor propagation is amplified by echo chambers in social media. *Scientific Reports* 10(1): 310.

Choi H (2020) Rumors and conspiracy theories hamper fight against COVID-19. *The Korea Herald*, 20 August. Available at: https://www.koreaherald.com/view.php?ud=20200820000670&ACE_SEARCH=1

Choi S (2014) Flow, diversity, form, and influence of political talk in social-media-based public forums. *Human Communication Research* 40(2): 209–237.

Choi S, Yang JS and Chen W (2018) Longitudinal change of an online political discussion forum: antecedents of discussion network size and evolution. *Journal of Computer-Mediated Communication* 23(5): 260–277.

Conroy NK, Rubin VL and Chen Y (2015) Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52(1): 1–4.

Diaz Ruiz C and Nilsson T (2023) Disinformation and echo chambers: how disinformation circulates on social media through identity-driven controversies. *Journal of Public Policy & Marketing* 42(1): 18–35.

Eom S-J, Hwang H and Kim JH (2018) Can social media increase government responsiveness? A case study of Seoul, Korea. *Government Information Quarterly* 35(1): 109–122.

Freiling I, Krause NM, Scheufele DA, et al. (2023) Believing and sharing misinformation, fact-checks, and accurate information on social media: the role of anxiety during COVID-19. *New Media and Society* 25(1): 141–162.

Go S-g and Lee M-r (2020) Analysis of fake news in the 2017 Korean presidential election. *Asian Journal for Public Opinion Research* 8(2): 105–125.

Gradoń KT, Hołyst JA, Moy WR, et al. (2021) Countering misinformation: a multidisciplinary approach. *Big Data & Society* 8(1): 20539517211013848.

Grover P, Kar AK and Dwivedi Y (2022) The evolution of social media influence: a literature review and research agenda. *International Journal of Information Management Data Insights* 2(2): 100116.

Gutiérrez JLM (2023) On actor-network theory and algorithms: ChatGPT and the new power relationships in the age of AI. *AI and Ethics*. Epub ahead of print 28 June. DOI: 10.1007/s43681-023-00314-4.

Hadden J and Jasny L (2019) The power of peers: how transnational advocacy networks shape NGO strategies on climate change. *British Journal of Political Science* 49(2): 637–659.

Ibrahim H, AlDahoul N, Lee S, et al. (2023) YouTube's recommendation algorithm is left-leaning in the United States. *PNAS Nexus* 2(8): pgad264.

Igwebuike EE and Chimuanya L (2021) Legitimating falsehood in social media: a discourse analysis of political fake news. *Discourse and Communication* 15(1): 42–58.

Jang SH, Jung KE and Yi YJ (2023) The power of fake news: Big Data analysis of discourse about COVID-19–related fake news in South Korea. *International Journal of Communication* 17: 27.

Kar AK and Aswani R (2021) How to differentiate propagators of information and misinformation– Insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences* 42(6): 1307–1335.

Kar AK, Tripathi SN, Malik N, et al. (2023) How does misinformation and capricious opinions impact the supply chain-A study on the impacts during the pandemic. *Annals of Operations Research* 327(2): 713–734.

Khan ML and Malik A (2022) Researching YouTube: methods, tools, and analytics. In: Quan-Haase A and Sloan L (eds) *The Sage Handbook of Social Media Research Methods*. New York: Sage, pp.651–663.

Kim D (2020) Before Trump, South Korean conservatives also claimed a "stolen" election. *The Diplomat*, 11 November. Available at: https://thediplomat.com/2020/11/before-trump-south-korean-conservatives-also-claimed-a-stolen-election/.

Ko J (2020) Police toughen stance on misinformation, obstruction of antivirus efforts. *The Korea Herald*, 23 August. Available at: https://www.koreaherald.com/view.php?ud=20200823000147&ACE_SEARCH=1

Kolli S and Khajeheian D (2020) How actors of social networks affect differently on the others? Addressing the critique of equal importance on actor-network theory by use of social network analysis. In: Williams (ed.) *Contemporary Applications of Actor Network Theory*. Singapore: Palgrave Macmillan, pp.211–230.

Kumar KK and Geethakumari G (2014) Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences* 4(1): 14.

Labafi S (2020) Iranian data protection policy in social media; an actor-network theory approach. In: Williams I (ed.) *Contemporary Applications of Actor Network Theory*. Singapore: Palgrave Macmillan, pp.121–139.

Latour B (2007) *Reassembling the Social: An Introduction to Actor-network-theory*. Oxford: Oxford University Press.

Law J (1992) Notes on the theory of the actor-network: ordering, strategy, and heterogeneity. *Systems Practice* 5: 379–393.

Lazer DM, Baum MA, Benkler Y, et al. (2018) The science of fake news. *Science* 359(6380): 1094–1096.

Lee DKL and Ramazan O (2021) Fact-checking of health information: the effect of media literacy, metacognition and health information exposure. *Journal of Health Communication* 26(7): 491–500.

Li J and Chang X (2023) Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media. *Information Systems Frontiers* 25(4): 1479–1493.

Nawararthne D and Storni C (2023) Black-boxing journalistic chains, an actor-network theory inquiry into journalistic truth. *Journalism Studies* 24(13): 1629–1650.

Newman N, Fletcher R, Kalogeropoulos A, et al. (2019) Digital news report 2019. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_0.pdf

Nolin J and Olson N (2016) The Internet of Things and convenience. *Internet Research* 26(2): 360–376.

Ognyanova K, Lazer D, Robertson RE, et al. (2020) *Misinformation in Action: Fake News Exposure is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is in Power*. Harvard Kennedy School Misinformation Review. Available at: https://misinforeview.hks.harvard.edu/article/misinformation-in-action-fake-news-exposure-is-linked-to-lower-trust-in-media-higher-trust-in-government-when-your-side-is-in-power/

Pantumsinchai P (2018) Armchair detectives and the social construction of falsehoods: an actor–network approach. *Information, Communication & Society* 21(5): 761–778.

Pariser E (2011) *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. New York: Penguin Books.

Park A and Youm K (2019) Fake news from legal perspective: the United States and South Korea compared. *Southwestern Journal of International Law* 25(1): 100–119.

Pham DV, Nguyen GL, Nguyen TN, et al. (2020) Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* 8: 78879–78889.

Rhodes SC (2022) Filter bubbles, echo chambers, and fake news: how social media conditions individuals to be less critical of political misinformation. *Political Communication* 39(1): 1–22.

Röchert D, Neubaum G, Ross B, et al. (2020) Opinion-based homogeneity on YouTube: combining sentiment and social network analysis. *Computational Communication Research* 2(1): 81–108.

Rodrigues da Cunha Palmieri E (2023) Social media, echo chambers and contingency: a system theoretical approach about communication in the digital space. *Kybernetes*. Available at: https://www.x-mol.net/paper/article/1643278470352744448

Rudgard O (2020) Why conspiracy theories are gaining ground in the pandemic. *The Telegraph*, 19 August. Available at: http://www.telegraph.co.uk/technology/2020/08/19/conspiracy-theories-gaining-ground-pandemic/

Schnackenberg AK and Tomlinson EC (2016) Organizational transparency: a new perspective on managing trust in organization-stakeholder relationships. *Journal of Management* 42(7): 1784–1810.

Serrat O and Serrat O (2017) Social network analysis. In: Serrat O (ed.) *Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance*. Singapore: Springer, pp.39–43.

Sharevski F, Huff A, Jachim P, et al. (2022) (Mis) perceptions and engagement on Twitter: COVID-19 vaccine rumors on efficacy and mass immunization effort. *International Journal of Information Management Data Insights* 2(1): 100059.

Sharifzadeh R (2016) Technology, agency and decision. *Culture Strategy* 34: 115–136.

Sheehy B, Choi S, Khan MI, et al. (2023). *Truths and Tales: Understanding Online Fake News Networks in South Korea*. University of Canberra. Available at: https://researchdata.canberra.edu.au/datasets/3xb4n9n6t4/1

Shin H, Kim M, Jo Y-M, et al. (2012) Annotation scheme for constructing sentiment corpus in Korean. In: *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, November, Bali, Indonesia, pp.181–190. Jawa Barat, Indonesia: Faculty of Computer Science, Universitas Indonesia.

Ulibarri N and Scott TA (2017) Linking network structure to collaborative governance. *Journal of Public Administration Research and Theory* 27(1): 163–181.

Unwin A (2020) Why is data visualization important? What is important in data visualization? *Harvard Data Science Review* 2(1): 1–7.

Vargo CJ, Guo L and Amazeen MA (2018) The agenda-setting power of fake news: a big data analysis of the online media landscape from 2014 to 2016. *New Media & Society* 20(5): 2028–2049.

Weikmann T and Lecheler S (2023) Cutting through the hype: understanding the implications of deepfakes for the fact-checking actor-network. *Digital Journalism*. Epub ahead of print 31 March. DOI: 10.1080/21670811.2023.2194665.

Whittle A and Spicer A (2008) Is actor network theory critique? *Organization Studies* 29(4): 611–629.

Yoo J, Kim D and Kim W-G (2022) Fake news on you, not me: the third-person effects of fake news in South Korea. *Communication Research Reports* 39(3): 115–125.

Zhang X and Ghorbani AA (2020) An overview of online fake news: characterization, detection, and discussion. *Information Processing & Management* 57(2): 102025.

Zhang Y, Chen F and Lukito J (2023) Network amplification of politicized information and misinformation about COVID-19 by conservative media and partisan influencers on Twitter. *Political* Communication 40(1): 24–47.

Zhen L, Yan B, Tang JL, et al. (2023) Social network dynamics, bots, and community-based online misinformation spread: lessons from anti-refugee and COVID-19 misinformation cases. *The Information Society* 39(1): 17–34.

Author biographies

Benedict Sheehy is a Professor of Law at Canberra Law School and an internationally recognised leader in the fields of corporate social responsibility, corporate law and regulatory theory. Benedict's research examines different aspects of how law works, how law fails and how society can more effectively achieve long term environmental and social sustainability using various types of law in a wide variety of contexts.

Sujin Choi (PhD from the University of Texas at Austin in the US) is an Associate Professor in the Department of Media at Kyung Hee University in Seoul, Korea. Her research explores the socio-political implications of journalistic/ algorithmic(AI)/social/personal curations in the digital sphere, utilizing inferential network analysis and computational methods.

Md Irfanuzzaman Khan is a Lecturer of Marketing at the Canberra School of Business, University of Canberra. His research interests include social media marketing, misinformation analysis, technology adoption theories, consumer behaviour and workplace behaviour.

Dr Bruce Baer Arnold is an Associate Professor in the School of Law at the University of Canberra. He has a strong interest in privacy, data protection, artificial intelligence and robotics, intellectual property and health sector regulation.

Yoonmo Sang is an Associate Professor in the Department of Media Communication at Sungshin Women's University in Seoul, South Korea. His primary research interests center on the intersection of new media technologies and the law, focusing on how socio-cultural and technological changes advantage and/or disadvantage different stakeholders

Jae- Jin Lee (Ph.D. from Southern Illinois University in the U.S.) is Professor of Dept. of Media & Communication, Hanyang University in Seoul, Korea. His research interests are in media law, ethics, and policy.

Article

new media & society
1–20

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448241253138
journals.sagepub.com/home/nms
Sage

What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape

Alena Birrer
University of Zurich, Switzerland

Natascha Just
University of Zurich, Switzerland

Abstract

The emergence of deepfakes has raised concerns among researchers, policymakers, and the public. However, many of these concerns stem from alarmism rather than well-founded evidence. This article provides an overview of what is currently known about deepfakes based on a systematic review of empirical research. It also examines and critically assesses regulatory responses globally through qualitative content analysis of policy and legal documents. The findings highlight gaps in our knowledge of deepfakes, making it difficult to assess the appropriateness and need for regulatory action. While deepfake technology may not introduce entirely new and unique regulatory problems at present, it can amplify existing problems such as the spread of non-consensual pornography and disinformation. Effective oversight and enforcement of existing rules, along with careful consideration of required adjustments will therefore be crucial. Altogether, this underscores the importance of more empirical research into the evolving challenges posed by deepfakes and calls for adaptive policy approaches.

Keywords

Deep fakes, deepfakes, state of empirical research, state of regulation

Corresponding author:
Alena Birrer, Media & Internet Governance Division, Department of Communication and Media Research, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland.
Email: [redacted]

Introduction

"You thought fake news was bad? Deepfakes are where truth goes to die" (Schwartz, 2018). Such headlines have been widely circulating with the advent of the deepfake phenomenon. The term "deepfake" first appeared in 2017, coined by a Reddit user to describe pornographic content apparently featuring the faces of famous women (McCosker, 2022). Since then, it has become a buzzword for manipulated media that rely on neural networks trained on extensive datasets to "learn" patterns that enable the imitation of real individuals and the synthesizing of fictional ones (Haller, 2022). Due to the use of this technology, deepfakes are said to be distinct from previous forms of falsified media, specifically in terms of scale, scope, and accessibility (Shahzad et al., 2022).

Deepfakes have caused widespread concern. However, much of the current debate is driven by anecdotal and speculative alarmism than by well-founded evidence and reasonable predictions (Kalpokas and Kalpokiene, 2022). Journalists have painted a dystopian picture and created a sense of impending doom (Gosse and Burkell, 2020; Wahl-Jorgensen and Carlson, 2021; Westerlund, 2019; Yadlin-Segal and Oppenheim, 2021). This is accompanied by fear-mongering by the Artificial Intelligence (AI) industry (Nature, 2023). Prominent tech executives have called for a temporary halt to the development of advanced AI (Pause Giant AI and Experiments: An Open Letter, 2023), and Microsoft's and Google's CEOs have publicly warned about the threats posed by deepfakes (Bartz, 2023). While this could be interpreted as a corporate response to demands for greater accountability, there may be a hidden agenda to stifle emerging competition (Bennett, 2023) and profit from "panic-marketing" (Weiss-Blatt, 2023).

Despite alarmist warnings that deepfakes will "wreak havoc on society" (Toews, 2020) and pressure on governments to intervene (e.g. Open Letter: Disrupting the Deepfake Supply Chain, 2024), regulators have been more hesitant in their responses. For example, the European AI Act[1] classifies deepfakes as "limited risk AI systems" and sets minimal transparency requirements. Legislation criminalizing the distribution of certain deepfakes has been enacted in the United States and in China, which was accompanied, however, by concerns that governments could use such rules to curtail free speech and control information flows (Hine and Floridi, 2022).

A growing research field discusses deepfakes' potential harm (see, for example, Chesney and Citron, 2019a for an overview), however, much less is known about the empirical evidence that substantiates these concerns. There are some literature reviews, but they focus exclusively on qualitative studies (Vasist and Krishnan, 2022b), or a limited number of empirical studies due to their publication date (Godulla et al., 2021; Vasist and Krishnan, 2022a). Moreover, there are no systematic overviews of dedicated regulatory responses to deepfakes. This article addresses these gaps and provides an up-to-date systematic literature review of what is currently empirically known about deepfakes and maps the emerging regulatory landscape through in-depth qualitative content analysis of policy and legal documents. This is to offer a comprehensive understanding of the deepfake phenomenon and to provide directions for future research and policymaking.

Systematic literature review

Considering calls for empirical evidence for regulation, we conducted a systematic literature review of *empirical* research on deepfakes to consolidate existing knowledge regarding their current uses, effects, consequences, and regulatory hurdles. Relevant literature was identified following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021; Figure 1). A detailed search of four databases (Scopus, Web of Science, EbscoHost, ProQuest) was conducted in June 2023. The scope was not confined to any specific research field as a systematic review aims to synthesize research from across disciplines. To ensure inclusion of all relevant literature, we conducted a broad search using the keywords "deepfake*" and "deep fake*." Search results were restricted to journal articles and conference proceedings in English. All identified records ($n = 3999$) were exported to Zotero and screened for duplicates, inaccessible, or obviously irrelevant studies. Next, eligibility criteria were defined and evaluated based on titles, abstracts, and, if necessary, full texts. During this process, theoretical essays, legal and literature reviews, and studies dealing exclusively with technical issues were excluded, as the goal was to synthesize *empirical* research on deepfakes. Emphasizing empirical research may inadvertently favor topics and regulatory challenges that are easier to investigate empirically, potentially overshadowing other issues discussed in the literature, such as financial fraud (e.g. Abbas et al., 2023;

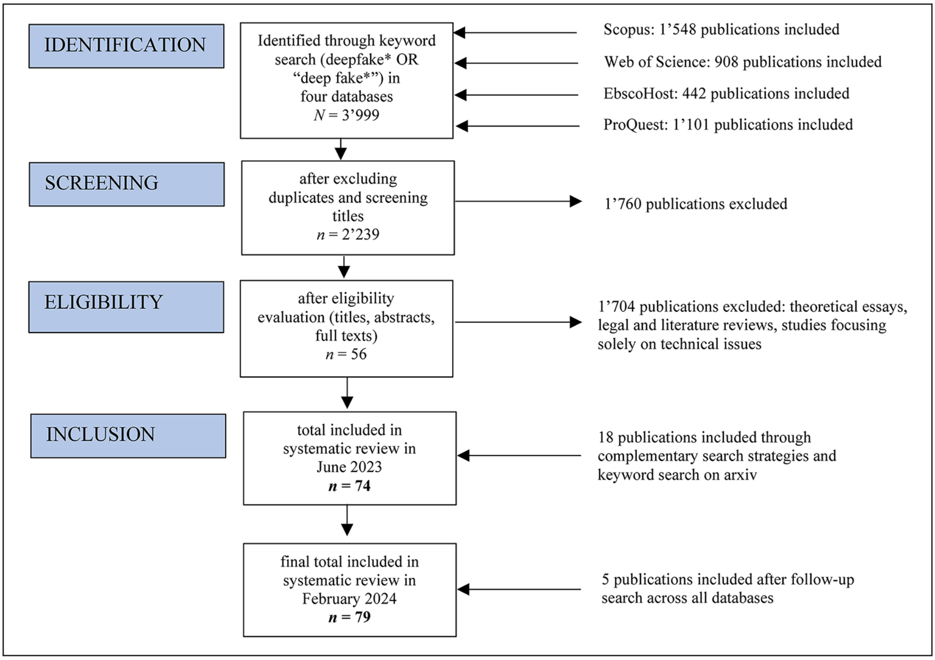

IDENTIFICATION
Identified through keyword search (deepfake* OR "deep fake*") in four databases N = 3'999
Scopus: 1'548 publications included
Web of Science: 908 publications included
EbscoHost: 442 publications included
ProQuest: 1'101 publications included
SCREENING
after excluding duplicates and screening titles n = 2'239
1'760 publications excluded
ELIGIBILITY
after eligibility evaluation (titles, abstracts, full texts) n = 56
1'704 publications excluded: theoretical essays, legal and literature reviews, studies focusing solely on technical issues
INCLUSION
total included in systematic review in June 2023 n = 74
18 publications included through complementary search strategies and keyword search on arxiv
final total included in systematic review in February 2024 n = 79
5 publications included after follow-up search across all databases

Figure 1. PRISMA flowchart of literature collection process.

Bateman, 2020). However, the analysis reveals that the regulatory problems identified as central in empirical literature align with those recognized as political priorities.

To enhance the comprehensiveness of our literature review and address database search limitations, we identified additional studies by reviewing reference lists of already included publications and conducting reverse searches on publications that cite important papers in the field. As this primarily revealed relevant computer-science studies available on arXiv, this database was searched for additional studies. In February 2024, following the initial review of this article, a follow-up search across all databases uncovered five additional articles, totaling 79 for analysis (see Annex I in the supplemental material for the list of studies).

Given the diversity of research questions and methods employed across the studies, we opted for a qualitative approach based on a deductive–inductive coding scheme to review the literature. Titles, authors, publication years, and disciplines[2] provided information on the field's evolution over time and across disciplines. The coding of the methods and the samples allowed comparison of results and methodological limitations. To shed light on the conceptual boundaries of the deepfake phenomenon, we also examined whether a definition of "deepfake" was provided and, if so, what conceptual elements were used in the definition. In addition, evidence of the prevalence of deepfakes and key regulatory challenges related to it were identified.

The following chapter offers a summary of the state of empirical research on deepfakes. Overall, empirical evidence remains limited, making it difficult to give informed assessments of current regulatory needs. Notably, there is limited research on how deepfakes are created, used, spread, and what potential positive and negative individual and societal impacts they can have. Thus we do not know whether the often-voiced concerns about the negative consequences of deepfakes align fully with the problems they may actually cause.

The research field over time and across disciplines

Empirical work on deepfakes emerged around 2020 and has steadily increased since then, with most studies being published in 2022. It comes from different disciplines, including the social sciences ($n = 40$), and information and computer science ($n = 37$). Empirical studies on deepfakes in the life sciences and law are scarce, with one study each.

Social-science research primarily focuses on the societal impact of deepfake disinformation and the psychological harm caused by deepfake pornography. These studies apply various methods, including online experiments, surveys, interviews, and content analyses. In contrast, information- and computer-science studies consist almost entirely of experimental studies that test automated versus human deepfake detection. Both the life science and law study rely on vignette surveys to assess user perceptions and ethical concerns related to deepfakes. A cross-citation analysis using LitMaps revealed that studies only rarely refer to each other across disciplinary boundaries, indicating a predominantly disciplinary rather than interdisciplinary nature. Future research could benefit from a more integrated knowledge on deepfakes.

With a few exceptions (e.g. Shahid et al., 2022), studies focus on the global north, especially the United States and European countries such as the United Kingdom, the Netherlands, and Germany, and only seven follow an internationally comparative approach. Because the potential negative effects of deepfakes could be greater in developing countries and under authoritarian regimes (Bateman, 2020; Gregory, 2021), there is need for more research in these specific contexts.

Definition of deepfake

Among the 79 articles, 61 provided a definition of the term "deepfake." There is, however, no universally accepted definition and studies diverged in their interpretation of common conceptual elements. For example, there is agreement on the use of technology as a key characteristic of deepfakes, but the lack of clarity on the specific technology required blurs the conceptual boundaries between deepfakes and less sophisticated audiovisual manipulations known as "cheap fakes" (Paris and Donovan, 2019) or "shallow fakes" (Johnson, 2019). Similarly, no consensus emerged on the media covered by the term; 36 definitions include videos, 17 images, and 15 audio. Finally, following the definitions by Westerlund (2019) and Chesney and Citron (2019a), 23 studies highlighted the hyper-realistic nature of the content and 20 specified that deepfakes involve a false depiction of someone saying or doing something they never did. Definitions typically do not refer to false representations of objects or events. Expanding the scope to include such depictions may prove crucial given the emergence of deepfakes not involving people, such as falsified satellite images (Zhao et al., 2021) or the widely shared deepfake showing an explosion at the US Pentagon (Marcelo, 2023). Future research should contribute to clarifying the meaning and conceptual limits of deepfakes to prevent it from degenerating into a conceptually ambiguous buzzword akin to "fake news." One potential approach could involve expanding the deepfakes concept to encompass the motives and actions of its creators, similar to the extension seen with "disinformation." So far however, only four studies have specified (malicious) intent as a conceptual element. In addition, contextual approaches to defining deepfakes could be helpful in more accurately capturing the circumstances in which deepfake technology is used and its distinct manifestations. Understanding the context can inform the development of more sophisticated detection and response strategies and can provide the flexibility needed to adapt definitions as new uses emerge.

Prevalence of deepfakes

Empirical evidence on the prevalence of deepfakes is largely missing. No clear conclusion can therefore be drawn about the extent of the deepfake phenomenon and its specific manifestations. However, initial research sheds light on how accessible deepfake technology currently is and what types of deepfakes are commonly circulating. In terms of accessibility, less data input is required compared to previous technology (Amezaga and Hajek, 2022). This may be reinforced by recent advancements such as OpenAI's "Sora" model, which generates video from text input (OpenAI, 2024). Nevertheless, studies suggest that advanced tools and time are required to create reasonably convincing deepfakes (Mehta et al., 2023; Weikmann and Lecheler, 2023). Accordingly, Gamage et al. (2022) observed an emerging marketplace for monetizing customized deepfake-production on Reddit. Regarding the circulation of deepfake content, a large part seems to be entertainment or humor (Cho et al., 2023; Dasilva et al., 2021), even during the Russo–Ukrainian war (Twomey et al., 2023). Such use of deepfakes has not been a primary concern for regulators, but this may be slowly shifting (see regulatory responses below). Some studies also referred to a report by Deeptrace, a Netherlands-based cybersecurity company, which found that 96% of the 14,678 identified deepfake videos in 2019 were pornographic (Ajder et al., 2019). According to a report by Home Security Heroes (2023), the number of deepfake videos rose to 95,820 in 2023, of which 98% were pornographic in nature. However, it is difficult to assess whether these are reliable numbers, and how they should be interpreted.

Key regulatory challenges

While deepfakes raise various concerns, the following three key regulatory challenges emerged in the analyzed empirical research: (1) people's (in)ability to detect deepfakes, (2) deepfake disinformation, and (3) deepfake pornography. The respective findings are discussed in the following. Special attention is paid to the effectiveness of countermeasures and the policy recommendations by scholars.

Detection. The first regulatory challenge relates to people's difficulties in detecting deepfakes—a concern that is frequently voiced in public discourse (Wahl-Jorgensen and Carlson, 2021). A range of experimental computer-science studies ($n=22$) investigated human deepfake detection, often compared to AI detectors. They often drew on large datasets containing deepfake *images* (Bray et al., 2023; Hulzebosch et al., 2020; Lago et al., 2022; Liu et al., 2020; Nightingale and Farid, 2022; Preu et al., 2022; Rössler et al., 2019; Shen et al., 2021) or *videos* (Chen et al., 2022; Groh et al., 2022; Khodabakhsh et al., 2019; Kim et al., 2018; Köbis et al., 2021; Korshunov and Marcel, 2020; Lovato et al., 2023; Prasad et al., 2022; Somoray and Miller, 2023; Tahir et al., 2021; Ternovski et al., 2021; Wöhler et al., 2021), while deepfake *audio* has not been sufficiently studied, with the exception of Müller et al. (2022). Across these studies, participants, on average, correctly identified 63.3% of deepfakes. Whether this is cause for concern is a matter of interpretation, depends on the specific context and requires more research. Research further suggests that detection varies greatly between different deepfakes. For example, lower image or video resolution made it harder for people to recognize whether content was authentic or not (Groh et al., 2022; Hulzebosch et al., 2020; Rössler et al., 2019; Tahir et al., 2021). This might be because low resolution impedes determining the authenticity of content, including spotting visual discrepancies (Lago et al., 2022; Preu et al., 2022; Tahir et al., 2021; Wöhler et al., 2021) and background inconsistencies (Lago et al., 2022; Preu et al., 2022; Tahir et al., 2021). In addition, Lovato et al. (2023) found that people were better at identifying deepfakes if the perceived demographic characteristics (age, gender, and ethnicity) of the person depicted matched their own.

No consistent patterns emerged as to who is particularly vulnerable to being fooled by deepfakes. While gender (Sütterlin et al., 2021; Tahir et al., 2021) and education (Tahir et al., 2021) had limited impact, there was evidence that older individuals had greater difficulty in detecting deepfakes (Ahmed, 2023; Müller et al., 2022). Studies also showed that people are often overly confident regarding their detection ability, especially those who were worse or equally bad at detecting deepfakes (Bray et al., 2023; Köbis et al., 2021; Lago et al., 2022; Preu et al., 2022). This is referred to as the Dunning-Kruger effect (Kruger and Dunning, 1999).

While there is a scarcity of empirical research concerning the effects of regulatory interventions on deepfake detection, a few studies offered initial insights and recommendations. For example, Köbis et al. (2021) found that both raising awareness and introducing financial incentives had no effect on people's detection accuracy. Similarly, providing immediate feedback to participants on whether they correctly identified a deepfake had limited effects in three experimental studies (Hulzebosch et al., 2020; Müller et al., 2022; Nightingale and Farid, 2022) and informing participants of common deepfake artifacts did not improve detection accuracy in another (Somoray and Miller, 2023). In contrast, introducing detailed walkthrough examples proved successful (Tahir et al., 2021), which may support the use of gamification approaches to literacy (see, for example, Glas et al., 2023 for a general overview of media literacy games). Groh et al. (2022) tested whether participants' ability to detect deepfakes improved when they were provided information on how AI detectors classified such content. They revealed that participants tended to over-trust AI, adjusting their own classification accordingly, even when AI was inaccurate. Despite the demonstrated AI detectors' superiority over human detection, this points to a religion-like, high share of (blind) faith-based versus knowledge-based trust in digital technology (Latzer, 2022). This may carry potential pitfalls, as AI detectors also have significant limitations and should therefore ideally be seen as Supplementary measures (Groh et al., 2022; Gupta et al., 2020; Korshunov and Marcel, 2020; Liu et al., 2020; Müller et al., 2022; Prasad et al., 2022; Tahir et al., 2021; Wöhler et al., 2021). Furthermore, the sole reliance on technical solutions to detect deepfakes could lead to a "cat-and-mouse" game (Lomtadze, 2019) as new technologies will find ways to circumvent current methods.

Overall, what is known about people's ability to detect deepfakes in computer-science research remains inconclusive. While it was confirmed that some people struggle to distinguish authentic from inauthentic content, whether this presents a regulatory problem depends on the specific context. To gain a more comprehensive understanding, the next section turns to social-science research and the deceptive potential of political deepfake disinformation and its societal consequences.

Deepfake disinformation. Social-science research has often conceptualized deepfakes as a form of disinformation and investigated its effects on politics and society. A first set of studies investigated whether deepfakes pose greater threats than other forms of disinformation, an assumption based on the argument that visual content typically holds greater persuasive power (e.g. Sundar, 2008). There is not enough research to conclusively assess whether this holds true, but initial findings did not support the uniquely deceptive nature of deepfakes. In two experimental studies (Hwang et al., 2021; Lee and Shin, 2022), deepfake videos were considered more vivid and credible compared to textual disinformation with the same message, but the differences were small. Furthermore, a few large experimental studies found that political deepfake videos are not perceived as more credible and emotionally appealing (Barari et al., 2021), not more effective in changing issue agreement or the evaluation of politicians (Appel and Prietzel, 2022; Hameleers et al., 2022, 2023), and not more likely to create false memories (Murphy and Flynn, 2022) than audio and/or textual disinformation. One explanation for why people might not be deceived as easily by political deepfakes is that they can spot "unnatural" behavior or expression when they are familiar with the person (Groh et al., 2023; Hameleers et al., 2024; Kim et al., 2018; Thaw et al., 2020; Vaccari and Chadwick, 2020). Preliminary findings thus suggest that the feared mass-deception by deepfake disinformation might be overstated. This may also hold true for the concern that they can be used to manipulate voters (Diakopoulos and Johnson, 2021), which was examined by Dobber et al. (2021). They found that showing participants a deepfake video featuring a politician of the Dutch Christian Party making jokes about Christ's crucifixion caused negative attitudes toward the politician. However, the effect was small and did not spill over to the politician's party. The authors also demonstrated that microtargeting could amplify negative attitudinal effects, but this effect was evident only within a small subgroup of highly religious Christians who had previously supported the Christian party. Similarly, Hameleers et al. (2024) found that partisanship is a likely driver of delegitimization of politicians through deepfakes. Thus, deepfake-based disinformation campaigns must be highly targeted to succeed. Together with the above finding that creating deepfakes still requires considerable resources and skills, this may suggest that, at least for now, they might not be the most appropriate tool for spreading disinformation. This is supported by fact-checkers interviewed by Weikmann and Lecheler (2023), who reported that deepfakes have so far caused far less turmoil than less sophisticated forms of visual disinformation and decontextualized images.

However, as indicated by research, the fundamental challenge posed by deepfake disinformation is its potential to contribute to a general climate of uncertainty and doubt. For example, Vaccari and Chadwick (2020) showed in an experiment that watching a deepfake video left people uncertain about what is real and what is not, which in turn reduced overall trust in social media and the news. Drawing from survey data, Ahmed (2021b) arrived at a similar conclusion, suggesting that deepfakes amplify overall skepticism toward the media. This could become one of the unintended consequence of raising public awareness of deepfakes, as Ternovski et al. (2021) and Lewis et al. (2023) showed. In their experiments, a "prebunking" intervention, that is, warning people about deepfakes, did not increase their detection accuracy, but instead made people more skeptical and led them to distrust all content presented, even if authentic. This in turn could be exploited by politicians to deflect accusations by delegitimizing facts as fiction. This is what Chesney and Citron (2019b) call the "liar's dividend" (p. 151). Accordingly, Twomey et al. (2023) found that during the Russo–Ukrainian war, Twitter users frequently denounced real content as deepfake, used "deepfake" as a blanket insult for disliked content, and supported deepfake conspiracy theories. Scholars have therefore recommended that mitigating measures must also focus on restoring trust in authentic content (Hameleers et al., 2024; Lewis et al., 2023; Tahir et al., 2021; Ternovski et al., 2021). Interventions could also focus on strengthening critical thinking, which—consistent with broader research on disinformation (Pennycook and Rand, 2019, 2020)—was identified as a relevant factor in preventing deception and the sharing of deepfakes (Ahmed, 2021c, 2023; Appel and Prietzel, 2022; Hameleers et al., 2024). In addition, corrective labels might provide another way of countering potential negative effects, as they reduced people's intention to share (Ahmed, 2021a; Lee and Shin, 2022) and mitigated ethical concerns regarding political deepfakes (Kugler and Pace, 2021). The latter did however not work for deepfake pornography, which is explored in the next section.

Deepfake pornography. The third regulatory challenge identified in empirical research relates to deepfake pornography and the resulting harm for individuals affected. Although the initial application of deepfakes was for pornography, empirical research on it is scarce, with only nine studies focusing on this area. A comprehensive, cross-country study by Flynn et al. (2022), which combined surveys ($N=6109$) and interviews ($N=118$) across the United Kingdom, New Zealand, and Australia, offered initial evidence into the pervasiveness of non-consensual deepfake pornography. Of the survey respondents, 14.1% reported being affected ($n=864$) by the creation, distribution, or threats of distribution of deepfake pornography featuring them; 7.6% reported having created or distributed such content ($n=466$). Belonging to a marginalized community and being younger and male predicted both victimization and perpetration. In addition, victims experienced a range of emotional, psychological, occupational, and relational effects, many of which continued long after the abuse had first taken place. This could lead to a constant "visceral fear" (Citron, 2019: 1925) over who has or will see the images in the future. In a broader sense, deepfakes could thus disrupt peoples' control over their own images, creating new forms of privacy invasions in terms of dignity, autonomy, and identity expression (Kugler and Pace, 2021). Studies further showed that victims are reluctant to speak up and report being affected by deepfake pornography due to a culture of victim blaming (Fido et al., 2022; Flynn et al., 2022; Winter and Salter, 2020) and normalization (Maddocks, 2020). Reporting was also hindered by missing or complex tools (De Angeli et al., 2021) and by authorities that discouraged victims from taking action because the perpetrator could not be identified (Flynn et al., 2022). Scholars therefore recommended creating better legal foundations and reporting mechanisms for deepfake pornography (Flynn et al., 2022; Kugler and Pace, 2021; Wang and Kim, 2022).

Altogether, the literature review indicated that deepfakes do not introduce fundamentally new and unique regulatory challenges. Instead, they add to the repertoire of tools available for spreading harmful or illegal content such as disinformation and non-consensual pornography. Consequently, the primary challenge lies in the effective oversight and enforcement of existing rules, along with careful considerations of required adjustments. This also necessitates consideration of potential unintended consequences when crafting countermeasures. Whether this is in line with emerging responses from regulators, is examined in the next section.

Regulatory responses to deepfakes

Research has generally begun to discuss the regulation of deepfakes and whether current laws are adequate to address them. In the United States, legal scholars have confirmed the general applicability of existing public and private law, but highlighted problems of enforceability such as identifying the responsible parties and cross-jurisdictional issues (e.g. Caldera, 2020; Chesney and Citron, 2019a; Hall, 2018; Langa, 2021; Meskys et al., 2020). In addition, discussions have intensified regarding the accountability of Internet platforms for the content they host, including deepfakes (O'Donnell, 2021), as have discussions on the need to strike a balance between regulatory measures and safeguarding freedom of speech (Bodi, 2021). Consequently, the emergence of deepfakes does not necessarily raise new regulatory questions, but intensifies existing ones (Barber, 2023). In the European Union (EU), problematic deepfakes fall under the scope of several European regulations designed to address harmful and illegal online content, notably the Digital Services Act (DSA, Regulation (EU) 2022/2065) and the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679), in addition to national law and the AI Act (Karaboga, 2023). Here, legal scholars have highlighted the need to strengthen enforcement, grant additional legal rights to victims, and foster public awareness (e.g. Van der Sloot and Wagensveld, 2022; Van Huijstee et al., 2021).

In contrast, there has been relatively little scholarly focus on dedicated regulatory responses to deepfakes and no global overview currently exists. To bridge this gap, we conducted a qualitative content analysis (Puppis, 2019) using the coding software MAXQDA to examine enacted and proposed regulatory measures and policy debates surrounding deepfakes. Initially, by November 2023, we found 50 documents through a review of existing research, in-depth searches of regulatory authorities' websites, and thorough monitoring of media coverage and policy blogs. Following the manuscript's first review, we added 50 more documents in February 2024, showcasing the highly dynamic policy landscape (see Annex II in the supplemental material for the list of 100 documents).

Overall, a diverse spectrum of regulatory responses to deepfakes emerged, ranging from market-driven initiatives to state-imposed command-and-control-regulation, with various forms of self- and co-regulation in between (Latzer et al., 2002). Some policymakers have chosen to refrain from regulatory action altogether, either due to limited research on deepfakes or the belief that current laws and industry self-regulation adequately addresses them. Others rely on self- and co-regulation aimed at raising awareness as well as hard regulations that require transparency or ban or otherwise limit the production or distribution of certain deepfakes. The measures target different stages of the deepfake lifecycle and consequently vary in their focus, applying to producers of deepfake technology, users who create or disseminate deepfakes, or the platforms that host them.

In the following sections, we present the main approaches taken by policymakers in response to deepfakes categorized according to the intensity of state involvement. We also compare them with the findings from our literature review, with particular emphasis on the rationale behind the need for regulatory action, the actors held accountable, and whether the enacted or proposed measures appear adequate to address the regulatory challenges identified.

No state regulation

Despite widespread concern among both the public and scholars, policymakers and legislators have generally been cautious, partly opting for a wait-and-see approach. This is largely justified by the lack of empirical research on deepfakes, as identified in our literature review. Accordingly, Austria (14) and Belgium (15) plan to intensify research on deepfakes before considering further action. In the United States, efforts are underway to institutionalize research on deepfakes through taskforces, regular reports (44, 45, 46, 47), and a mandated intelligence assessment by the Secretary of Defense regarding national security threats posed by deepfakes (98). However, concurrently several deepfake laws have been adopted, criminalizing the dissemination of certain deepfakes (see below).

In some countries, policymakers have explicitly decided against regulatory action. In the Netherlands, the Ministry of Justice and Security reviewed the need for deepfake regulation prompted by a 2021 report that suggested regulatory options (Van der Sloot et al., 2021), but in 2023 decided not to criminalize all or even specific types of deepfakes. This decision was based on the belief that existing laws are sufficient, coupled with concerns about potential constraints on freedom of expression (23). Similarly, in 2023, the Swiss Federal Council denied a motion to regulate deepfakes, asserting that the use of deepfake applications does not create legal loopholes in criminal and civil law (100). It remains to be seen whether this will be reconsidered after the Swiss Foundation for Technology Assessment (TA-SWISS, 2023) publishes the results of an ongoing study on the impact of deepfakes.

In the absence of dedicated regulatory action, some of the aforementioned countries (14, 15, 23) have expressed their intent to strengthen public awareness of deepfakes to preemptively counter potential negative consequences, as discussed in the following.

Strengthening public awareness and fostering transparency

A second approach centers around soft measures to enhance public awareness of deepfakes. This includes increased efforts to educate the public on how to identify deepfakes and self- and co-regulatory approaches to transparency.

A first set of measures under discussion or already implemented in some countries involves raising awareness and improving people's abilities to recognize deepfakes. The European Parliament has, for example, consistently advocated such action (8, 9). In addition, some member states, including Austria (14), Belgium (15), and the Netherlands (23), plan to adopt measures aimed at strengthening deepfake-specific literacy. Furthermore, the Italian Data Protection Authority (22) and the German Federal Office for Information Security (21) have already released information on how users can protect themselves from deepfakes, although the focus is primarily on general information about digital artifacts in deepfake content.

These measures have been justified by the assumption that the public cannot differentiate between deepfakes and authentic content, which was not fully supported by the literature review. While such awareness measures can serve as a starting point to mitigate some potential negative impacts of deepfakes, the literature review has further indicated their limited efficacy. For example, raising awareness in general might not significantly improve people's ability to detect deepfakes and can sometimes even backfire and create uncertainty. Hence, measures aimed at strengthening public literacy should also focus on rebuilding trust in authentic content and recognize people's tendency to overestimate their abilities.

Major Internet platforms have also made efforts to contribute to strengthening public deepfake literacy. For example, Meta and Google have created large public deepfake datasets to advance research on deepfake detection, which were used in several of the computational studies quoted above in the "detection" section. Platforms have also successively instituted deepfake policies and created technology designed to detect, label, or remove deepfakes. Early on, their primary focus was on deepfake pornography. Accordingly, Reddit banned deepfake pornography in 2018, followed by Pornhub, Discord, and X. The focus has since shifted to combating electoral interference and developing industry-wide standards. For example, 20 large tech companies signed a joint "tech accord" to tackle deceptive AI use in 2024 elections around the world (AI elections accord, 2024). Furthermore, the Coalition for Content Provenance and Authenticity (C2PA) (2023) introduced the "Content Credentials" watermarking system to trace the sources and integrity of digital content.

In addition, the EU continues to promote self-regulation aimed at fostering transparency, which has a long tradition for combating disinformation and other harmful content and which has recently been extended to include deepfakes. The 2018 Code of Practice on Disinformation, a self-regulatory framework initiated by the European Commission to tackle disinformation, was strengthened in 2022 (6) and now includes recommendations for labeling deepfake content. Although an evaluation of the initial code revealed that labels alone might not be an effective measure (European Regulators Group for Audiovisual Media Services, 2020), this approach was continued and legally backed up by the recently enacted Digital Services Act (7) and the AI Act (5), which is discussed in the next section.

Transparency, criminalization, and ex-ante control

While industry self-regulation has been a cornerstone in Internet governance, confidence in it has slowly declined and complementary hard regulation has been gradually adopted (Floridi, 2021; Shattock, 2021), also regarding deepfakes.

In the EU, deepfakes are addressed as part of broader frameworks regulating online platforms and AI. The DSA (7) requires providers of very large online platforms and search engines to label deepfakes. In addition, *illegal* deepfake content is subject to its stricter notice-and-action-procedures and systemic risk mitigation. A legally binding transparency approach was also included in the AI Act (5). Under its risk-based framework, AI systems generating deepfakes are classified as "limited risk." Accordingly, deployers, defined as "any natural or legal person using an AI system under its authority," must disclose whether content has been artificially generated or manipulated in a clear, timely, and accessible manner. Notably, the term "deployer" explicitly does not encompass the personal use of deepfakes. In addition, disclosure is not required if deepfakes are used by law enforcement and when such use is essential for exercising freedom of expression, arts, and sciences. An EU AI Office shall encourage codes of practice to aid rule implementation, while the European Commission is empowered to adopt further implementing acts. These mechanisms could alleviate concerns about platforms' extensive discretion in rule implementation. Furthermore, AI system providers may be mandated to adopt technical detection and labeling solutions. Still, uncertainties and criticism persist regarding the enforcement of the transparency obligation and the efficacy of available mechanisms for sanctions (e.g. Karaboga, 2023; Van Huijstee et al., 2021). For example, labels are inadequate to mitigate harm related to deepfake pornography, as shown in the literature review. Therefore, exploring context-specific transparency measures could be an option.

Law enforcement's use of deepfake-detection software was initially considered high risk in the European Commission's AI Act proposal (1), but was later removed at the European Parliament's request (2) due to an unreasonable distinction between private and public uses of deepfakes (4). However, certain deepfake applications may be classified as high risk in the future as the list is subject to adaptation. This aligns well with the need for adaptive policy approaches (Latzer, 2013) in the light of limited controllability and predictability of deepfake technology. Future research could thus help identify high-risk deepfake applications.

Other measures focus on criminalizing specific deepfake applications and adopting preemptive rules to curb the harmful use of deepfake technology. Such measures address the regulatory challenges described in the literature, focusing primarily on electoral manipulation through deepfake disinformation and deepfake pornography. However, they often rely on unverified assumptions about the prevalence and deceptive capacity of deepfakes and fuel unsubstantiated alarmist narratives. Moreover, the enforceability and appropriateness of some measures may be contested considering the findings of the literature review.

China was among the first countries to adopt rules relating to deepfakes. The "Provisions on the Administration of Deep Synthesis of Internet Information Services" (16), effective from 10 January 2023, consist of a range of instruments that target all "synthesized media," including deepfakes. They generally prohibit the creation of synthesized media that violate national law or threaten national security and interests, harm the national image, or disrupt the economy. In addition, they implement a notice-and-action system that requires synthetic-media providers, that is, providers of apps for the production of synthetic content, to prominently label synthetic content, and platform operators to identify and remove content deemed "undesirable," especially disinformation. Both rules have raised concerns of excessive censorship, because companies may over-police content to avoid legal liability and such rules can easily be abused to exert excessive government control of information (Kölling, 2023; Sheehan, 2023). App providers must also verify users' identity to access their services, further increasing privacy concerns (Hine and Floridi, 2022). In addition, the "Interim Measures for the Management of Generative Artificial Intelligence Services" (17) require new generative AI products with "public opinion attributes" or "capabilities for social mobilization" to undergo a security review prior to release. App providers would also be tasked with helping users understand and responsibly use deepfakes to avoid harming others. How these demanding ex-ante requirements will be enforced remains to be seen.

In the United States, several state and federal bills are in force or under consideration that target specific harmful uses of deepfakes. They primarily target individual users and address two concerns: the dissemination of election-interfering deepfakes and non-consensual pornography. This approach aligns with the United States' traditionally high level of free speech protection and its liability shield for platform providers (Geng, 2023).

The federal "Protect Elections from Deceptive AI Act" (55) and the US Federal Election Commission (52) aim to ban deceptive deepfake content in political ads. South Korea (99) recently enacted a similar ban, facing criticism that it might be misused for controlling elections (Park, 2024). Moreover, legislation criminalizing the distribution of deepfakes during elections has been successfully passed in Texas (42), California (41), Minnesota (29), Maryland (39), Washington (57) and Michigan (69). Proposals are pending in 30 more states in anticipation of the 2024 elections (28, 56, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 70, 71, 72, 73, 74, 75, 76, 78, 79, 80, 90, 91, 92, 95, 96, 97). Amid concerns over election interference, President Biden also signed an executive order advocating for the watermarking of AI-produced content (54), and the US Federal Communications Commission has prohibited AI-generated robocalls (87). Although these measures are more narrowly targeted, they have been criticized for potentially interfering with the right to free speech (Tashman, 2021), particularly in light of previous court decisions that rejected attempts to restrict election-related lies (Chesney and Citron, 2019a). At the same time, many have questioned their enforceability, particularly given the difficulty in identifying deepfake producers and determining harmful intent (e.g. Williams et al., 2019).

Another set of US bills is dedicated to deepfake pornography. At the federal level, the "Preventing Deepfakes of Intimate Images Act" (32) was introduced in the House of Representatives in 2023 and would make the dissemination of pornographic deepfakes illegal and provide additional legal options for victims. The "DEEP FAKES Accountability Act," which initially failed twice (48, 49) before being reintroduced in 2023 (50), would require users to digitally watermark pornographic deepfake content, along with content shared to "incite violence, physical harm, provoke armed or diplomatic conflict, or disrupt official proceedings." In addition, the bipartisan "DEFIANCE Act" (85) was introduced in early 2024 to establish civil remedies for victims of deepfake pornography, prompted by falsified explicit images of Taylor Swift circulating online. At the state level, legislation banning deepfake pornography is already in effect in Virginia (43), California (40), Minnesota (29), New York (31, 34), Illinois (81), Texas (82), Hawaii (83), and Georgia (84). Furthermore, a pending Maryland bill seeks to establish a dedicated taskforce to prevent and deal with deepfake pornography (38). Beyond the United States, deepfake pornography has also legally been declared a civil offense in Australia (12), granting the Australian eSafety-Commissioner the authority to require service providers or users to remove deepfake pornography posted online. Similar discussions have been underway in the United Kingdom, with the UK Law Commission recommending criminalizing the sharing of deepfake pornography (24), a measure included in the much-anticipated Online Safety Act passed in October 2023 (26). Concurrently, a coalition of bipartisan politicians in the United Kingdom has called for a comprehensive ban on all harmful deepfakes across the entire production and distribution process (27). France is also discussing an amendment to its penal code to specifically include the sharing of deepfake pornography (20), and Belgium has expressed its intention to develop a legal framework to prosecute and enforce the misuse of deepfakes, including pornography (15). Furthermore, the EU directive on combating violence against women would require all Member States to make the non-consensual sharing of deepfake pornography a criminal offense (10). Overall, the need for improved reporting mechanisms and stricter law enforcement could be supported by research on the long-lasting effects on victims and the lack of reporting tools described earlier. However, despite these measures, the burden of proof still falls on victims, who may be discouraged from reporting abuse due to victim blaming, raising doubts about their effectiveness.

While the mentioned proposals primarily tackle malicious deepfakes and often exclude those created or shared for entertainment purposes, there are increasing concerns in the United States regarding the protection of the creative industry against AI-generated content (Rose, 2024). In response, the "NO FAKES Act of 2023" (86) and "No AI FRAUD Act" (88) proposals aim to strengthen individuals' right to publicity by protecting their voice and visual likeness from unauthorized AI recreation, even beyond their death. Similar legislation has been passed in New York (35) and is pending in Tennessee (77). In addition, the US Federal Trade Commission has proposed new rules to prohibit AI impersonation of individuals (89). However, some have questioned the necessity of these additional rules considering existing personality rights and warned about their potential to benefit big labels over individual artists (Rothman, 2023).

In sum, the content analysis revealed a spectrum of responses to deepfakes. Most prioritize public awareness and transparency over the criminalization and control of deepfake applications. In addition, there seems to be an understanding that existing laws—although sometimes extended in scope—are generally equipped to address deepfakes. Yet concerns about enforcement and efficacy persist, particularly when evaluated against the findings of the literature review.

Conclusion

This article offers a comprehensive review of existing empirical research on deepfakes and the regulatory responses to this emerging technology. The findings indicate that our understanding and knowledge of deepfakes is not yet sufficient to determine whether the commonly held concerns about their harmful impacts are materializing and how to effectively address them. At present, it seems that the challenges posed by deepfakes are not entirely unprecedented but rather an extension of ongoing discussions regarding the dissemination of harmful and illegal content. We therefore advocate the need for evidence-based knowledge and empirical research, over rushed and anecdotal assumptions. The ever-evolving landscape of deepfake technology necessitates adaptive policy approaches (Latzer, 2013) aimed at mitigating harm while safeguarding individual rights and addressing broader societal issues related to trust and truth. Risk-based approaches, as adopted in the AI Act, seem promising in striking this balance. Nonetheless, existing tools may not fully resolve current and future challenges, making critical oversight and periodic review essential. In addition, it is crucial to give careful consideration to adequate governance arrangements, considering both appropriate state and private involvement as highlighted in the governance-choice approach (Latzer et al., 2019). Altogether, this underscores the importance of conducting more empirical research to effectively address and understand the evolving regulatory challenges posed by deepfake technology. In particular, as noted in the respective sections above, future research should clearly define deepfakes and its diverse applications and explore the harmful and beneficial individual and societal impact they can have, also beyond the global north. Moreover, further research is needed to understand the intended and unintended consequences of countermeasures, thus strengthening evidence-based policymaking. Given the rapid advancement of technology, future research should also consolidate and integrate knowledge on deepfakes and novel phenomena in synthetic-media production across disciplinary boundaries.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Swiss National Science Foundation [grant number 209250].

ORCID iD

Alena Birrer https://orcid.org/0000-0002-7141-7028

Supplemental material

Supplemental material for this article is available online.

Notes

1. The final compromise text, published in January 2024, is expected to be adopted in the upcoming months.
2. Based on the discipline of journals according to the ISI Web of Science Journal Citation Report (JCR) and the discipline of the authors according to their institutional affiliation.

References

References of included studies are given in the supplementary file.

Abbas NN, Ahmad R, Qazi S, et al. (2023) Impact of deepfake technology on fintech applications. In: Saeed S, Almuhaideb A, Kumar N, et al. (eds) *Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications*. Hershey, PA: IGI Global, pp. 225–242.

AI Elections accord (2024) A tech accord to combat deceptive use of AI in 2024 elections. Available at: https://www.aielectionsaccord.com/ (accessed 20 February 2024).

Ajder H, Patrini G, Cavalli F, et al. (2019) The State of Deepfakes: Landscape, Threats, and Impact. Amsterdam: Deeptrace. Available at: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf (accessed 21 June 2023).

Barber A (2023) Freedom of expression meets deepfakes. *Synthese* 202(2): 1–17.

Bartz D (2023) Microsoft chief says deep fakes are biggest AI concern. *Reuters*, 25 May. Available at: https://www.reuters.com/technology/microsoft-chief-calls-humans-rule-ai-safeguard-criticalinfrastructure-2023-05-25/ (accessed 24 August 2023).

Bateman J (2020) *Deepfakes and synthetic media in the financial system: assessing threat scenarios*. Cyber Policy Initiative Working Paper Series, “Cybersecurity and the Financial System.” Washington, DC: Carnegie Endowment for International Peace. Available at: https://carnegieendowment.org/files/Bateman_FinCyber_Deepfakes_final.pdf (accessed 3 July 2023).

Bennett MT (2023) No, AI probably won’t kill us all—and there’s more to this fear campaign than meets the eye. *The Conversation*, 1 June. Available at: http://theconversation.com/no-aiprobably-wont-kill-us-all-and-theres-more-to-this-fear-campaign-than-meets-the-eye-206614 (accessed 23 August 2023).

Bodi M (2021) The first amendment implications of regulating political deepfakes. *Rutgers Computer and Technology Law Journal* 47(1): 143–172.

Caldera E (2020) “Reject the evidence of your eyes and ears”: Deepfakes and the law of virtual replicants. *Seton Hall Law Review* 50(1): 177–205.

Chesney R and Citron D (2019a) Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review* 107(6): 1753–1820.

Chesney R and Citron D (2019b) Deepfakes and the new disinformation war the coming age of post-truth geopolitics. *Foreign Affairs* 98(1): 147–155.

Citron D (2019) Sexual privacy. *The Yale Law Journal* 128(7): 1870–1960.

Coalition for Content Provenance and Authenticity (C2PA) (2023) Introducing official content credentials icon. Available at: https://c2pa.org/post/contentcredentials/ (accessed 30 January 2024).

Diakopoulos N and Johnson D (2021) Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society* 23(7): 2072–2098.

European Regulators Group for Audiovisual Media Services (2020) ERGA report on disinformation: assessment of the implementation of the code of practice. Available at: https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf (accessed 19 July 2023).

Floridi L (2021) The end of an era: from self-regulation to hard law for the digital industry. *Philosophy & Technology* 34(4): 619–622.

Geng Y (2023) Comparing “deepfake” regulatory regimes in the United States, the European Union, and China. *Georgetown Law Technology Review* 7(1): 157–178.

Glas R, Van Vught J, Fluitsma T, et al. (2023) Literacy at play: an analysis of media literacy games used to foster media literacy competencies. *Frontiers in Communication* 8: 1–16.

Godulla A, Hoffmann CP and Seibert DMA (2021) Dealing with deepfakes—an interdisciplinary examination of the state of research and implications for communication studies. *Studies in Communication and Media* 10(1): 73–96.

Gosse C and Burkell J (2020) Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication* 37(5): 497–511.

Gregory S (2021) Authoritarian regimes could exploit cries of “deepfake.” *Wired*, 14 February. Available at: https://www.wired.com/story/opinion-authoritarian-regimes-could-exploit-cries-of-deepfake/ (accessed 20 July 2023).

Hall HK (2018) Deepfake videos: when seeing isn’t believing. *Catholic University Journal of Law and Technology* 27(1): 51–76.

Haller E (2022) The two faces of deepfakes: cybersecurity & identity fraud. *Security Magazine*, 15 February. Available at: https://www.securitymagazine.com/articles/97085-the-two-faces-ofdeepfakes-cybersecurity-and-identity-fraud (accessed 24 March 2023).

Hine E and Floridi L (2022) New deepfake regulations in China are a tool for social stability, but at what cost? *Nature Machine Intelligence* 4: 608–610.

Home Security Heroes (2023) 2023 state of deepfakes: realities, threats, and impact. Available at: https://www.homesecurityheroes.com/state-of-deepfakes/ (accessed 31 January 2024).

Johnson B (2019) Deepfakes are solvable—but don’t forget that “shallowfakes” are already pervasive. *MIT Technology Review*, 25 March. Available at: https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/ (accessed 23 June 2023).

Kalpokas I and Kalpokiene J (2022) On alarmism: between infodemic and epistemic anarchy. In: Kalpokas I and Kalpokiene J (eds) *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation* (SpringerBriefs in Political Science). Cham: Springer, pp. 41–53.

Karaboga M (2023) Die Regulierung von Deepfakes auf EU-Ebene: Überblick eines Flickenteppichs und Einordnung des Digital Services Act- und KI-Regulierungsvorschlags.

In: Jaki S and Steiger S (eds) *Digitale Hate Speech: Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation*. Berlin and Heidelberg: Springer, pp. 197–220.

Kölling M (2023) Zum Schutz oder zur Zensur? China erlässt Gesetz für Deepfakes. *Heise Online*, 26 January. Available at: https://www.heise.de/hintergrund/Zum-Schutz-oder-zur-Zensur-China-erlaesst-Gesetz-fuer-Deepfakes-7470247.html (accessed 17 August 2023).

Kruger J and Dunning D (1999) Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77(6): 1121–1134.

Langa J (2021) Deepfakes, real consequences: crafting legislation to combat threats posed by deepfakes notes. *Boston University Law Review* 101(2): 761–802.

Latzer M (2013) Medienwandel durch Innovation, Ko-Evolution und Komplexität: Ein Aufriss. *Medien & Kommunikationswissenschaft* 61(2): 235–252.

Latzer M (2022) The digital trinity—controllable human evolution—implicit everyday religion. *Kzfss Kölner Zeitschrift für Soziologie und Sozialpsychologie* 74(1): 331–354.

Latzer M, Just N, Saurwein F, et al. (2002) *Selbst- und Ko-Regulierung im Mediamatiksektor*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Latzer M, Saurwein F and Just N (2019) Assessing policy II: governance-choice method. In: Van den Bulck H, Puppis M, Donders K, et al. (eds) *The Palgrave Handbook of Methods for Media Policy Research*. Cham: Springer, pp. 557–574.

Lomtadze A (2019) The WSJ on DeepFakes: "it's a cat & mouse game." *Global Editors Network*, 10 October. Available at: https://medium.com/global-editors-network/wsj-on-deepfakes-its-a-cat-mouse-game-212c3c7c6a4 (accessed 3 October 2023).

Marcelo P (2023) FACT FOCUS: Fake image of Pentagon explosion briefly sends jitters through stock market. AP News, 23 May. Available at: https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4 (accessed 29 February 2024).

Meskys E, Liaudanskas A, Kalpokiene J, et al. (2020) Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice* 15(1): 24–31.

Nature (2023) Stop talking about tomorrow's AI doomsday when AI poses risks today. *Nature* 618: 885–886.

O'Donnell N (2021) Have we no decency? Section 230 and the liability of social media companies for deepfake videos. *University of Illinois Law Review*. Available at: https://www.illinoislawreview.org/wp-content/uploads/2021/03/ODonnell.pdf

OpenAI (2024) Sora: creating video from text. Available at: https://openai.com/sora (accessed 29 February 2024).

Open Letter: Disrupting the Deepfake Supply Chain (2024). Available at: https://openletter.net/l/disrupting-deepfakes (accessed 21 February 2024).

Page MJ, McKenzie JE, Bossuyt PM, et al. (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal* 372(71): 1–9.

Paris B and Donovan J (2019) Deepfakes and cheap fakes. *Data & Society*, 18 September. Available at: https://datasociety.net/library/deepfakes-and-cheap-fakes/ (accessed 23 June 2023).

Park C-K (2024) South Korea's Yoon accused of using "fake news" crackdown to gag dissent ahead of polls. *South China Morning Post*, 26 February. Available at: https://www.scmp.com/week-asia/politics/article/3253245/south-koreas-yoon-accused-using-fake-news-crackdown-gag-dissent-ahead-polls (accessed 29 February 2024).

Pause Giant AI and Experiments: An Open Letter (2023). Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (accessed 25 May 2023).

Pennycook G and Rand DG (2019) Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188: 39–50.

Pennycook G and Rand DG (2020) Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality* 88(2): 185–200.

Puppis M (2019) Analyzing talk and text I: qualitative content analysis. In: Van den Bulck H, Puppis M, Donders K, et al. (eds) *The Palgrave Handbook of Methods for Media Policy Research*. London: Palgrave Macmillan, pp. 367–384.

Rose J (2024) Congress is trying to stop AI nudes and deepfake scams because celebrities are mad. *Vice*, 16 January. Available at: https://www.vice.com/en/article/5d9az5/congress-is-trying-to-stop-ai-nudes-and-deepfake-scams-because-celebrities-are-mad (accessed 6 February 2024).

Rothman JE (2023) Draft digital replica bill risks living performers' rights over AI-generated replacements. Available at: https://rightofpublicityroadmap.com/news_commentary/draft-digital-replica-bill-risks-living-performers-rights-over-ai-generated-replacements/ (accessed 19 February 2024).

Schwartz O (2018) You thought fake news was bad? Deep fakes are where truth goes to die. *The Guardian*, 12 November. Available at: https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth (accessed 6 October 2023).

Shahzad SA, Hashmi A, Khan S, et al. (2022) Lip sync matters: a novel multimodal forgery detector. In: *Proceedings of 2022 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)*, Chiang Mai, Thailand, 7–10 November, pp. 1885–1892. New York: IEEE.

Shattock E (2021) Self-regulation 2:0? A critical reflection of the European fight against disinformation. *Harvard Kennedy School Misinformation Review* 2(3): 1–8.

Sheehan M (2023) China's AI regulations and how they get made. *Carnegie Endowment for International Peace*. Available at: https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117 (accessed 21 October 2023).

Sundar SS (2008) The MAIN model: a heuristic approach to understanding technology effects on credibility. In: Metzger MJ and Flanagin AJ (eds) *Digital Media, Youth, and Credibility*. Cambridge, MA: MIT Press, pp. 73–100.

Tashman A (2021) "Malicious deepfakes"—how California's A.B. 730 tries (and fails) to address the Internet's burgeoning political crisis. *Loyola of Los Angeles Law Review* 54(4): 1391–1421.

TA-SWISS (2023) Deepfakes und manipulierte Realitäten. *TA-SWISS*. Available at: https://www.ta-swiss.ch/deepfakes (accessed 22 October 2023).

Toews R (2020) Deepfakes are going to wreak Havoc on society. We are not prepared. *Forbes*, 25 May. Available at: https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/ (accessed 6 October 2023).

Van der Sloot B and Wagensveld Y (2022) Deepfakes: regulatory challenges for the synthetic society. *Computer Law & Security Review* 46: 1–15.

Van der Sloot B, Wagensveld Y and Koops B-J (2021) *Deepfakes. De juridische uitdagingen van een synthetische samenleving*. Tilburg: Tilburg Institute for Law, Technology, and Society, Tilburg University. Available at: https://repository.wodc.nl/handle/20.500.12832/3134 (accessed 22 June 2023).

Van Huijstee M, Van Boheemen P, Das D, et al. (2021) Tackling Deepfakes in European Policy. *Strasbourg: European Parliament*. Available at: https://data.europa.eu/doi/10.2861/325063 (accessed 9 August 2023).

Vasist PN and Krishnan S (2022a) Deepfakes: an integrative review of the literature and an agenda for future research. *Communications of the Association for Information Systems* 51(1): 556–602.

Vasist PN and Krishnan S (2022b) Engaging with deepfakes: a meta-synthesis from the perspective of social shaping of technology theory. *Internet Research* 33(5): 1670–1726.

Wahl-Jorgensen K and Carlson M (2021) Conjecturing fearful futures: journalistic discourses on deepfakes. *Journalism Practice* 15(6): 803–820.

Weiss-Blatt N (2023) The AI doomers' playbook. *Techdirt*, 14 April. Available at: https://www.techdirt.com/2023/04/14/the-ai-doomers-playbook/ (accessed 18 August 2023).

Westerlund M (2019) The emergence of deepfake technology: a review. *Technology Innovation Management Review* 9(11): 39–52.

Williams J, McKinney I and Tsukayama H (2019) Congress should not rush to regulate deepfakes. *Electronic Frontier Foundation*, 24 June. Available at: https://www.eff.org/de/deeplinks/2019/06/congress-should-not-rush-regulate-deepfakes (accessed 24 August 2023).

Wöhler L, Zembaty M, Castillo S, et al. (2021) Towards understanding perceptual differences between genuine and face-swapped videos. In: *CHI '21: Proceedings of the 2021 CHI conference on human factors in computing systems*, Yokohama, Japan, 8–13 May, pp. 1–13. New York: ACM.

Yadlin-Segal A and Oppenheim Y (2021) Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence* 27(1): 36–51.

Zhao B, Zhang S, Xu C, et al. (2021) Deep fake geography? When geospatial data encounter Artificial Intelligence. *Cartography and Geographic Information Science* 48(4): 338–352.

Author biography

Alena Birrer is a Research and Teaching Associate in the Media and Internet Governance Division at the Department of Communication and Media Research of the University of Zurich. Her research interests are related to communications policy and Internet governance. They include privacy and identity, communication rights, and media concentration and market power control.

Natascha Just is Professor of Communication and Chair of the Media and Internet Governance Division at the Department of Communication and Media Research of the University of Zurich. Her research interests are related to communications policy and Internet governance. They center on competition policy, market power control, changing governance structures, governance of and by technology, algorithms on the Internet, and Internet platforms.

DISINFORMATION, DEEPFAKES AND DEMOCRACIES: THE NEED FOR LEGISLATIVE REFORM

ANDREW RAY*

Rapid technological advancement is changing the way that political parties, voters, and media platforms engage with each other. This along with cultural change has led to an emerging era of disinformation and misinformation driven by both domestic and foreign actors. Political deepfakes, videos created through the use of artificial intelligence, allow individuals to rapidly create fake videos indistinguishable from true content. These videos have the capacity to undermine voter trust and could alter electoral outcomes. Regulating disinformation however raises significant free speech concerns, as well as questions about where liability should fall. In particular, holding large technology and media platforms accountable for content could lead to unintended chilling effects around freedom of expression, harming rather than protecting democratic institutions. Proposed regulations should therefore be carefully analysed through the framework of the implied freedom of political communication, ensuring that any new laws are proportionate and tailored to the threat they seek to prevent. This article analyses how current Australian law interacts with political deepfakes and proposes two targeted amendments to our federal electoral regulations to reduce the threat they pose to elections.

I INTRODUCTION

The rapid advancement of artificial intelligence ('AI') and machine-learning algorithms ('MLAs') is disrupting the way that we operate and do business.[1] The

[*] BSc/LLB (Hons I) (ANU), Visiting Fellow at Australian National University College of Law. The author would like to thank Associate Professor Philippa Ryan and the anonymous reviewers and editors whose comments strengthened this article. This article reflects the author's personal views.

[1] While much of the underpinning logic behind AI and MLAs has been understood since the 1970s, it is the rapid advancement in computing power, combined with increasing data gathering and analysis capabilities that is driving the growth in AI: see Andrea Zanella et al, 'Internet of Things for Smart Cities' (2014) 1(1) *Internet of Things Journal* 22; Monika Zalnieriute, Lyria Bennett Moses and George Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82(3) *Modern Law Review* 425; Will Bateman, 'Algorithmic Decision-Making and Legality: Public Law Dimensions' (2020) 94(7) *Australian Law Journal* 520. Given the rapidly moving field of technology law (and deepfake technology in particular), this article draws on grey literature to supplement peer-reviewed research. For discussion on grey literature in the context of evolving medical technology, see Louisa Degenhardt et al, 'Searching the Grey Literature

interaction between AI and law, and the day-to-day operation of government are posing unique challenges, given the speed at which AI operates and the threat it presents to accountability and transparency of government. This has been demonstrated in an Australian context through the challenges driven by automated decision-making,[2] including the ongoing Robodebt debacle.[3] While much has been written about the application of AI to government,[4] little analysis has been conducted regarding the threat AI poses to elections, and by extension to the foundations of representative democracies. In turn, this means few protections are available to combat this threat.

This article outlines the limitations of existing law as applied to the emerging problem of 'political deepfakes', a subtype of political disinformation. Deepfakes are videos created using AI, which allow creators to superimpose images and audio from one video to another.[5] In effect, deepfake technology allows a user to create a fake video of a person saying or doing almost anything, only limited by their creativity and the footage of the subject they can source. Regulating deepfakes poses unique challenges in an Australian context through the operation of the implied freedom of political communication. Similarly, there remain significant challenges when designing regulations to ensure that speech is not overburdened and that regulations are proportionate and tailored to the threat they seek to prevent.

This article proceeds in four parts. Part II analyses the threat posed to Australian elections by political deepfakes. Parts III and IV explore current private and public remedies available to legitimate political actors and the Australian Electoral Commission ('AEC') to combat political deepfakes. The insufficiency of these available remedies to mitigate the harms caused by political deepfakes is then examined. Part V proposes legislative reform via a model law that could be enacted by the Commonwealth, state and territory governments to combat political deepfakes. In doing so, the article recommends against broader regulation of misinformation and disinformation which may lead to a significant chilling effect on political communication.

to Access Research on Illicit Drug Use, HIV and Viral Hepatitis' (Technical Report No 334, National Drug and Alcohol Research Centre, University of New South Wales, 2016).

[2] Andrew Ray, 'Implications of the Future Use of Machine Learning in Complex Government Decision-Making in Australia' (2020) 1(1) *Australian National University Journal of Law and Technology* 4.

[3] Richard Glenn, Acting Commonwealth Ombudsman, 'Centrelink's Automated Debt Raising and Recovery System' (Report No 2, April 2017) 7–8 [3.2]–[3.6] <https://www.ombudsman.gov.au/__data/assets/pdf_file/0022/43528/Report-Centrelinks-automated-debt-raising-and-recovery-system-April-2017.pdf>; Order of Davies J in *Amato v Commonwealth* (Federal Court of Australia, VID611/2019, 27 November 2019). The settlement was approved by the Federal Court in *Prygodicz v Commonwealth [No 2]* [2021] FCA 634; however, accountability issues remain as the opposition pushes for review of the decisions leading to the class action.

[4] See, eg, Zalnieriute, Bennett Moses and Williams (n 1).

[5] Kristina Libby, 'Deepfakes Are Amazing. They're Also Terrifying for Our Future', *Popular Mechanics* (online, 13 August 2020) <https://www.popularmechanics.com/technology/security/a28691128/deepfake-technology/>.

II DEEPFAKES AND DEMOCRACIES

In the context of elections, AI combined with key datasets (commonly referred to as Big Data) is being used by political parties to better target swing voters and to assess the palatability of policy positions.[6] Similarly, electoral agencies are using algorithms to manage the increasingly complex process of counting votes.[7] These algorithms are not subject to public scrutiny.[8] While these issues are concerning, the threats they pose can largely be mitigated through open, fair and transparent electoral processes. This is because electoral agencies are responsible to Parliament, and therefore the population can decide whether the actions of political parties (and the AEC) should be punished at the ballot box.[9] It is therefore the influence of AI on the conduct and results (rather than the management) of elections that is the primary focus of this article.

A Political Deepfakes

The use of AI technologies represents a significant and growing threat to electoral security. In particular, deepfake technology when deployed by experts can create videos of politicians so realistic they cannot be distinguished from a real video by humans or computers designed to detect them.[10] Deepfakes are created using 'neural networks that analyze large sets of data … to learn to mimic a person's facial expressions, mannerisms, voice, and inflections'.[11] By way of a popular example, similar technology was used to create scenes in which the late Carrie Fisher appeared in the recent Star Wars film: *Rogue One*.[12]

Historically, individuals wishing to make a useful (or, perhaps more accurately described, *undetectable*) deepfake, required hundreds of images of their 'subject' to train an MLA.[13] However, recent advances in technology have meant that only

[6] Jennifer Lees-Marshment et al, 'Vote Compass in the 2014 New Zealand Election' (2015) 67(2) *Political Science* 94.

[7] Ben Raue, 'Looking Out for No 1: Why the Senate Vote Count Needs Greater Transparency', *The Guardian* (online, 20 July 2016) <https://www.theguardian.com/australia-news/2016/jul/20/looking-out-for-no-1-why-the-senate-vote-count-needs-greater-transparency>.

[8] *Cordover and Australian Electoral Commission (Freedom of information)* [2015] AATA 956 (11 December 2015); Ray (n 2) 13–14.

[9] Brian Galligan, 'Parliamentary Responsible Government and the Protection of Rights' (Papers on Parliament No 18, Parliament of Australia, December 1992).

[10] Mika Westerlund, 'The Emergence of Deepfake Technology: A Review' (2019) 9(11) *Technology Innovation Management Review* 39, 45–6.

[11] Ibid 40.

[12] Erin Winick, 'How Acting as Carrie Fisher's Puppet Made a Career for Rogue One's Princess Leia', *MIT Technology Review* (online, 16 October 2018) <https://www.technologyreview.com/2018/10/16/139739/how-acting-as-carrie-fishers-puppet-made-a-career-for-rogue-ones-princess-leia/>. In an Australian context, fans have inserted the Joker into *A Knight's Tale* (Columbia Pictures, 2001): Ben Gilbert, 'An Incredible Series of Videos Swap Famous Hollywood Faces to Demonstrate How Convincing "Deepfake" Tech Has Gotten: Take a Look', *Business Insider Australia* (online, 31 May 2019) <https://www.businessinsider.com.au/deepfakes-of-famous-movies-youtube-channel-2019-5?r=US&IR=T>.

[13] See, eg, Supasorn Suwajanakorn, Steven M Seitz and Ira Kemelmacher-Shlizerman, 'Synthesizing Obama: Learning Lip Sync from Audio' (2017) 36(4) *ACM Transactions on Graphics* 1.

a small number of images are required to generate realistic videos of the subject.[14] This, combined with the fact that videos shot front-on in consistent light are the easiest to replicate,[15] makes political figures a ripe target for deepfakes. This is due to the wide availability of footage of political figures in which they are positioned forward-facing, under similar lighting conditions.[16] This ease of creation is demonstrated by the fact that deepfakes can now be created on a smartphone, using only a few images of the intended subject.[17]

The targeting of politicians with deepfake technology is more than an academic hypothesis. Indeed, deepfakes have been made featuring Donald Trump,[18] Barack Obama,[19] Manoj Tiwari,[20] Vladimir Putin[21] and Sophie Wilmès.[22] These examples, while well-known, are not exhaustive. The targeting of then Belgian Prime Minister Sophie Wilmès by Extinction Rebellion[23] in mid-2020 is of particular concern as it appears to be the *first* adverse targeting of a politician: previous examples of political deepfakes were generally educational, comedic or satirical.[24] The video in question, which showed Wilmès giving a fictitious speech about the link between COVID-19 and climate change, was widely shared on social media. Critically, at least some users were tricked into believing the video was real.[25] Regardless of whether you agree with the motivation behind the video, the use of deepfake technology to falsely attribute a speech to an elected Prime Minister is of grave concern.

[14] Egor Zakharov et al, 'Few-Shot Adversarial Learning of Realistic Neural Talking Head Models', *arXiv* (submitted 20 May 2019, revised 25 September 2019) <https://arxiv.org/abs/1905.08233>.

[15] 'How to Create the Perfect DeepFakes', *Alan Zucconi* (Blog Post, 14 March 2018) <https://www.alanzucconi.com/2018/03/14/create-perfect-deepfakes/>.

[16] For example, politicians regularly appear at press conferences and in news segments where they are often filmed looking directly at the camera in a well lit environment.

[17] See, eg, NEOCORTEXT, INC., 'Reface: Face Swap Videos', *Apple App Store* (Application, 2020) <https://apps.apple.com/app/id1488782587>.

[18] Helena Skinner, 'French Charity Publishes Deepfake of Trump Saying "AIDS is over"', *Euronews* (online, 9 October 2019) <https://www.euronews.com/2019/10/09/french-charity-publishes-deepfake-of-trump-saying-aids-is-over>.

[19] James Vincent, 'Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA about Fake News', *The Verge* (online, 17 April 2018) <https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed>.

[20] Regina Mihindukulasuriya, 'Why the Manoj Tiwari Deepfakes Should Have India Deeply Worried', *The Print* (online, 29 February 2020) <https://theprint.in/tech/why-the-manoj-tiwari-deepfakes-should-have-india-deeply-worried/372389/>. This video differs from the other examples as it was made by the subject to help them communicate to voters with different language backgrounds.

[21] Karen Hao, 'Deepfake Putin Is Here to Warn Americans about Their Self-Inflicted Doom', *MIT Technology Review* (online, 29 September 2020) <https://www.technologyreview.com/2020/09/29/1009098/ai-deepfake-putin-kim-jong-un-us-election/>.

[22] 'The Truth about COVID-19 and the Ecological Crisis: A Speech for Sophie Wilmès', *Extinction Rebellion Belgium* (Web Page, April 2020) <https://www.extinctionrebellion.be/en/tell-the-truth>.

[23] Ibid.

[24] Westerlund (n 10) 43.

[25] Gerald Holubowicz, 'Extinction Rebellion S'empare des Deepfakes en Belgique' [Extinction Rebellion Takes over Deepfakes in Belgium], *Mediapart* (Blog Post, 15 April 2020) <https://blogs.mediapart.fr/geraldholubowicz/blog/150420/extinction-rebellion-s-empare-des-deepfakes-en-belgique>.

B Impact on Elections

This article will focus on two primary threats posed to elections by deepfakes: the use of deepfakes to alter voter preferences, and the impact of deepfakes on trust generally in elections and democratic institutions.[26] First, through their potential impact on voter preferences, deepfakes may be used to obfuscate or undermine a politician's (or political party's) stance on a given issue, or to target their credibility. Given the shift to longer periods of pre-polling in Australia (and other democracies),[27] the release of a deepfake within this period or just before election day will make it extremely challenging for politicians to respond before any votes are cast. For example, a deepfake of a politician with a strong anti-drug platform consuming an illicit drug could be both impactful, and difficult to disprove.[28] A deepfake could be made as part of a candidate's official campaign, by an overseas actor attempting to sway an election, or even by an individual disconnected from the political process.

While there is no evidence that deepfakes have impacted an Australian election to date, compromising (albeit true) video footage has previously led to federal candidates dropping out of an electoral race.[29] Meanwhile, doctored footage has been used in the United States ('US') by the Republican Party to attack House Speaker Nancy Pelosi by slowing down real video clips of her speeches to slur her words and make her appear drunk.[30] Similar videos were also used to target President Joe Biden in the 2020 Presidential election, with experts warning prior to the election that the worst was yet to come as 'cutting-edge methods such as deepfakes are best suited to … predictable moment[s] of public uncertainty'.[31] Such a moment, they posited, would occur following the election, with Trump hinting

[26] Secondary threats could include undermining diplomacy and jeopardising national security. These threats can be viewed as subsidiary to the primary threats identified above in that they rely on either convincing a particular actor a fake video is real or in eroding public trust in video content, for example, fake news about nuclear attacks could cause general panic and reduce trust in future warnings.

[27] Stephen Mills and Martin Drum, 'Surge in Pre-poll Numbers at 2019 Federal Election Changes the Relationship between Voters and Parties', *The Conversation* (online, 19 August 2019) <https://theconversation.com/surge-in-pre-poll-numbers-at-2019-federal-election-changes-the-relationship-between-voters-and-parties-121929>. This trend has increased in recent elections: Damon Muller, 'Trends in Early Voting in Federal Elections', *Parliament of Australia* (Web Page, 8 May 2019) <https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/FlagPost/2019/May/Trends_in_early_voting_in_federal_elections>.

[28] Further possibilities could include footage of candidates withdrawing from a race and endorsing another candidate, a politician committing an offence, accepting a bribe, or outlining a fake policy position. Given the ease of use of the technology, users are limited only by their creativity.

[29] Josh Bavas, 'One Nation Election Candidate Steve Dickson Resigns over Strip Club Videos', *ABC News* (online, 30 April 2019) <https://www.abc.net.au/news/2019-04-30/one-nation-candidate-steve-dickson-quits-over-strip-club-video/11056676>.

[30] Hannah Denham, 'Another Fake Video of Pelosi Goes Viral on Facebook', *The Washington Post* (online, 3 August 2020) <https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/>.

[31] Clint Watts and Tim Hwang, 'Deepfakes Are Coming for American Democracy: Here's How We Can Prepare', *The Washington Post* (online, 10 September 2020) <https://www.washingtonpost.com/opinions/2020/09/10/deepfakes-are-coming-american-democracy-heres-how-we-can-prepare/>.

that he would not accept electoral defeat.[32] That set of circumstances unfolded partly as predicted with Trump declaring the election results 'fake news' and his supporters storming the Capitol in circumstances condemned as terrorism by US security agencies.[33] There was however no detectable use of deepfake videos, with the potential for a faked video of then President-elect Biden accepting 'defeat' remaining only a possibility. It is noteworthy that despite public institutions, inquiries and courts all labelling the fraud claims false, Trump and the Republican Party more broadly continue to push the electoral fraud claims publicly.

1 Changing Voter Preferences

Exactly how many voters could be misled by a deepfake remains unclear. However, if marginal seats were targeted during an election, even swaying as few as 100 voters could be impactful.[34] In this context, a 2020 study found that approximately 15% of viewers in a controlled trial believed a deepfake of Obama was real.[35] While it is unlikely that everyone who believes a deepfake will alter their vote because of it (in part due to the strength of party allegiance),[36] the possibility should not be discounted. Indeed, it may not be necessary for voters to alter their vote for a deepfake video to impact an election. For example, deepfake videos could force candidates to withdraw or impact a candidate's or party's fundraising ability – these results themselves having an indirect effect on electoral outcomes. Further, while some authors have found that disinformation generally has little direct impact on elections,[37] disinformation has been shown to have (at least some) impact in Australian elections. For example, the Australian Labor Party acknowledged the impact of the (false) 'death tax' ads on its 2019 campaign, although they accepted that this alone did not decide the election.[38] Additionally, while disinformation (and specifically, in the context of this article, the use of deepfakes) may not alter which party secures a majority of seats, it may play a larger role in deciding *individual* electoral contests. This is especially the case with deepfakes, where, as discussed

[32] 'Donald Trump Refuses to Commit to Peaceful Transfer of Power if He Loses US Election', *ABC News* (online, 24 September 2020) <https://www.abc.net.au/news/2020-09-24/donald-trump-wont-commit-to-transfer-of-power-after-election/12696786>.

[33] See generally 'FBI Chief Calls Capitol Attack Domestic Terrorism and Rejects Trump's Fraud Claims', *The Guardian* (online, 11 June 2021) <https://www.theguardian.com/us-news/2021/jun/10/capitol-attack-fbi-christopher-wray-congress>.

[34] For example, in the 2020 Northern Territory election 11/25 seats would have changed hands if 100 voters had been swayed by a deepfake: 'NT Summary of Two Candidate Preferred Votes by Division', *Northern Territory Electoral Commission* (Web Page, 2020) <https://ntec.nt.gov.au/elections/2020-territory-election/results/nt-summary-of-two-candidate-preferred-votes-by-division>. The average turnout for each division was 4,235 voters, so swaying ~2.5% of voters could have altered 11/25 contests.

[35] Cristian Vaccari and Andrew Chadwick, 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News' (2020) 6(1) *Social Media + Society* 1, 6.

[36] Spencer McKay and Chris Tenove, 'Disinformation as a Threat to Deliberative Democracy' (2020) (July) *Political Research Quarterly* 1, 1.

[37] Ibid. However, the authors went on to assess other harms that disinformation may pose, including degrading trust in media organisations and academic think tanks.

[38] See, eg, Craig Emerson and Jay Weatherill, 'Review of Labor's 2019 Federal Election Campaign' (Report, 7 November 2019) 79–80.

above, it is possible for actors to target individual politicians by, for example, creating a deepfake of them engaging in illegal conduct. In this context, critically, at a federal level Australia remains vulnerable to targeted attacks: 36 lower house seats are currently held by a margin of less than 5%, 84 by less than 10% and 129 by less than 15%.[39]

It is however the secondary threat that is likely of greater concern. In addition to the percentage who believed the deepfake was real, the 2020 study found that only 50.8% of the participants were *not* deceived by the video.[40] The remainder were *unable* to determine if the video was real or fake. It is this segment of individuals that highlights the second threat posed by deepfakes to elections: a reduction in trust in video footage and news impacting our perception of democracy more broadly.

2 Decreasing Trust in Democracy and Democratic Institutions

Increasingly, Australians are turning to digital platforms such as Facebook to access news content.[41] This mirrors a global trend towards accessible and shareable content,[42] which is making it easier for fake news to be distributed widely. The shift to digital content has coincided with decreasing trust in politicians and politics in general.[43] Political deepfakes will further erode trust by allowing candidates to deride real footage as fake news, feeding into increasing claims by politicians that they have been set up.[44] It is this threat that most alarms political scientists as, after all, threats to a single election are of themselves a threat to democracy.[45] However, the rise of disinformation more broadly has the capacity to fundamentally undermine 'truth' in elections with disastrous consequences. For example, in the US, disproven rumours of electoral fraud are supporting a wave of electoral reforms that will make it harder to vote to 'safeguard' future elections.[46] These laws

[39] Corresponding to 24%, 56% and 85% of lower house seats accordingly. Analysis conducted on AEC data from the recent 2019 federal election and 2020 Eden-Monaro by-election: Australian Electoral Commission, 'Seat Summary', *Tally Room 2019 Federal Election* (Web Page, 2019) <https://results.aec.gov.au/24310/Website/HouseSeatSummary-24310.htm> (results on file with author).

[40] This was described as 'surprising given the statement [an unsophisticated insult about Donald Trump] was highly improbable': Vaccari and Chadwick (n 35) 6.

[41] Australian Competition and Consumer Commission, 'Digital Platforms Inquiry' (Final Report, June 2019) ch 1; See also Christopher Hughes, 'News Sources in Australia in 2021', *Statista* (online, 12 July 2021) <https://www.statista.com/statistics/588441/australia-news-sources/>.

[42] Katie Elson Anderson, 'Getting Acquainted with Social Networks and Apps: Combating Fake News on Social Media' (2018) 35(3) *Library Hi Tech News* 1.

[43] Simon Tormey, 'The Contemporary Crisis of Representative Democracy' (Papers on Parliament No 66, Parliament of Australia, October 2016) 90 <https://www.aph.gov.au/About_Parliament/Senate/Powers_practice_n_procedures/pops/Papers_on_Parliament_66/The_Contemporary_Crisis_of_Representative_Democracy>; Russell J Dalton, *Democratic Challenges, Democratic Choices: The Erosion of Political Support in Advanced Industrial Democracies* (Oxford University Press, 2004).

[44] See, eg, comments made by then President Donald Trump during the 2020 election: David Smith, 'Wounded by Media Scrutiny, Trump Turned a Briefing into a Presidential Tantrum', *The Guardian* (online, 14 April 2020) <https://www.theguardian.com/us-news/2020/apr/13/trump-coronavirus-meltdown-media-authority>.

[45] McKay and Tenove (n 36).

[46] Sam Levine, 'The Republicans' Staggering Effort to Attack Voting Rights in Biden's First 100 Days', *The Guardian* (online, 28 April 2021) <https://www.theguardian.com/us-news/2021/apr/28/republicans-voter-suppression-biden-100-days>.

have been held constitutional by the US Supreme Court,[47] and may, along with gerrymandering, decide the outcome of future elections alone notwithstanding for whom people vote on voting day. Deepfakes may exacerbate these underlying issues and cause distrust amongst voters themselves who may not know *whom* or *what* they can actually trust, allowing lawmakers to pass anti-democratic laws to 'safeguard' elections.

These threats are not insignificant, especially as deepfakes can be generated and shared from within or outside of Australia by anyone with a desktop computer or smartphone.[48] It is this accessibility that makes the threat most concerning, as once the videos have been created and shared, they can be re-uploaded rapidly making it almost impossible for them to be taken down (even if proven false). For example, the widely discredited video *Plandemic* was repeatedly re-uploaded to alternative hosting sites after being taken down by Facebook and YouTube, with commentators suggesting the attempt to shut down the video led to it being viewed by a wider audience.[49]

C Increasing Challenge of Electoral Interference

The threat posed by deepfakes is heightened by the increasing level of foreign interference in elections. The threat posed by foreign actors is unique, in that they can operate outside a target jurisdiction, while still being able to spread fake news through social media. This rise in foreign interference both increases the likelihood that deepfakes will be used and makes them harder to combat due to limitations of domestic law. Despite these limitations, difficulties in attributing disinformation to a state mean that domestic regulations are likely more useful than pursuing action internationally.[50]

Foreign interference impacted the outcome of the 2016 US Presidential election,[51] and has been of increasing concern to the Australian Government. For example, the Government has recently launched Senate inquiries into foreign interference,[52] proposed a widening of the Australian Security Intelligence Organisation's powers

[47] *Brnovich v Democratic National Committee*, 594 US ___ (2021). For commentary: see, eg, Lauren Fedor, 'US Supreme Court Upholds Arizona Law in Voting Rights Challenge', *Financial Times* (online, 2 July 2021) <https://www.ft.com/content/35e67872-e1eb-449d-8745-3d0c13db1526>.

[48] Best results require a mid-high end graphics card: Timothy B Lee, 'I Created My Own Deepfake: It Took Two Weeks and Cost $552', *ARS Technica* (online, 16 December 2019) <https://arstechnica.com/science/2019/12/how-i-created-a-deepfake-of-mark-zuckerberg-and-star-treks-data/>.

[49] Andrea Bellemare, Katie Nicholson and Jason Ho, 'How a Debunked COVID-19 Video Kept Spreading after Facebook and YouTube Took It Down', *CBC News* (online, 21 May 2020) <https://www.cbc.ca/news/technology/alt-tech-platforms-resurface-plandemic-1.5577013>.

[50] Björnstjern Baade, 'Fake News and International Law' (2019) 29(4) *European Journal of International Law* 1357, 1361–2. This article will therefore focus on domestic rather than international law.

[51] United States Senate Select Committee on Intelligence, *Russian Active Measures Campaigns and Interference in the 2016 US Election* (Report, 2020) vol 5 <https://www.intelligence.senate.gov/sites/default/files/documents/report_volume5.pdf>; 'Russia Worked to Help Trump in 2016 Election: Senate Panel', *Aljazeera* (online, 18 August 2020) <https://www.aljazeera.com/news/2020/8/18/russia-worked-to-help-trump-in-2016-election-senate-panel>. The US federal government has implemented laws encouraging research deepfakes but is yet to legislate to directly combat the threat: *National Defense Authorization Act for Fiscal Year 2020*, Pub L No 116-92, §§ 5709, 5724, 133 Stat 1790 (2019).

[52] The Senate Select Committee on Foreign Interference through Social Media was established in 2019: 'Select Committee on Foreign Interference through Social Media', *Parliament of Australia* (Web Page)

to investigate foreign interference,[53] and passed sweeping new laws to target the same in state governments and at universities.[54] Meanwhile, the link between foreign interference and political deepfakes has been highlighted by academic commentators in submissions to both parliamentary and departmental inquiries.[55] Commentators have also highlighted the need for anticipatory reform, particularly given that elections generally cannot be 'redone' without overcoming significant legal hurdles.[56] In the absence of a new election, there is no practical remedy a court could offer post-election once a deepfake has been viewed. Reform is therefore needed *prior* to any impact on an Australian election. This is especially the case as the use of deepfakes may benefit a particular political party (whether or not they supported the use of the technology) and that party may then be unwilling to support a review into the impact of deepfake technology on their electoral victory.

D The Need for Law to Capture (and Combat) Political Deepfakes

Protection against deepfakes cannot be left to the social media platforms on which they are shared. While some platforms have developed policies to combat deepfakes,[57] this type of remedy is insufficient for three reasons. First, even where a video is removed by the platform this does not necessarily counter the harm, and without legal powers to compel the social media platforms, an affected party cannot seek a retraction or public recognition that the video was fake. Second, not all social media companies' current disinformation policies address deepfakes, nor is there a guarantee that existing policies are sustainable. Third, definitions of 'deepfake' may vary between social media platforms and may not capture *all* videos that have been edited to mislead viewers – for example, current disinformation policies do not capture the Nancy Pelosi example discussed above.[58] In order to

<https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Foreign_Interference_through_Social_Media>.

[53] Australian Security Intelligence Organisation Amendment Bill 2020 (Cth).

[54] See, eg, Australia's Foreign Relations (State and Territory Arrangements) Bill 2020 (Cth); Australia's Foreign Relations (State and Territory Arrangements) (Consequential Amendments) Bill 2020 (Cth).

[55] News and Media Centre University of Canberra and the Virtual Observatory for the Study of Online Networks Australian National University, Submission No 8 to Senate Select Committee on Foreign Interference through Social Media, Parliament of Australia, *Foreign Interference through Social Media* (2020) 3; The Allens Hub for Technology, Law and Innovation, Submission No 2 to Department of Foreign Affairs and Trade, Australian Government, *International Cyber and Critical Technology Engagement Strategy* (16 June 2020) 2.

[56] In the US context the Supreme Court has blocked recounts in close presidential races: *Bush v Gore*, 531 US 98 (2000); Jack M Balkin, 'Bush v. Gore and the Boundary between Law and Politics' (2001) 110(8) *Yale Law Journal* 1407; Richard Posner, 'Bush v Gore: Prolegomenon to an Assessment' (2001) 68(3) *University of Chicago Law Review* 719, 736. Subsequent analysis revealed that Gore should have won Florida and the presidential election had a *state-wide* review of all contested ballots been conducted. However, this was not the remedy Gore had sought: Wade Payson-Denney, 'So, Who Really Won? What the Bush v. Gore Studies Showed', *CNN* (online, 31 October 2015) <https://edition.cnn.com/2015/10/31/politics/bush-gore-2000-election-results-studies/index.html>.

[57] Aaron Holmes, 'Facebook Just Banned Deepfakes, but the Policy Has Loopholes – And a Widely Circulated Deepfake of Mark Zuckerberg Is Allowed to Stay Up', *Business Insider* (online, 8 January 2020) <https://www.businessinsider.com/facebook-just-banned-deepfakes-but-the-policy-has-loopholes-2020-1?r=AU&IR=T>.

[58] 'Facebook Refuses to Remove Doctored Nancy Pelosi Video', *The Guardian* (online, 4 August 2020) <https://www.theguardian.com/us-news/2020/aug/03/facebook-fake-nancy-pelosi-video-false-label>.

ensure consistent, and therefore fair, treatment of political deepfakes, measures must be captured in law rather than left to discretionary company policy. This approach also ensures that Parliament can set appropriate limits on what type of videos are or are not captured by the law, and tailor appropriate exemptions.

III EVALUATION OF PRIVATE PROTECTIONS

This Part analyses the scope of current Australian laws and regulations to combat deepfakes, and the *private* remedies that are available to the subjects of a deepfake. *Public* remedies will be discussed in Part IV. This Part explores two general areas of private law: copyright law and tort law. These feature in the bulk of analysis by US commentators who have considered the legal options currently afforded to individuals who are the subject of a deepfake. Such commentary is, however, often relatively brief, forming only a small part of a larger article.[59] Additionally, little analysis has, to date, been conducted in an Australian context.

Before embarking on this analysis, it is worth noting some general points. Intellectual property and tort law provide private remedies allowing victims to bring personal actions to have deepfakes taken down, and to seek damages for any loss or injury they have suffered. Electoral regulations, discussed in Part IV, instead form a hybrid private-public remedy given the work of both the AEC and political parties and candidates in enforcing electoral regulations. The relevance of this distinction will be discussed when analysing a possible remedy, but ultimately the identity of the person bringing the action, and the speed at which they can do so are critical in the context of political deepfakes. This is because, as adverted to above, damages are unlikely to be an appropriate remedy for cases involving political deepfakes. Instead, the preferred remedy is the removal of the deepfake in a timely manner, so as to avoid any adverse impact on a politician's performance in an election.[60] More simply put, it is impossible to put a price on political power.

A Copyright Law

Copyright law has been suggested by some commentators as a potential solution to the threat posed by deepfakes.[61] In a recent high profile example, the US reality television stars 'the Kardashians' were successful in an action to remove a deepfake from YouTube using existing copyright infringement procedures.[62] The

[59] See, eg, Edvinas Meskys et al, 'Regulating Deep Fakes: Legal and Ethical Considerations' (2020) 15(1) *Journal of Intellectual Property Law & Practice* 24, 29.

[60] See, eg, the concern raised at the 2019 federal election about the use of signs that mimic AEC colours: Paul Karp, 'Oliver Yates May Take Liberals to Court of Disputed Returns over "Deceptive" Election Signs', *The Guardian* (online, 21 May 2019) <https://www.theguardian.com/australia-news/2019/may/21/oliver-yates-may-take-liberals-to-court-of-disputed-returns-over-deceptive-election-signs>.

[61] Meskys et al (n 59) 29.

[62] Mathew Katz, 'Kim Kardashian Can Get a Deepfake Taken off YouTube. It's Much Harder for You', *Digital Trends* (online, 17 June 2019) <https://www.digitaltrends.com/social-media/kim-kardashian-deepfake-removed-from-youtube/>. The original footage used in the video was featured in *Vogue*.

deepfake is however still accessible on other platforms including Instagram.[63] Given its potential, this section explores the application of Australian intellectual property law to political deepfakes by analysing copyright subsistence, before addressing infringement, exceptions and limitations of copyright law.

Deepfakes pose a number of challenges to copyright law, including the novel question about whether copyright would, or *should*, subsist in the final work. This is important, as, if copyright subsists in a deepfake, laws that purported to strip this copyright may raise issues surrounding the acquisition of property on just terms.[64] Laws that merely regulated the *use* of the videos would however not be limited.[65] Given the requirement for human authorship for copyright to subsist in a work under Australian copyright law,[66] it is likely that copyright would *not* currently subsist in deepfakes.[67] This does not, however, mean that creators will not be liable if they infringe on another's copyright.

1 Subsistence of Copyright

In assessing whether copyright subsists in a work, a court needs to assess whether the work is original. This is a question of fact,[68] which requires courts to determine whether a *human* author exercised 'independent intellectual effort' in the production of the material work.[69] In *Telstra Corporation Ltd v Phone Directories Co Pty Ltd*, the Federal Court applied this test to a written work created through a largely automated process, finding that copyright did not subsist in the resulting work.[70] In discussing how the test applied to computer programs and automated processes, Perram J stated:

> So long as the person controlling the program can be seen as directing or fashioning the material form of the work there is no particular danger in viewing that person as the work's author. … [However] the performance by a computer of functions ordinarily performed by human authors will mean that copyright does not subsist in the work …[71]

[63] Ibid.

[64] *JT International SA v Commonwealth* (2012) 250 CLR 1.

[65] The key issue being whether an interest, benefit or advantage of a proprietary nature is acquired by the Commonwealth or another party: ibid.

[66] *Copyright Act 1968* (Cth) s 32(1); *IceTV Pty Ltd v Nine Network Australia Pty Ltd* (2009) 239 CLR 458, 493–6 [95]–[106] (Gummow, Hayne and Heydon JJ) ('*IceTV*'); *Telstra Corporation Ltd v Phone Directories Co Pty Ltd* (2010) 194 FCR 142 ('*Phone Directories*'); Sam Ricketson, 'The Need for Human Authorship: Australian Developments: *Telstra Corp Ltd v Phone Directories Co Pty Ltd*' (2012) 34(1) *European Intellectual Property Review* 54; Dilan Thampapillai, 'If Value Then Right? Copyright and Works of Non-human Authorship' (2019) 30(2) *Australian Intellectual Property Journal* 1; Dilan Thampapillai, 'The Gatekeeper Doctrines: Originality and Authorship in the Age of Artificial Intelligence' (2019) 10 *WIPO-WTO Colloquium Papers* 1.

[67] There are however open questions regarding whether the provisions of the *Copyright Act 1968* (Cth) should be amended to capture works created through automated processes. Similar amendments were made in the United Kingdom: *Copyright, Designs and Patents Act 1988* (UK) s 9.

[68] *IceTV* (2009) 239 CLR 458, 494–5 [99] (Gummow, Hayne and Heydon JJ).

[69] *Sands & McDougall Pty Ltd v Robinson* (1917) 23 CLR 49, 52 (Isaacs J).

[70] *Phone Directories* (2010) 194 FCR 142.

[71] Ibid 178–9 [118]. The other members of the Court made similar statements: see 171 [89]–[90] (Keane CJ), 191 [169] (Yates J).

To determine whether copyright subsists in a deepfake, a court will need to determine whether this test should extend to artistic works. It is likely that a court would find this to be the case. Relevantly, artistic works are afforded *less* protection than literary works in the *Copyright Act 1968* (Cth) ('*Copyright Act*'),[72] and the Act does not distinguish between literary and artistic works in terms of the requirement for originality.[73] In applying the test, a court would need to determine the extent to which a person operating a neural network to create a deepfake 'direct[ed] or fashion[ed] the material [final] form of the work'.[74] This question is complex as an individual is involved at various stages of the process, including: selecting the images used to train the neural net, deciding when the neural net is ready, and selecting the video to 'swap' the face onto. Despite this involvement, it is the trained neural net that performs most of the decision-making. It is therefore likely that in Australia, copyright *would not* subsist in a deepfake.

2 Copyright Infringement

The question of copyright infringement is distinct from whether copyright subsists in a work. The *Copyright Act* prevents individuals who do not own the copyright in a particular work from 'the doing in Australia of, any act comprised in the copyright'.[75] The burden of proving infringement lies on the copyright holder. In the context of *political* deepfakes, it is likely that, in creating a deepfake, an individual will draw on news content, as this is where video and audio of politicians are most accessible. Notably, the protections afforded to television and sound broadcasts by the *Copyright Act* are not as extensive as those afforded to artistic works.[76] Nonetheless the Act still prohibits the communication of sound recordings and television broadcasts to the public.[77] This could prima facie be established where news footage was used to create a deepfake. To establish infringement, a copyright holder must prove that the works are objectively similar, there was a causal connection between the original work and the infringing work, and that a substantial part of the copyright work was infringed.[78]

How these tests will apply to deepfakes has not yet been resolved by a court or explained in existing academic literature. What is clear is that there will be significant challenges in applying the tests due to the 'black box' nature of machine-learning systems. This nature means that while the inputs to the system are known (ie, the training data and the video into which the face will be swapped), the precise steps it takes to create the deepfake are not.[79] It will therefore be unclear precisely

[72] For comparison, see ss 31(1)(a), 31(1)(b).

[73] Ibid s 32(1): 'copyright subsists in an original literary, dramatic, musical or artistic work'.

[74] *Phone Directories* (2010) 194 FCR 142, 178 [118] (Perram J).

[75] *Copyright Act 1968* (Cth) s 36(1).

[76] Ibid s 87.

[77] Ibid ss 85, 87.

[78] See, eg, *Elwood Clothing Pty Ltd v Cotton On Clothing Pty Ltd* (2008) 172 FCR 580, 588 [41] (the Court). In assessing whether a substantial part of the work was infringed, what is relevant is the quality of the work. This requires an assessment of the independent intellectual effort put into the relevant material: *IceTV* (2009) 239 CLR 458, 479 [49]–[50] (French CJ, Crennan and Kiefel JJ).

[79] This was discussed in relation to automated decision-making and the resulting transparency and accountability issues that arise: Bateman (n 1).

what material was used by the machine-learning system, or the extent to which it is replicated in the final form of the deepfake. This will pose significant challenges to copyright holders (the news companies) – especially given that they will need to demonstrate that a substantial part of *their* work was infringed. Where a deepfake swaps a face into a video clip owned by a single copyright holder this issue would not arise.[80] However, if the deepfake creator stages their own scene and merely swaps an individual's face or voice into this video (using a compilation of other copyright holders' work to perform the face swap), then establishing infringement will be complex. Further, the potential compensation that would be awarded to an individual copyright holder would likely be small,[81] making bringing an action (and bearing the resulting risk of an adverse costs order) unattractive.

3 Copyright Exemptions

In addition to the challenge of establishing infringement, in certain cases, deepfake creators or distributers may be able to avail themselves of exemptions in the *Copyright Act*. The Act provides an exemption where a work is a 'fair dealing … for the purpose of parody or satire'.[82]

While courts have historically used dictionaries to aid in statutory interpretation,[83] academic commentators have suggested that a broader definition of comedy and satire would give effect to the legislative intent behind the provisions.[84] These academic commentators have suggested that 'ordinary definitions', that is, the use of comedy and satire in practice, would better achieve the stated purpose of the exemptions: promoting 'free speech and Australia's fine tradition of satire by allowing our comedians and cartoonists to use copyright material for the purposes of parody or satire'.[85] Other academics have suggested that the exemption should be read broadly, with the primary test to be applied being whether the work 'adds significant new expression so as not to be substitutable for the original work'.[86] Regardless of the approach adopted by the courts, it is likely that at least some deepfakes could fall within a comedy and satire exemption, with many of them

[80] This is often the case for example with regard to pornographic deepfakes, where an individual's face is swapped into a video owned by a single entity.

[81] This is analogous to individual copyright infringement claims against individuals who pirate movies. Collectively the action is worth bringing but where courts limit the options of copyright holders, they may abandon the action: see, eg, *Dallas Buyers Club LLC v iiNet Ltd* (2015) 245 FCR 129. While Dallas Buyers Club LLC was successful in getting preliminary discovery over IP addresses, Perram J attached conditions relating to what could be communicated to the individuals identified to limit the possibility of 'speculative invoicing': at 148–9 [83]. The court later rejected the proposed letter in *Dallas Buyers Club LLC v iiNet Limited [No 3]* (2015) 327 ALR 695.

[82] *Copyright Act 1968* (Cth) s 103AA.

[83] The *Macquarie Dictionary* being preferred: Michael Kirby, 'Statutory Interpretation: The Meaning of Meaning' (2011) 35(1) *Melbourne University Law Review* 113, 124.

[84] Conal Condren et al, 'Defining Parody and Satire: Australian Copyright Law and Its New Exception' (2008) 13(3) *Media and Arts Law Review* 273 ('Defining Parody and Satire Part 1'); Conal Condren et al, 'Defining Parody and Satire: Australian Copyright Law and Its New Exception: Part 2: Advancing Ordinary Definitions' (2008) 13(4) *Media and Arts Law Review* 401.

[85] Commonwealth, *Parliamentary Debates*, House of Representatives, 19 October 2006, 2 (Philip Ruddock, Attorney-General) quoted in Condren et al, 'Defining Parody and Satire Part 1' (n 84) 274.

[86] Nicolas Suzor, 'Where the Bloody Hell Does Parody Fit in Australian Copyright Law?' (2008) 13(2) *Media and Arts Law Review* 218, 220.

made for the purpose of ridiculing or critiquing politicians using very little copyrighted material. If the broad approach is taken, deepfakes would not be viewed as 'substitutable' to the original work, with the creator/author effectively using collated images for tell their own story.

Ultimately, it is unlikely that the exemption would be determinative in the overall protection afforded by copyright law, but it is worth acknowledging that its utility would, at least in some cases, be limited by the fair dealing for comedy or satire exemption.

4 Limitations of Copyright Law

Whether an individual could prove that a deepfake infringed their copyright is uncertain given the black box nature of neural networks, and the possible application of the fair dealing exemptions. There are, however, additional limitations to the protection afforded by copyright law to political deepfakes as, quite often, the politician or political party will not be the relevant copyright holder. For example, politicians often give public speeches that are recorded by broadcasters and published online. The use of this footage to train a neural net, even if it did infringe copyright, would not provide a remedy to the politician or political party. At best, the politician could request that the relevant copyright holder(s) pursue the creator of the deepfake.

It is unclear whether media companies would be willing to pursue such action, as they suffer no real harm from the infringement, and may in fact see a benefit in terms of viewer engagement. Even if they did so, the length of this process would eliminate any utility to the politician. This is especially the case where deepfakes are published on the eve of an election. In that scenario, a politician's ability to respond to a deepfake may in fact be limited by electoral blackout laws. These laws bar television and radio electoral advertising close to elections.[87] As such, deepfakes communicated over social media would not be captured by the restrictions while politicians would be limited in how they could respond to disinformation in the deepfake. The blackout laws have previously been critiqued due to the inconsistent treatment of different forms of advertising, but amendments have not yet been proposed.[88] While, in the author's view, amendments equalising the treatment of different forms of political advertising are desirable, they will not, of themselves, address the challenge posed by political deepfakes. Further analysis of the blackout laws therefore is outside the scope of this article.

B Tort Law

There are two potential torts that may provide a remedy to the subjects of a political deepfake: defamation[89] and passing off. [90]

[87] *Broadcasting Services Act 1992* sch 2 s 3A.

[88] Jordan Guiao, 'Distorting the Public Square: Political Campaigning on Social Media Requires Greater Regulation' (Discussion Paper, Australia Institute, November 2019) 5.

[89] See, eg, Meskys et al (n 59) 26.

[90] Emma Perot and Frederick Mostert, 'Fake It Till You Make It: An Examination of the US and English Approaches to Persona Protection as Applied to Deepfakes on Social Media' (2020) 15(1) *Journal of Intellectual Property Law & Practice* 32, 35–6.

1 Defamation Law

Australian defamation law has evolved from statute passed by the New South Wales Legislative Council in 1847,[91] through to the adoption of a national uniform law.[92] This evolution has been accompanied by a significant increase in the number of defamation proceedings launched. Indeed, despite the common stereotype of the Australian larrikin, Australia is seen as the defamation capital of the world.[93] This growth has coincided with the rise of social media, and is driven by a significant number of low-value claims.[94] Given this, in terms of legal actions politicians may seek to rely on to combat deepfakes, defamation is a likely candidate. Australian politicians have regularly used defamation to try to remove content harmful to their reputations. For example, Pauline Hanson was successful in obtaining an injunction against the Australian Broadcasting Corporation preventing them from playing the satirical song 'Backdoor Man'. The injunction was upheld unanimously on appeal.[95]

Broadly, to succeed in an action for defamation, a plaintiff must prove that:

1. The material was published by the defendant;
2. It identified the plaintiff; and
3. The material is defamatory (that is, it contains one or more defamatory imputations).[96]

In relation to deepfakes, the first element will be heavily fact dependent. Where a deepfake is created and published by someone in Australia, the element will be clearly established. This may not be the case where the deepfake is created by an overseas actor. In such cases, it may be possible for an individual to bring an action against the social media platform on which the deepfake was published. Australia-based media companies have been found liable in defamation for material published to their public Facebook pages.[97] Similarly, Google has been held to be liable for defamatory material published as part of its search results.[98] This suggests that where political deepfakes defame politicians, there may already be a number of prospective defendants, including web platforms and media platforms that promulgate the content.

[91] Paul Mitchell, 'The Foundations of Australian Defamation Law' (2006) 28(3) *Sydney Law Review* 477.

[92] For discussion see Andrew T Kenyon, 'Six Years of Australian Uniform Defamation Law: Damages, Opinion and Defence Meanings' (2012) 35(1) *University of New South Wales Law Journal* 31.

[93] Matt Collins, 'Nothing to Write Home about: Australia the Defamation Capital of the World' (Speech, National Press Club, 4 September 2019). For analysis of the growth of low-scale cases see, eg, Centre for Media Transition, 'Trends in Digital Defamation: Defendants, Plaintiffs, Platforms' (Report, University of Technology Sydney, 2018) <http://s3.amazonaws.com/arena-attachments/1918329/e636f1839b7687241f593933d2770018.pdf?1521525181>.

[94] Centre for Media Transition (n 93). Recent amendments to defamation laws passed in some states aim to reverse this trend; however their impact is yet to be seen: see, eg, Defamation Amendment Bill 2020 (NSW). For discussion about the laws, see Michaela Whitbourn, 'Uniformity at Risk as Defamation Reforms Set to Start in Three States on July 1', *Sydney Morning Herald* (online, 1 April 2021) <https://www.smh.com.au/national/uniformity-at-risk-as-defamation-reforms-set-to-start-in-three-states-on-july-1-20210401-p57fu5.html>.

[95] *Australian Broadcasting Corporation v Hanson* [1998] QCA 306.

[96] *Radio 2UE Sydney Pty Ltd v Chesterton* (2009) 238 CLR 460, 467 (French CJ, Gummow, Kiefel and Bell JJ).

[97] *Fairfax Media Publications Pty Ltd v Voller* [2020] NSWCA 102. The NSW Court of Appeal decision was upheld on appeal by the High Court: *Fairfax Media Publications Pty Ltd v Voller* [2021] HCA 27.

[98] *Defteros v Google LLC* [2020] VSC 219 ('*Defteros*').

Critically, as intent is irrelevant, defamation can be established even where '[t]he communication … [is] unintentional, and the publisher … [is] unaware of the defamatory matter'.[99] While the defence of innocent dissemination may apply, such a defence was found not to be available with respect to material published by Google in their image and text search results after it was made aware that such material was produced by its search results.[100]

How a court would apply these principles to a question concerning a political deepfake is uncertain, especially in circumstances where a media platform was unaware the video was fake (and therefore defamatory). Such a question will be significantly affected by proposed (but not yet introduced reforms) to defamation law to limit the liability of media companies for defamation.[101] If such laws are passed, then individuals or political parties impacted by deepfakes created by overseas actors may lack any remedy under defamation law.

The second and third elements would be easy to establish in relation to political deepfakes. This is because an ordinary reasonable person would likely believe a deepfake video portrayed the individual depicted, even where slight imperfections were present. This accords with previous judicial reasoning concerning doctored images, which were of a significantly lower quality than is achievable in a deepfake.[102] Finally, given that the purpose of using a political deepfake is to lower the likelihood of an individual voting for a particular individual or party it is probable that in many cases a deepfake would contain a defamatory imputation. However, where a deepfake was *only* targeted at a political party it would fall outside the protection afforded by defamation law – which only protects the reputation of natural persons.

2 Passing Off

The classical elements of the tort of passing off under Australian law are drawn from the United Kingdom ('UK') case of *Reckitt & Colman Products Ltd v Borden Inc* ('*Reckitt & Colman*').[103] The broad test requires the establishment of the 'classical trinity', the elements of which are:

1. Reputation within Australia;
2. Misrepresentation; and
3. Damage.[104]

[99] *Lee v Wilson* (1934) 51 CLR 276, 288 (Dixon J).

[100] *Defteros* [2020] VSC 219, [134] (Richards J).

[101] Michael Douglas, 'Australia's Proposed Defamation Law Overhaul Will Expand Media Freedom – But at What Cost?', *The Conversation* (online, 1 December 2019) <https://theconversation.com/australias-proposed-defamation-law-overhaul-will-expand-media-freedom-but-at-what-cost-128064>. Reforms to limit liability of media companies and intermediary platforms are currently being considered by government: see Attorneys-General, 'Review of Model Defamation Provisions: Stage 2' (Discussion Paper, 2021) <https://www.justice.nsw.gov.au/justicepolicy/Documents/review-model-defamation-provisions/discussion-paper-stage-2.pdf>.

[102] See, eg, *Hanson-Young v Bauer Media Ltd [No 2]* [2013] NSWSC 2029.

[103] *Reckitt & Colman Products Ltd v Borden Inc* [1990] 1 WLR 491 ('*Reckitt*'). *Reckitt* was applied by the High Court in *ConAgra Inc v McCain Foods (Aust) Pty Ltd* (1992) 33 FCR 302 ('*ConAgra*').

[104] *Reckitt* [1990] 1 WLR 491, 499 (Lord Oliver); *ConAgra* (1992) 33 FCR 302, 355–6 (Gummow J).

In Australia, the *Reckitt & Colman* test has been regularly used to protect celebrities' images where individuals or businesses have implied that their goods or services have been approved or endorsed by the celebrity. For example, Ita Buttrose was successful in recovering damages where her image was used in a false endorsement.[105] Similarly, Paul Hogan was successful in recovering damages where an advertisement used an actor dressed in similar attire to his costume in Crocodile Dundee and used the now-famous line 'that's not a knife'.[106] This suggests that (similar to the analysis above in terms of defamation law) a deepfake could meet the requirements of this test, even if it contains slight glitches or imperfections. This is because courts are not assessing whether an individual is likely to believe that the celebrity portrayed really did say the words attributed to them, but instead whether an individual would form a connection in their mind such that they would believe 'the goods are … endorsed by the [celebrity]'.[107] In contrast, UK courts have historically been less willing to extend the doctrine of passing off beyond its traditional business roots,[108] although this has recently begun to shift.[109]

In the case of political deepfakes, the critical issues are whether a subject had a significant enough reputation in Australia, and whether a misrepresentation in the *commercial* sense protected by the tort had occurred. This case would differ from the traditional endorsement cases discussed above, as it is unlikely that a political deepfake would be used to advance a business interest. Instead, the deepfake would likely target a political interest: to affect public opinion regarding a politician, or the platform of a given politician or party. This analysis is analogous to the position adopted by Perot and Mostert who suggested that passing off may afford protections to individuals for certain categories of deepfakes in the UK.[110] The authors did not discuss the application of the test to political deepfakes. Where an opposing political party utilises a deepfake to further their political interests, this link may be easier to establish. In most cases involving political deepfakes, however, the current test for passing off is unlikely to serve as a suitable protection.

3 Limitations of Tort Law

As outlined in the above analysis, the efficacy of either defamation or passing off in combatting political deepfakes is limited. In addition to the gaps identified above, the primary limitation of tort law pertains to the remedies available to an aggrieved plaintiff. While courts are able to grant injunctions to prevent ongoing

[105] *Buttrose v The Senior's Choice (Australia) Pty Ltd* [2013] FCCA 2050 ('*Buttrose*').

[106] *Pacific Dunlop Ltd v Hogan* (1989) 23 FCR 553. Hogan has been an active celebrity in this space, also bringing an action against a company selling a 'Crocodile Dundee Koala Bear': *Hogan v Koala Dundee Pty Ltd* (1988) 83 ALR 187. See also 'Grill'd Settles Dispute with Paul Hogan', *SBS News* (online, 5 February 2018) <https://www.sbs.com.au/news/grill-d-settles-dispute-with-paul-hogan>.

[107] *Buttrose* [2013] FCCA 2050, [48] (Jones J).

[108] See, eg, *Elvis Presley Trade Marks* [1999] RPC 567, 598 (Brown LJ): 'there should be no … assumption that only a celebrity … may ever market … [their] own character'.

[109] See *Irvine v Talksport Ltd [Nos 1 and 2]* [2003] 2 All ER 881; *Fenty v Arcadia Group Brands Ltd* [2015] EWCA Civ 3.

[110] Perot and Mostert (n 90) 35–6.

damage, their use is limited.[111] This is especially the case for interlocutory applications where a court will only interfere in exceptional cases.[112] The reasons for this were summarised by the Federal Court in *Rush v Nationwide News Pty Ltd [No 9]*:[113]

> There are essentially three reasons why caution is warranted … [first that] free speech might be unnecessarily curtailed or restricted … [second that] it is not known whether publication of the matter will in fact invade the legal right of the applicant; and third, the fact that the defence of justification is ordinarily a matter for decision by a jury, not by a judge sitting alone …[114]

Additionally, defamation cases – the more useful remedy for individual politicians – are extremely costly and lengthy to run. Indeed, costs have been estimated to be as high as $80,000–$100,000 for cases involving only $10,000 in damages, leading to the introduction of legislation that would have removed the ability of parties to recover costs in low-value matters.[115] The cost-benefit analysis in the case of a deepfake affecting only 100–200 votes may be against bringing an action. Similarly, as a deepfake can be generated in a matter of days, a politician who embarked on a 'defend all cases' strategy may find themselves endlessly appearing in court. Fatigue, or mounting costs, would likely force the end to such action. In essence, the actions are limited by their personal nature, and the fact that parties may struggle to seek an injunction to prevent the ongoing harm.

Further, an award of damages would do little to restore trust in political and democratic institutions. Indeed, bringing an action can lead to increased media focus on the defamation case itself, allowing the allegedly defamatory claims to spread further. A more appropriate solution may be to empower impartial actors to secure the integrity of the voting process.

C Summary of Applicable Private Law

As outlined above, the remedies available in private law with respect to political deepfakes are insufficient. In particular, copyright law will only protect the relevant copyright holders – who are more likely to be media companies than the politicians impacted. Additionally, even where media companies were inclined to bring an action, the black box nature of deepfake technology would make identifying whose copyright had been infringed impossible in many cases. While defamation law would provide politicians with the strongest remedy, the time and costs needed to bring a defamation action limit its utility. Similar issues pervade the tort of passing off. Ultimately, rather than a private law action for damages,

[111] See, eg, *Australian Broadcasting Corporation v O'Neill* (2006) 227 CLR 57, 66 [16] (Gleeson CJ and Crennan J).

[112] Benedict Bartl and Dianne Nicol, 'The Grant of Interlocutory Injunctions in Defamation Cases in Australia following the Decision in *Australian Broadcasting Corporation v O'Neill*' (2006) 25(2) *University of Tasmania Law Review* 156.

[113] [2019] FCA 1383.

[114] Ibid [8] (Wigney J).

[115] New South Wales, *Parliamentary Debates*, Legislative Assembly, 18 September 2003, 3586–7 (David Barr). The laws were not passed in 2003; however, a Bill that will likely have a similar effect has now been passed in some states: see, eg, Defamation Amendment Bill 2020 (NSW).

those impacted by a deepfake likely want a 'public law' protection allowing them to take down harmful deepfakes.

IV EVALUATION OF PUBLIC LAW PROTECTIONS

Federal elections are governed by the *Commonwealth Electoral Act 1918* (Cth) ('*Electoral Act*'). While some Australian states have moved to prohibit specific uses of deepfake technology, notably in the context of intimate partner violence,[116] there are no specific laws or regulations concerning their use in federal, state or local elections.[117] Instead, the *Electoral Act* creates a number of general electoral offences that *may* apply to political deepfakes.[118] Where an offence has occurred, the *Electoral Act* creates a hybrid public-private enforcement regime, with both the AEC and candidates in an election able to seek an injunction to prevent conduct that would contravene the *Electoral Act*.[119] While there is some controversy concerning the availability of general administrative review rights against the AEC,[120] this question is not concerned with jurisdiction over electoral offences.[121] Therefore, while it remains unclear what remedies, if any, a private citizen has under the *Electoral Act*, this question is beyond the scope of this article although exploration of that topic may yield additional (and novel) remedies to the challenges posed by political deepfakes.

Relevantly, if requested by a candidate (during an election period), or the AEC, the Federal Court may grant an injunction where an offence has occurred or is likely to occur 'if in the opinion of the [Court] it is desirable to do so'.[122] Therefore, if the publication or distribution of a deepfake contravened a section of the *Electoral Act*, a court would be able to prohibit its publication through an injunction. This is exactly the remedy that the subject of a deepfake would be likely to seek. The below analysis highlights how two relevant offences would apply to political deepfakes.

A Misleading and Deceptive Conduct

Section 329 of the *Electoral Act* creates an offence for misleading and deceptive publication, which, on its face, would appear to apply to political deepfakes. The offence is however limited in its application. Section 329 relevantly states:

[116] *Crimes Act 1900* (NSW) ss 91N, 91Q. The use of deepfake technology would fall within the definition of 'altered image'.

[117] The latter two are beyond the scope of this article; however, state electoral regulations would provide guidance if they regulated deepfakes.

[118] *Electoral Act 1918* (Cth) pt XXI.

[119] Ibid s 383.

[120] Graeme Orr, 'Judicial Review of Electoral Affairs' (Conference Paper, AIAL National Administrative Law Forum, July 2011). See also Graeme Orr and George Williams, 'Electoral Challenges: Judicial Review of Parliamentary Elections in Australia' (2001) 23(1) *Sydney Law Review* 53.

[121] Orr (n 120).

[122] *Electoral Act 1918* (Cth) s 383(1).

> **329 Misleading or deceptive publications etc.**
> (1) A person shall not, *during the relevant period* in relation to an election under this Act, print, publish or distribute, or cause, permit or authorize to be printed, published or distributed, any matter or thing that is likely to mislead or deceive an elector *in relation to the casting of a vote*.[123]

While 'matter or thing' would likely include deepfake videos, and the term 'publish' includes distribution over the internet,[124] section 329 would be of limited use for two reasons. First, the section only applies during the relevant period – which is defined under the *Electoral Act* to be the period from the issue of writs to the conclusion of the election.[125] This means that the section would not apply to any communications or materials before the issuing of the writs. This limitation is not, however, critical. As noted above, the primary concern regarding deepfakes is their release close to an election where insufficient time remains to verify whether the contents of the video are true. As such, the limitation of section 329 to the time between the issue of writs and the end of the election would not be fatal to its use. More significant, however, is the limitation of the section to conduct 'in relation to the casting of a vote'. Courts have consistently held that this language limits section 329 to only apply to cases where the misleading or deceptive conduct relates to *how* an elector (having already decided who will be receiving their vote) would number the boxes on a ballot paper.[126] For example, the Full Federal Court in *Garbett v Liu*[127] stated:

> The provision is not concerned with a matter or thing which is misleading or deceptive and which might influence an elector in forming a judgment … It is concerned with the casting of the vote … The distinction is one between the formation of the political or voting judgment of the elector, and *its recording or expression*.[128]

Section 329 therefore does not guard against misleading or deceptive conduct in relation to electoral choices.[129] This can be contrasted with various state and territory electoral Acts which contain (or will soon contain)[130] prohibitions on false and misleading statements in advertising. For example, the South Australian *Electoral Act* creates an offence where:

> A person who authorises, causes or permits the publication of an electoral advertisement (an advertiser) is guilty of an offence if the advertisement contains a statement purporting to be a statement of fact that is inaccurate and misleading to a material extent.[131]

[123] Ibid s 329(1) (emphasis added).
[124] Ibid s 329(6).
[125] Ibid s 322.
[126] See, eg, *Evans v Crichton-Browne* (1981) 147 CLR 169.
[127] (2019) 273 FCR 1.
[128] Ibid 8 [31], 10 [36] (emphasis added).
[129] Historically, the section *did* engage with generally misleading and deceptive conduct – but the former provision was repealed: George Williams, 'Truth in Political Advertising Legislation in Australia' (Research Paper No 13, Parliamentary Library, Parliament of Australia, 24 March 1997).
[130] *Electoral Amendment Act 2020* (ACT) s 13, which will insert a new section 297A into the *Electoral Act 1992* (ACT).
[131] *Electoral Act 1985* (SA) s 113(2).

This provision, as of 2019, was the strongest 'truth in political advertising' law globally.[132] Notably, the University College London Report, in making this finding, outlined that amendments to the South Australian legislation in 1997 allowing the Electoral Commissioner to intervene to request an advertisement be immediately withdrawn meant that action could be taken before 'the election was over'.[133]

The utility of the South Australian provision has, however, been called into question.[134] For example, a former South Australian Electoral Commissioner outlined to a Federal parliamentary inquiry that:

> [H]e did not believe the South Australian legislation had had any appreciable effect on the nature of electoral advertising in the State. Instead, he considered that the legislation opened up opportunities for individual candidates to disrupt the electoral process by lodging nuisance complaints.[135]

Additionally, as the South Australian and Australian Capital Territory provisions apply only to *paid* advertising, they would not cover the use of deepfakes spread through social media by individuals not connected to a political campaign.

Nevertheless, absent such a provision, at a federal level, a deepfake falsely showing a candidate engaging in criminal activity, or outlining a false policy position which may mislead a voter as for whom they *wish* to vote would not be captured through the operation of section 329. In contrast, section 329 would prohibit the creation of a deepfake which, for example, falsely suggested which box a voter should number if they wished to vote for a particular party.[136]

B Publication of Matter regarding Candidates

The second provision that, on its face, appears to apply to political deepfakes is section 351, which relevantly states:

> **351 Publication of matter regarding candidates**
>
> (1) If, in any matter announced or published by any person, or caused by any person to be announced or published, on behalf of any association, league, organization or other body of persons, it is:
>
> (a) claimed or suggested that a candidate in an election is associated with, … that association, league, organization or other body of persons; or
>
> (b) expressly or impliedly advocated or suggested:
>
> (i) … that a voter should place in the square opposite the name of a candidate on a ballot paper a number not greater than the number of Senators to be elected; or

[132] Alan Renwick and Michela Palese, 'Doing Democracy Better: How Can Information and Discourse in Election and Referendum Campaigns in the UK Be Improved?' (Report, University College London, March 2019) 22.

[133] As was the case where courts had to make a determination: ibid 23.

[134] Ibid.

[135] Senate Standing Committee on Finance and Public Administration, Parliament of Australia, *Inquiry into Bills Concerning Political Honesty and Advertising* (Report, August 2002) 88 [5.60].

[136] This is analogous to the creation of a false how-to-vote card, which the AEC has stated would be captured by the section: Australian Electoral Commission, Submission No 1 to Joint Standing Committee on Electoral Matters, Parliament of Australia, *Inquiry into Allegations of Irregularities in the Recent South Australian State Election* (June 2010) 2–3.

> (ii) … that that candidate is the candidate for whom the first preference vote should be given;
>
> that person commits an offence.

A survey of results from two databases was not able to find any cases where the section has been used.[137] However, the section does appear to prohibit certain types of political deepfakes. This is because a deepfake of a candidate speaking may suggest to viewers that they hold the views outlined in the video. A key limitation of the provision is that the deepfake would have to be published *on behalf of* an organisation (or the associated terms used in the *Electoral Act*). The deepfake would then also have to suggest that the candidate is linked to the organisation, or suggest to voters how they should number their ballot paper (this part of section 351(1)(b) is similar to section 329). While it would be possible for a deepfake to fall within the section, it would be straightforward to design a deepfake to avoid such an outcome. Similarly, the section would not prohibit an individual, of their own volition, creating or disseminating political deepfakes.

C Summary of Applicable Public Law

As the above section has outlined, there are only limited public law protections available to political actors or the AEC as a means of pursuing those responsible for political deepfakes. While some limited types of deepfake will be captured, sophisticated actors will be able to avoid the subject matter areas that may run afoul of electoral regulation. The lack of remedy creates a gap in the law highlighting that the current legal framework is not fit for purpose at least insofar as it deals with the threat posed by political deepfakes.

V PROPOSED REFORM

Given the analysis above, reform is needed to combat political deepfakes. The following section discusses the constitutional limitations that would apply to federal laws developed to combat the threat of political deepfakes, before outlining a proposed model law.

A Commonwealth Powers

The Federal Government has a wide array of constitutional heads of power to draw on to regulate against the creation or distribution of *political* deepfakes. For example, the Commonwealth has the power to legislate with respect to elections,[138]

[137] With the usual caveats around use of available databases, the search terms "Electoral Act 1918 (Cth)" AND "351" AND "misleading" were used across two databases. No relevant cases were found.

[138] *Constitution* s 51(xxxvi).

copyright,[139] telecommunications,[140] corporations,[141] defence[142] and external affairs.[143] In combination these powers would likely allow[144] the Commonwealth to:

1. regulate the creation and content of political deepfakes by political parties or related entities within the context of federal elections using the elections power;
2. extend current copyright law to prohibit the creation of deepfakes;
3. ban the distribution of political deepfakes within and outside an electoral period through the use of a carriage service (including the internet);[145]
4. create offences relating to the creation or dissemination of deepfakes for the purpose of influencing elections due to the threat they pose to security;
5. impose duties on corporations acting in Australia to prevent the distribution of political deepfakes;[146] and
6. extend any offence provisions overseas.[147]

Given the wide array of options identified above, the key question to answer in determining what can be done to regulate against the threats identified in Part II is what limits, if any, the *Constitution* imposes with respect to these laws. Given the focus of this article on *political* deepfakes, the relevant limit is the operation of the implied freedom of political communication ('IFPC').

B Limits Imposed by the IFPC

The IFPC is a limitation on legislative and executive power derived from the text and structure of the *Constitution*.[148] The current test was applied by a majority

[139] Ibid s 51(xviii).

[140] Ibid s 51(v).

[141] Ibid s 51(xx).

[142] Ibid s 51(vi).

[143] Ibid s 51(xxix).

[144] The list is not intended to be an exhaustive statement regarding government power, merely to provide several examples identified by the author. No comment is made regarding the desirability of these regulations.

[145] This could likely be done through the telecommunications powers under which similar regulations barring the dissemination of child exploitation material have been passed: see *Criminal Code Act 1995* (Cth) sch div 474 sub-div D ('*Commonwealth Criminal Code*').

[146] This could be done using the corporations power contained in section 51(xx) of the *Constitution*, and would mirror current laws regarding child exploitation material: for discussion, see below n 171 and accompanying text.

[147] This could be done using the external affairs power in section 51(xxix) of the *Constitution*, analogous to current foreign interference laws: *Foreign Influence Transparency Scheme Act 2018* (Cth) s 7. Albeit the utility of such laws would be questionable, as foreign states can limit the utility of prosecution by not allowing their citizens to be extradited: see, eg, Amy Maguire, 'MH17 Charges: Who the Suspects Are, What They're Charged With, and What Happens Next', *The Conversation* (20 June 2019) <https://theconversation.com/mh17-charges-who-the-suspects-are-what-theyre-charged-with-and-what-happens-next-119155>. Notably both Russia and the People's Republic of China (nations which have been condemned internationally for their foreign interference efforts) have domestic laws that would prevent Australia from seeking extradition of their nationals: article 61 of the *Constitution of the Russian Federation*; «中华人民共和国引渡法» [Extradition Law of the People's Republic of China] (People's Republic of China) National People's Congress, Order No 42, 28 December 2000, art 8.

[148] *Australian Capital Television Pty Ltd v Commonwealth* (1992) 177 CLR 106. For discussion, see *Comcare v Banerji* (2019) 267 CLR 373, 395 [20] (Kiefel CJ, Bell, Keane and Nettle JJ) ('*Banerji*').

of the High Court in *McCloy v NSW*[149] and further clarified in *Brown v Tasmania*.[150] It requires a court to answer three questions:

> 1. Does the law effectively burden the implied freedom … ?
> 2. … is the purpose of the law legitimate … ?
> 3. … is the law reasonably appropriate and adapted to advance that legitimate objective … ?[151]

In assessing this third question a court must consider whether the law is suitable, necessary and adequate in its balance.[152] If question (1) is answered in the affirmative and either of questions (2) or (3) are answered in the negative the law will be invalid.[153]

In terms of regulating political deepfakes, the first question a court would need to assess is whether deepfakes are political speech. If not, then the IFPC would not apply. The key issue here is whether the IFPC protects *false* speech. While it does not appear that the High Court has made a direct finding on this issue, comments in obiter from both the High Court and the South Australian Supreme Court support the proposition that false speech is protected. This is especially the case where the speech is related to a core political matter. For example, in *Roberts v Bass*[154] Gaudron, McHugh and Gummow JJ stated that defamation (which inherently is concerned with untrue statements) is limited by the IFPC.[155] A similar finding was made by the Court in *Lange v Australian Broadcasting Corporation*.[156] Even in cases concerned with false statements, courts have stepped through the entirety of the *McCloy* test to assess whether a law is adequate in its balance.[157] For example, in *Cameron v Becker*, in holding that section 113 of the South Australian *Electoral Act* did not breach the IFPC, Olsson J (with whom Bollen J agreed) appeared to hint that false speech would not attract the protection of the IFPC.[158] However, Olsson J went on to assess whether the law was 'reasonably appropriate and adapted'.[159]

This approach prevents courts from unnecessarily assessing whether speech is or is not true. Especially within the context of elections, Australian courts have taken care when applying the IFPC. For example, Kirby J in *Roberts v Bass* stated:

> Because this is the real world in which elections are fought in Australia, any applicable legal rule … must be fashioned … to reflect such electoral realities. Otherwise, before or after the conduct of elections, attempts will be made to bring to courts of law, under the guise of legal claims, the very disputes that it was the

[149] *McCloy v NSW* (2015) 257 CLR 178 ('*McCloy*').

[150] *Brown v Tasmania* (2017) 261 CLR 328.

[151] *Clubb v Edwards* (2019) CLR 171, 186 [5] (Kiefel CJ, Bell and Keane JJ). See also *Banerji* (2019) 267 CLR 373, 398–400 [29]–[32] (Kiefel CJ, Bell, Keane and Nettle JJ).

[152] *Brown v Tasmania* (2017) 261 CLR 328, 368 (Kiefel CJ, Bell and Keane JJ), 376 (Gageler J), 416–17 (Nettle J), 476–7 (Gordon J); *Banerji* (2019) CLR 373, 400 [32] (Kiefel CJ, Bell, Keane and Nettle JJ).

[153] *McCloy* (2015) 257 CLR 178, 193–5 [2]–[3] (French CJ, Kiefel, Bell and Keane JJ).

[154] (2002) 212 CLR 1.

[155] Ibid 40–1 [102].

[156] (1997) 189 CLR 520.

[157] In the context of the earlier tests predating the *McCloy* test, see, eg, *Cameron v Becker* (1995) 64 SASR 238, 248 (Olsson J, Bollen J agreeing at 239).

[158] Ibid 247.

[159] Ibid 248.

> purpose of the representative democracy, established by the *Constitution*, to commit to the decision of the electors.[160]

This approach balances the need for laws to comply with the IFPC with the risk that courts could become an electoral and political battleground, subverting the will of the people. The United States, in contrast, has a very litigious electoral system with state and federal courts often called on to settle political controversies around voting rights, access to voting and the legitimacy of electoral results. This approach culminated in the *Bush v Gore* decision where the Supreme Court split on party lines to elect George W Bush as President.[161]

While deepfakes are a form of *false* speech in that they portray individuals making false statements, there are also many legitimate uses of deepfakes as discussed above. Deepfakes can be used as a form of parody or satire, or to educate the general population about the threat of fake news. Additionally, deepfakes can be used by politicians to make videos of *themselves* speaking in different languages in efforts to appeal to a greater share of the voting base. Laws that purport to prohibit the creation or dissemination of political deepfakes would impact on these *legitimate* uses of the technology. As members of the High Court have recently made clear, laws which impact on future communications may have a significant chilling effect.[162]

As such, and especially given the statements in *Cameron v Becker*, it seems likely that a court would find that laws that limit the publication and dissemination of deepfakes are a burden on the implied freedom. Therefore, any laws prohibiting the creation or dissemination of a political deepfake would need to be compatible with the system of representative government, and be reasonably and appropriately adapted to that legitimate purpose. The purpose underlying the laws has been addressed earlier in this article; however, in sum, the laws would aim to safeguard Australian elections and ensure that voter preferences were not subverted by deepfakes. This purpose aims to strengthen legitimate political communication and protect elections from both foreign interference and domestic threats and would likely be compatible with Australia's system of representative government. The key issue in designing such laws is therefore in ensuring that the laws are suitable, necessary and adequate in their balance. This analysis will depend on the specific measures adopted and will accordingly be discussed further below alongside the proposed legislative scheme.

C Possible Reforms

As outlined above, there are several potential avenues for reform. In determining which approach should be taken, three questions need to be answered:

[160] *Roberts v Bass* (2002) 212 CLR 1, 63 [172].

[161] See, eg, *Bush v Gore*, 531 US 98 (2000).

[162] See, eg, *LibertyWorks Inc v Commonwealth* [2021] HCA 18, [95] (Gageler J) (noting that his Honour was in dissent on this issue with the plurality finding that there was no such acceptance of strict scrutiny for prior restraint: at [50] (Kiefel CJ, Keane and Gleeson JJ)). Regardless, however the burden is analysed it is clear that laws which impact on an individual's ability to communicate about politicians using deepfakes will likely fall afoul of the first element of the *McCloy* test.

1. On whom should obligations be imposed?
2. What type of remedy is appropriate?
3. Who should be able to seek the remedy?

In answering these questions, it is important to outline the purpose of these proposed reforms: to safeguard *elections* by preventing voters from being swayed by misinformation and disinformation. While ordinarily a certain amount of misinformation is anticipated in the context of a contested election campaign, the need to combat deepfakes has been clearly articulated. The situation can be distinguished from false claims generally as there is no practical way for the subject of a deepfake to correct the record. Either people will believe the video is real, or they will not. This is different, for example, from false advertising regarding death taxes[163] or Medicare funding,[164] as these policy-based arguments can, at least in theory, be debated and corrected on the public record.[165]

In contrast, the difficulty in disproving a deepfake and the increasing ease of deepfake creation[166] justifies intervention. Care, however, must be taken to ensure that any new regulations are not used by political parties to decide electoral contests through litigation.[167] Such an outcome could erode the trust of electors in elections, and undermine the separation of powers in Australia by giving courts the ability to decide electoral contests.[168] It would also mean the law would be more likely to breach the IFPC as not being reasonably and appropriately adapted to the purpose it is seeking to achieve. Finally, the outcome could also increase the political profile of federal courts, and increase the influence of a judge's political persuasion in appointment decisions. The dangers of this potential outcome are on full display in the US, where the confirmation of Justice Amy Coney Barrett took place eight days before the election, with President Trump admitting he hoped the appointment would establish a sympathetic bench to rule upon electoral issues such as mail voter fraud.[169]

[163] Katharine Murphy, Christopher Knaus and Nick Evershed, '"It Felt Like a Big Tide": How the Death Tax Lie Infected Australia's Election Campaign', *The Guardian* (online, 8 June 2019) <https://www.theguardian.com/australia-news/2019/jun/08/it-felt-like-a-big-tide-how-the-death-tax-lie-infected-australias-election-campaign>.

[164] See, eg, Nicholas Reece, 'Why Scare Campaigns Like "Mediscare" Work: Even if Voters Hate Them', *The Conversation* (online, 14 July 2016) <https://theconversation.com/why-scare-campaigns-like-mediscare-work-even-if-voters-hate-them-62279>.

[165] For example, while Labor acknowledged the impact of the 'death tax' ads on its campaign, its report into the 2019 election admits that much of the blame lay with an unwieldy policy platform and an inability to respond to the claims in a way that voters could understand: see, eg, Emerson and Weatherill (n 38) 19, 74.

[166] For discussion, see Part II.

[167] *Roberts v Bass* (2002) 212 CLR 1, 63 [172] (Kirby J).

[168] While currently the Court of Disputed Returns can void an election, the grounds on which they can do so are extremely limited: *Electoral Act 1918* (Cth) pt XXII. Such powers have never been used.

[169] Jordyn Phelps, 'Trump Argues His Nominee Needed on Supreme Court in Time to Vote on Election Legal Challenges', *ABC News* (online, 24 September 2020) <https://abcnews.go.com/Politics/trump-argues-nominee-needed-supreme-court-time-vote/story?id=73192756>. See also ABC News, 'Donald Trump's Nominee Amy Coney Barrett Confirmed to Supreme Court of the United States', *ABC News* (online, 27 October 2020) <https://www.abc.net.au/news/2020-10-27/amy-coney-barrett-confirmation-senate-supreme-court-donald-trump/12815614>.

1 Where Should Obligations Fall?

This question is likely the most contentious of the three, especially given recent attempts by the Commonwealth to impose obligations on social media companies to pay for news have led to threats by social media and internet companies to withdraw or limit their Australian operations.[170] What is clear is that individuals or political parties who share deepfakes with the intention of impacting elections should be captured by the regulations. The laws should also account for scenarios where the precise author of the deepfake remains unknown (at least when action is commenced). As highlighted above, it is possible for anonymous actors overseas to be responsible for the creation and dissemination of deepfakes. What is less clear is how to regulate news media companies and social media platforms who may unknowingly assist in the distribution of a deepfake. In the context of terror attacks or child exploitation material, obligations have been imposed (both in Australia and internationally) on media companies to prohibit the sharing or uploading of content.[171] This has led to platforms creating automated tools that flag and then delete any such content.[172] While some commentators have suggested imposing obligations on social media platforms prohibiting the spread of fake news, even in the context of deepfakes, such an approach may be unwieldy and overbroad.[173] This is because defining misinformation and disinformation is much harder than defining child exploitation material, or abhorrent violent material, and as such additional content may be captured by automated detection tools (and unnecessarily censored).[174]

Obligations imposing significant penalties on service providers or content hosts where their platform is used to access that material may therefore lead to unnecessary restrictions on free speech, with providers removing more content than necessary. For example, videos that were clearly identified as deepfakes and were uploaded for educational purposes may be removed by risk averse companies using automated tools to detect and remove *all* deepfakes. This in turn would hurt the democratic process by unnecessarily restricting political communication.

To balance the need for media freedom (and avoid unnecessarily burdening social media platforms), a two-pronged approach could be used. At first instance, action could be taken against the original creator or disseminator of the deepfake. Then, if that action is successful, obligations could be imposed on social media platforms

[170] Matthew Doran and Jordan Hayne, 'Facebook Threatens to Ban Australians from Sharing News after Google Launches Attack on Government Plans', *ABC News* (online, 1 September 2020) <https://www.abc.net.au/news/2020-09-01/facebook-threatens-to-ban-australians-from-sharing-news-content/12616216>.

[171] Further measures were introduced following the live-streaming of the Christchurch terror attack to insert (among other provisions) sections 474.33 and 474.34 into the *Commonwealth Criminal Code*: Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019 (Cth). Sections 474.33 and 474.34 drew on the approach in section 474.25 which imposes obligations where an internet service provider or content host is aware that the service can be used to access child abuse material to impose similar obligations with respect to abhorrent material.

[172] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7(1) *Big Data & Society* 1.

[173] Ibid 10–12.

[174] Ibid.

and news companies to prevent them from *knowingly* allowing this content to be shared. This would have the effect of reducing the burden on social media companies – who would not need to decide whether material did or did not need to be removed at the first instance. They would instead be able to rely on a court determination and *then* use automated detection tools to remove any re-uploaded deepfake videos. Such an approach mirrors that taken in relation to extremist content following the Christchurch terror attacks,[175] and circumvents much of the ongoing debate around the extent of safe harbour provisions[176] as the regulation will be limited to a defined set of videos of which technology companies are aware.

By limiting the restrictions imposed in this manner, the government could leave decisions about less harmful cases (including when content should be downgraded in searches or flagged as misleading or false) to social media companies themselves, who can manage these issues under internal policies.[177] Under this approach, the government's efforts will be tailored to focus on the greater threat posed by deepfakes, ensuring that the law is not overbroad and more likely to be held to be reasonably and appropriately adapted.

2 What Type of Remedy Is Needed and When Should It Be Available?

It is clear from the preceding analysis that damages are not a sufficient remedy to combat political deepfakes. Instead, what is needed is an ongoing injunction restraining the publication or republication of the relevant political deepfake. Given that deepfakes can be easily re-uploaded, a further remedy should be available: the ability to request or compel a public correction of the record by the party responsible for publishing the deepfake. This approach mirrors that contained in the South Australian and Australian Capital Territory electoral Acts regarding false political advertising.[178] Where a public retraction is required, legislation should, as a matter of course, require the retraction be in the same form and shared as widely as the original post or video. While this power is likely already captured in the wide array of orders a court may grant under the *Electoral Act*,[179] an express statement would clearly indicate its availability and help tailor the conditions attached to the order. It is worth noting that existing provisions in the *Electoral Act* require courts

[175] See, eg, *Commonwealth Criminal Code* sub-div 474(H).

[176] See, eg, Peter Leonard, 'Building Safe Harbours in Choppy Waters: Towards a Sensible Approach to Liability of Internet Intermediaries in Australia' (2010) 29(3) *Communications Law Bulletin* 10; Danny Friedmann, 'Sinking the Safe Harbour with the Legal Certainty of Strict Liability in Sight' (2014) 9(2) *Journal of Intellectual Property Law & Practice* 148. This approach accords with the safe harbour scheme contained in clause 91 of schedule 5 of the *Broadcasting Services Act 1991* (Cth) which requires *knowledge* to impose liability on an internet service provider. In the author's view, safe harbour protections should generally not be afforded to internet service providers in relation to electoral offences where they are aware that the content infringes electoral law.

[177] For discussion on the measures already taken by social media companies, see, eg, Emma Llansó et al, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' (Working Paper, Transatlantic Working Group on Content Moderation Online and Freedom of Expression, 26 February 2020).

[178] *Electoral Act 1985* (SA) s 113; *Electoral Act 1992* (ACT) s 297A.

[179] *Electoral Act 1918* (Cth) s 360, noting that the section is framed as an inclusive list of powers.

to make decisions as quickly as possible given the circumstances of the case.[180] This further strengthens the appropriateness of the proposed remedy.

3 Who Should Be Able to Seek the Remedy?

This third question is likely the easiest of the three to answer. In line with current practice, the hybrid public-private model created by the *Electoral Act* should be applied. This would allow candidates affected (if the video occurs during an election campaign) and the AEC to bring an action. Limiting the action to candidates only during an election period further tailors the law, as it prevents overuse of the courts for political point-scoring. Allowing the Electoral Commissioner to issue notices will enable action to be taken rapidly rather than requiring court action in every case. It will also enable the Commissioner to issue take-down notices in situations where the creator or disseminator remains anonymous and a civil action against that person may not be possible. While this alone will not resolve the issue of attribution of actions taken online, especially where actions are taken by state-sponsored actors, it will go some way to providing the Commissioner with powers to remove deepfake content. Of course, in an Australian context, current electoral laws already require the identification of the individual(s) authorising electoral communications.[181]

D Proposed Amendments

To give effect to the above, two proposed amendments to the federal *Electoral Act* are set out below:

Section 329A Publish or distribute altered images etc.

(1) This section applies to altered images published by any means.

(2) A person who authorises, causes or permits the publication of any matter or thing is guilty of an offence if the matter or thing contains a statement regarding electoral matters that is inaccurate or misleading to a material extent.

(3) In prosecuting a person for an offence under this section, it is a defence if:

(a) the person proves that they did not know and could not reasonably be expected to have known, that the matter or thing was:

(i) likely to mislead or deceive an elector to a material extent; or

(ii) an altered image; or

(b) the person proves that:

(i) the material or thing was published for the purpose of education, comedy, or satire; and

(ii) the material or thing was identified as an altered image.

[180] Ibid s 363A.

[181] Ibid pt XXA. One possible alternate to the proposal in Part V(D) would be to require an authorised individual to be nominated for every political deepfake published in Australia and to take down any deepfakes that are not authorised; however, such a measure would have far greater impact on political communication and accordingly in the author’s view this approach is likely more proportionate to the threat it seeks to prevent.

(4) If the Electoral Commissioner is satisfied, on the balance of probabilities, that a matter or thing has been published in relation to electoral matters that is inaccurate or misleading to a material extent, the Electoral Commissioner may request a person who has authorised, caused or published the matter or thing, to do one or more of the following:
   (a) withdraw the matter or thing from further publication;
   (b) publish a retraction in specified terms and in a specified manner and form.

(5) In deciding the terms, manner and form of a retraction requested under section 329A(4), the Electoral Commissioner must consider:
   (a) the terms, manner and form of the matter or thing published; and
   (b) the number of times the matter or thing had been viewed.

(6) If the Court is satisfied, on the balance of probabilities, that a matter or thing has been published in relation to electoral matters that is inaccurate or misleading to a material extent, the Court may order any person who has authorised, caused or published the matter or thing, to do one or more of the following:
   (a) withdraw the matter or thing from further publication;
   (b) publish a retraction in specified terms and a specified manner and form.

(7) Where a person consents to comply to a request under subsection (4) or an order is made under subsection (5) the Electoral Commissioner must publish a notice, in the manner prescribed by the regulations, notifying internet service providers and internet content hosts that such an order has been made or a request consented to.

Note: A person can consent to comply with a request on a without-admissions basis.

(8) The Electoral Commissioner may make regulations for the purpose of establishing a notification scheme where members of the public or candidates may draw the Electoral Commissioner's attention to a purported offence under this section.

Section 329B Obligations of internet service providers and internet content hosts relating to altered images

(1) A person commits an offence if the person:
   (a) is an internet service provider or an internet content host; and
   (b) is aware that the service provided by the person can be used to access material that has been subject to an order or request under section 329A; and
   (c) does not refer details of the material to the Electoral Commissioner within a reasonable time after becoming aware of the existence of the material; or
   (d) if requested by the Electoral Commissioner, does not take reasonable steps to take down or remove access to that material.

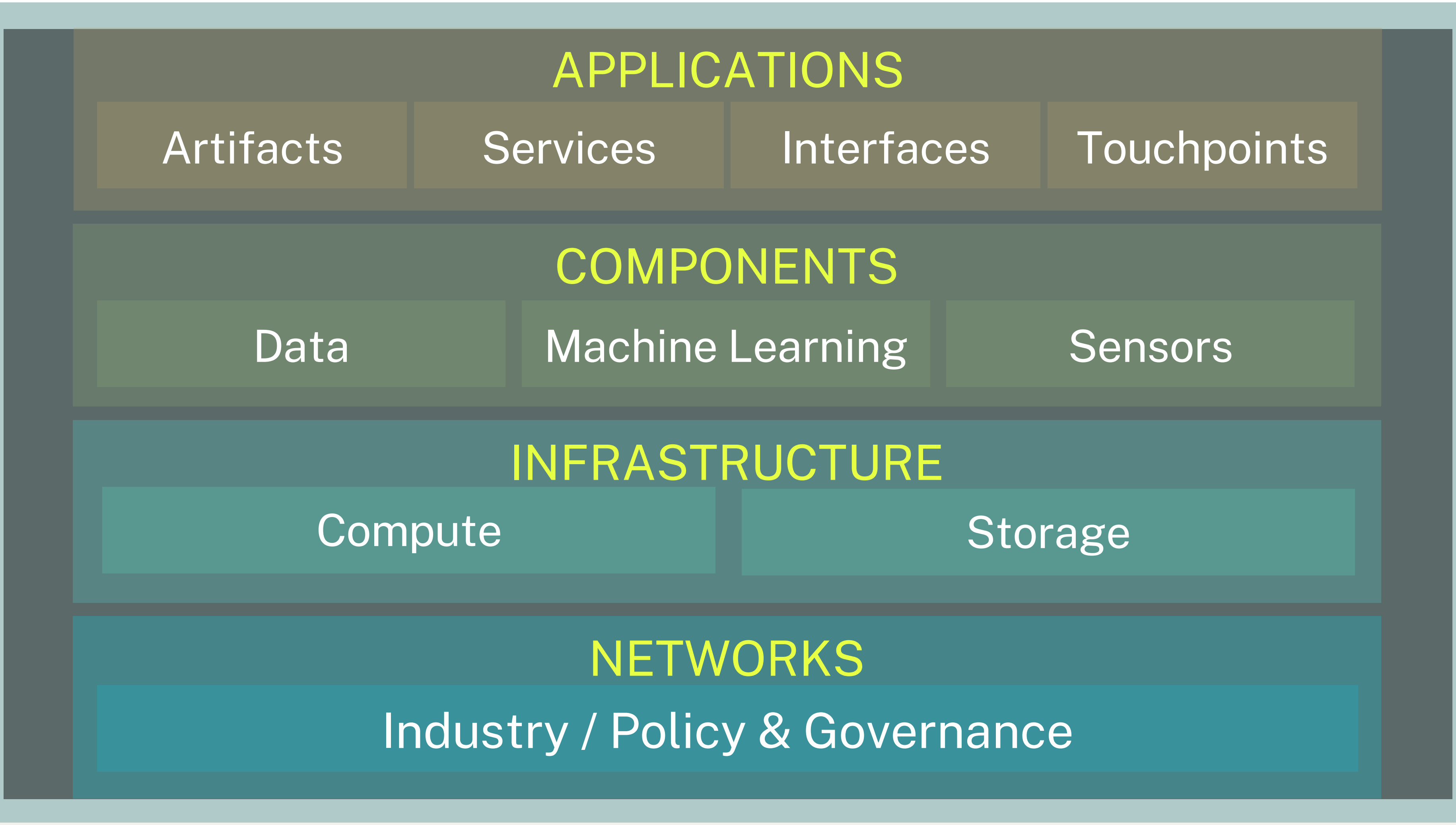
(2) A person is presumed to be aware that a service they provide can be used to access material subject to an order or request under section 329A, where a

notice under section 329A(7) has been published and a reasonable period of time has elapsed.

(3) The Electoral Commissioner may provide guidance to organisations relating to their obligations under subsection (1). Any such guidance must be published in the manner prescribed in the regulations.

VI CONCLUSION

Regulations that limit free speech must be suitable, necessary, and adequate in their balance. This article has considered the current protections available to politicians, political parties and the AEC to combat the growing threat posed by deepfake technology to elections, and by extension to democracy. It concludes that there are current gaps in the law, with copyright, tort and electoral law only offering very limited protections that could be readily avoided. These protections remain unclear, ill-defined and are inadequate to prevent the use of deepfakes to directly sway voter preferences, or to undercut truth in political discourse. In response, it proposes two targeted amendments to the *Electoral Act*. The amendments are, critically, both tailored proportionate to the threat posed by deepfakes. This article concludes that these measures are distinguishable from (appropriately rejected) calls for general regulations concerning misinformation or disinformation. While it would likely be possible to craft such laws, they would overburden free speech in Australia and lead to a significant chilling effect for media organisations, internet content platforms and everyday citizens, and reduce the strength of our democratic institutions.



APPLICATIONS

Artifacts | Services | Interfaces | Touchpoints

COMPONENTS

Data | Machine Learning | Sensors

INFRASTRUCTURE

Compute | Storage

NETWORKS

Industry / Policy & Governance

The AI tech stack