JUTLP
Journal of University Teaching & Learning Practice

Deepfakes and Higher Education: A Research Agenda and Scoping Review of Synthetic Media

Dr Jasper Roe[a], Associate Professor Mike Perkins[b], and Mr Leon Furze[c]
[a] James Cook University Singapore, Singapore; [b] British University Vietnam, Vietnam [c] Deakin University, Australia

Abstract

The availability of software which can produce convincing, yet synthetic media poses both threats and benefits to tertiary education globally. While other forms of synthetic media exist, this study focuses on deepfakes: advanced Generative AI (GenAI) imitations of real people's likeness or voice. This conceptual paper assesses the current literature on deepfakes across multiple disciplines by conducting an initial review of 182 peer-reviewed publications. The review reveals three major trends: detection methods, malicious applications, and potential benefits, although no specific studies on deepfakes in the tertiary educational context were found. Following a discussion of these trends, we apply the findings to postulate the major risks and potential mitigation strategies of deepfake technologies in higher education, as well as potential beneficial uses to aid the teaching and learning of both deepfakes and synthetic media. This culminates in the proposal of a research agenda to build a comprehensive, cross-cultural approach to investigate deepfakes in higher education.

Editors

Section: Educational Technology
Editor-in-Chief: Dr Joseph Crawford
Senior Editor: A/Prof Rachel Fitzgerald

Publication

Received: 24 July 2024
Revised: 2 September 2024
Accepted: 5 September 2024
Online First: 6 September 2024

Citation

Roe, J., Perkins, M., Furze, L. (2024) Deepfakes and Higher Education: A Research Agenda and Scoping Review of Synthetic Media. *Journal of University Teaching and Learning Practice,* Advanced Online Publication. https://doi.org/10.53761/2y2np178.

Introduction

The pace of the development of Artificial Intelligence (AI) technologies has led to significant concern in many areas of society, including educational contexts. As a result, research agendas on Generative AI (GenAI) in tertiary education have been established (Lodge et al., 2023); however, to date, no review or research agenda has specifically focused on deepfakes in tertiary education. Deepfakes are GenAI outputs which comprise realistic audio, visual, or media outputs that depict false or inaccurate information (Akhtar, 2023). The major consequence of deepfakes is that they can portray an individual doing something or saying something that they have never done, marking an unprecedented shift in the ability to distort reality (Appel & Prietzel, 2022). As tertiary education institutions are centres of learning, the potential implications of such false information are highly important for students, teachers, and university leadership, thus warranting stakeholder attention.

Our study contributes to the literature on educational technology in tertiary education by addressing three distinct aims. First, to conduct a rapid review to assess the literature on deepfakes and synthetic media across multiple fields and to relate the impact of these findings to the context of tertiary education. Second, to identify potential strategies and best practices that can serve as a starting point for crafting coherent institutional approaches to dealing with deepfakes. The third aim is to propose a research agenda establishing priority areas for addressing the potential challenges and benefits of deepfake technology in higher education. To achieve these objectives, our study addresses the following research questions:

1. What is the current state of research on deepfakes and synthetic media across various disciplines, and how do these findings relate to tertiary education?
2. What are the potential risks and benefits of deepfake technology in higher education contexts?
3. What strategies can higher education institutions implement to address the challenges posed by deepfakes?
4. What are the priority areas for future research on deepfakes in higher education?

By addressing these research questions, our study makes a pioneering contribution to the field of educational technology in tertiary education. This article lays the foundation for the study of deepfakes and synthetic media in tertiary education by offering the first review and research agenda specifically focused on this context. Our approach synthesises findings from a number of diverse fields, providing a holistic understanding of deepfakes that bridges the gap between technology and educational research, which we use to develop evidence-based strategies and best practices for higher education institutions to address deepfake-related challenges, filling a critical need in the sector. By proposing a detailed research agenda, we identify key areas which we believe require further investigation to equip stakeholders with the knowledge needed to anticipate and prepare for future technological developments. Through these contributions, we aim to stimulate dialogue, inform policy, and inspire further research on the intersection of deepfakes and education.

The remainder of this paper is structured to address our research questions and objectives. We begin with a literature review providing an overview of deepfake technology, its history, and current applications, followed by a synthesis of existing research across various fields. The methodology section details our approach to conducting an initial review broadly in line with a scoping review procedure, including search strategies, inclusion criteria, and analysis methods. We then present the findings of our review, highlighting key themes and their relevance to tertiary education before analysing the implications of deepfakes for higher education, including potential risks and opportunities. Following this, we propose evidence-based strategies and best practices

for higher education institutions to address deepfake-related challenges. The paper concludes with a proposed research agenda outlining priority areas for future investigation and a summary of key findings and their significance for educational technology in tertiary education.

Reviewing The Landscape of Deepfakes and Deepfake Research

The term deepfake was coined in 2017 by a user on the social media site Reddit, to refer to a combination of the terms 'deep learning' and 'fakes' (Kietzmann et al., 2020), in contrast to 'shallow fakes' or 'cheap fakes' which revolve around simpler, less convincing methods of media manipulation, such as slowing videos or using photo editing software. While we use the English terminology 'deepfake' in this article, it is important to note that deepfakes are a global phenomenon, do not only occur in Western contexts, and have their own terms in other languages (e.g. "huanlian" in Chinese (de Seta, 2021)). Cases of deepfakes spreading on social media have been documented in Kenya (Mayoyo, 2023), India (Nema, 2021), Russia (Samoilenko & Suvorova, 2023), and China (de Seta, 2021).

Although deepfakes have been recognised since 2017, the underlying technology required to successfully produce these manipulated media outputs has only recently been sufficiently developed to become widely accessible (Busacca & Monaca, 2023). Public awareness of the issue has also grown through several recent high-profile scandals, including pornographic deepfakes of celebrities such as Taylor Swift (Saner, 2024). These cases have led to Reddit banning the original 'deepfake' subreddit in which the term was coined (Appel & Prietzel, 2022), and in the United Kingdom creation of nonconsensual, explicit deepfake images has been criminalised, attracting fines and even jailtime for creators (Milligan, 2024).

We differentiate between the terms "deepfake" and "synthetic media" to acknowledge the potential creative, economic, and educative benefits of AI-generated media (Feher, 2024). While we primarily focus on deepfakes in this paper, it is also important to note that purely synthetic media – i.e., not created with a visual or audio likeness of a real person – is important in higher education and forms part of our research agenda. Several platforms exist which can create purely synthetic media, including synthetic speech (ElevenLabs, 2024), music (e.g., Suno, 2024; Udio, 2024), and video avatars created using diffusion-based image models (HeyGen, 2024; Synthesia, 2024).

Deepfakes are produced through deep learning procedures which typically involve Generative Adversarial Networks (GANs) or more recently, advanced diffusion models (Appel & Prietzel, 2022). Historically, deep learning to produce visual materials has required large sets of training data. This is one reason why celebrities have been common targets, as huge libraries of images and video material on these individuals already exist (Kietzmann et al., 2020). More recently, end-to-end systems which generate videos of talking heads using only a single still image and an audio clip have been developed (Vougioukas et al., 2019; Xu et al., 2024), suggesting the abilities of synthetic media and deepfakes will continue to improve, and that deepfakes of anyone with a single image may be achievable. Similar to large language models (LLMs) and GenAI image generators, deepfake technologies are increasingly accessible to the general population. There are multiple open source applications for deepfake creation, including DeepFaceLab, FaceApp, Wombo, Zao, and FaceApp (de Rancourt-Raymond & Smaili, 2022). One of the most popular applications, FakeApp, makes the creation of deepfakes possible for users with basic technological capabilities quickly and easily (Ali et al., 2021). Therefore, deepfakes are expected to become more common soon.

There are no agreed upon common categories for deepfakes, but there are areas of commonality across definitions and frameworks. Farid (2022) outlines three categories of visual deepfake: 'face-swap' such as swapping the face of an actor into another movie; lip synchronisation (lip-sync), in which the modification of a video aims to match the speaker's lips with audio recording;

and ‘puppet-master,’ animating a person’s image through another performer (the ‘puppet’). Akhtar (2023) suggested four alternate types of visual deepfake manipulation: swapping of identities, face reenactment, manipulation of attributes, and synthesis of the face in its entirety. Kietzmann et al. (2020)’s taxonomy of deepfakes extends beyond visual media, encompassing face and body swapping, audio deepfakes including voice-swapping, video, and combined video/audio deepfakes. Regardless of these minor differences in classification, deepfakes of all varieties produce similar output: believable, realistic media which is predicated on some aspect of ‘fakery’ or inaccurate representation, thus distinguishing it from synthetic media which can use a non-existent or artificial identity.

Methods

While our review aims to understand the scope of deepfake and synthetic media research from the perspective of tertiary education, there have been many broader reviews of the deepfake literature in other disciplines. Bode et al. (2021) points out that there has been an ‘explosion’ of multidisciplinary research on deepfakes, often focusing on erosion of truth and trust, the creation of fake pornography, and political manipulation. A systematic review of deepfakes identified that much of this research is within the fields of computer science, politics, and law (Godulla et al., 2021), while a bibliometric analysis of 331 research articles surrounding deepfakes on Web of Science and Scopus databases found similar patterns (Gil et al., 2023).

Given that we aimed to ascertain initially whether there was research on deepfakes in tertiary education, we adopted a scoping review procedure (Mak & Thomas, 2022). We used the academic databases Google Scholar, Scopus, and Web of Science for the operators ‘deepfakes’, ‘education’, ‘higher education’ and ‘synthetic media’. After screening the abstracts for those relevant to the topic of deepfakes, we obtained 182 peer-reviewed journal publications, conference proceedings, and chapters in edited books regarding deepfakes and synthetic media. We opted to include all studies relevant to the topic of deepfake technology, as long as they met the guidelines of being peer-reviewed, scholarly literature. Although we excluded grey literature from this review, we also noted a distinct lack of grey literature in Google Scholar (which is the only database utilised which contains such literature), suggesting that deepfakes as a topic has yet to reach mainstream public attention. We then read the collected papers in detail and found no works that could be described as specifically focusing on the tertiary education context. This indicated there was not sufficient literature for a detailed scoping review on deepfakes in higher education, although there is limited existing literature regarding deepfake education among younger cohorts (Ali et al., 2021; Blankenship, 2021; Murillo-Ligorred et al., 2023). As a result, we decided to take a more integrative approach to synthesising the existing knowledge on deepfakes, before attempting to relate the findings to a tertiary education context to begin the foundations of scholarly work in this area. To do so, we undertook an inductive thematic analysis to identify the current themes of extant scholarship on deepfakes and synthetic media.

Results

We identified three broad themes in the literature: deepfake detection, malicious applications of deepfake technologies, and positive applications of deepfake technologies. This literature is synthesised in the results section, and then analysed and related specifically to a tertiary educational context.

Table 1

Categories of Deepfake Literature

Theme	Description
Deepfake Detection	Most studies relating to this theme fell into the disciplines off computer science and Artificial Intelligence (AI). This theme covers studies which describe novel techniques for detecting deepfake media, or evading detection technologies through adversarial means.
Malicious Applications of Deepfakes	This theme relates to works which sought to describe the potential malicious applications of deepfake technology across multiple fields and areas of society, with most studies focusing on the implications in politics, public voting, and disinformation, or the generation of nonconsensual pornography.
Positive Applications of Deepfake Technologies	Although representing a small minority of the papers analysed, we found several articles which did describe potential positive applications of deepfake technology, including in guest experience at tourism attractions, or in benefiting accessibility for those with vocal impairments.

Detection of Deepfakes

Most deepfake literature focuses on detection and evasion, mirroring scholarship on GenAI text detection (Anderson et al., 2023; Perkins et al., 2023; Weber-Wulff et al., 2023). While early, less sophisticated deepfakes had detectable inconsistencies like blinking frequency (Agarwal et al., 2019), recent iterations are harder for humans to detect (Chadha et al., 2021). For example, one bilingual study (n = 529) found listeners correctly identified audio deepfakes only 73% of the time, regardless of language (Mai et al., 2023), while other studies have found that people may overestimate their detection abilities (Köbis et al., 2021).

Other methods of forensic analysis which centre on human-led detection, such as checking for inconsistency between frames, have proven unsuccessful (Guarnera et al., 2020), not only because deepfakes are increasingly sophisticated, but also because of the volume of deepfake content being created and the time needed to apply these methods (Silva et al., 2022). Given difficulties in relying on human-led detection of deepfakes, it is unsurprising that computational or deep learning methods of detection have become a popular topic of enquiry. There is a 'cat and mouse game' of detection and creation (de Seta, 2021), noted in other forms of high-technology detection processes in an AI-driven higher education landscape (Roe & Perkins, 2022). Many studies continue to posit novel detection techniques to improve accuracy of deepfake detection (Giudice et al., 2021; He et al., 2021; Hernandez-Ortega et al., 2020; Huang et al., 2020; Jung et al., 2020; Porcile et al., 2024; Qi et al., 2020; Silva et al., 2022; Tolosana et al., 2020). However, some argue that because of this tension between identification and evasion, deepfakes are not reducible to a problem solvable by detection (Jacobsen & Simpson, 2023).

Malicious Applications of Deepfake Technologies

Following deepfake detection, the negative potential impacts of deepfake technologies seem to account for much of the extant literature outside of the domain of computer science. These include the dangers of deepfake misinformation on political processes and elections (Diakopoulos &

Johnson, 2021; Dobber et al., 2021; Hameleers et al., 2022; Vaccari & Chadwick, 2020), pornography (Arhiptsev et al., 2021; Burkell & Gosse, 2019; Delfino, 2019; Karasavva & Noorbhai, 2021; Öhman, 2020) legal and ethical aspects of deepfakes (Citron & Chesney, 2019; Franks & Waldman, 2018; Meskys et al., 2019) and the philosophy of deepfakes in moral terms (de Ruiter, 2021). A common theme that recurs in describing the risks of deepfakes is the ability for actors to take advantage of the 'liar's dividend' i.e. the ability for individuals to deny factual content as fake (Ahmed, 2023).

One of the dominant concerns of deepfake disinformation in social and political contexts is the ability for it to sow social divisions (Harris, 2021), which can take place along the lines of political affiliation, gender, race, and class, while simultaneously lowering trust in institutions and authorities (Helmus, 2022). As tertiary educational institutions are tasked with providing education and care for students, this is also of great potential impact. Social discontent, instability, and lowering of trust in the institution are all specific effects of deepfakes which may be amplified within the context of a university or college campus.

Negative impacts also extend to the individual. Voice authentication of secure information such as bank accounts is no longer viable with the advent of vocal deepfakes (OpenAI, 2024), and deepfake technology can be used to produce non-consensual media about a person or persons, which can invade privacy, damage individuals and families or be used for identity theft (de Rancourt-Raymond & Smaili, 2022). This abuse of voice synthesis technology has already been witnessed in education, with the deepfake of a US principal's voice circulated to the community by a disgruntled employee (Associated Press, 2024). Image deepfakes can be used for the creation of non-consensual explicit images, which may be specifically used to objectify and degrade women, or for cyber-bullying or exploitation (Burkell & Gosse, 2019; Farid, 2022). In one study on social media use and deepfakes in Kenya, a majority of respondents (n = 978) had been victims of sexual violence on social media, which included being victims of deepfakes (Mayoyo, 2023). Moreover, as legal regulations and procedures surrounding deepfakes in much of the world have not yet developed, there is little opportunity to achieve justice (Kietzmann et al., 2020).

There are also growing concerns about deepfakes being used to create child sexual abuse material (CSAM) (Hern, 2024). Consequently, there is a growing consensus among political and legal stakeholders that deepfakes require some form of regulation (Hern, 2024, p. 24; Langa, 2021), lest the veracity of all audio and visual media come under question, a scenario labelled an 'infocalypse' (Schick, 2020). In the absence of clear regulation in most jurisdictions, it is telling that developers themselves are reticent to release powerful models, with Microsoft currently planning no API access or applications built on its VASA-1 video avatar model (Xu et al., 2024), and OpenAI's hesitance to release voice synthesis and generation (OpenAI, 2024).

At the time of writing, no coherent legal and institutional framework is forthcoming; therefore, we contend that it is up to practitioners and researchers in educational technology to take the lead in addressing deepfakes in a tertiary education context, and that this represents a professional obligation to meaningfully engage with GenAI, so as not to let the agenda be controlled by other actors (Thompson et al., 2023).

Positive Applications of Deepfake Technologies

Despite often being associated with malicious applications, deepfake technology has numerous potential positive uses, including in university teaching and learning practice. Deepfakes can enhance video conferencing with digital avatars (Gambín et al., 2024), and are used in entertainment for recreating deceased celebrities and interactive experiences (Gambín et al., 2024; Kietzmann et al., 2020). They show promise in tourism (Gambín et al., 2024; Kleine, 2022)

and retail by enabling virtual product try-on (Kietzmann et al., 2020). Synthetic audio allows voice changes for audiobooks, replacing misspoken words (Kietzmann et al., 2020), and potentially providing voices to those who have lost the ability to speak (de Ruiter, 2021; OpenAI, 2024; Whittaker et al., 2021). In healthcare and journalism, deepfakes can anonymise data and protect identities (Pandey et al., 2021). Synthetic media can support educational campaigns (Farid, 2022), advertising (Campbell et al., 2022), and e-commerce (Bode et al., 2021), which may indirectly benefit educational institutions (for instance, in recruitment and marketing efforts).

Our review also found there are potential beneficial applications of deepfake technology and synthetic media which apply to teaching and learning in a higher education context. For example, Caporusso (2021) describes the potential for "digital twins" which could be used in personalised learning, healthcare, and entertainment contexts and Westerlund (2019) points to the use of deepfake voice synthesis to create educational media. In the arts, Kwok & Koh (2021) discuss Florida's Salvador Dali museum, which uses a deepfake of the artist to educate visitors, and Cheres & Groza (2023) produced an AI-based at installation to raise awareness of the threats of social media, using deepfake technology to highlight the ways in which users provide substantial training data to AI-models. Pataranutaporn's (2024) doctoral thesis and concurrent studies explore numerous positive and educative applications of the technology, including AI-generated mentors, digital health counselling, and virtual study buddies. These studies explore the potential of using deepfake audio/visual models of celebrities, fictional characters, and inspiring role models, finding benefits for students' motivation, positive affect towards courses, and overall rating of the efficacy of AI-based tutors. As sites of research and scientific enquiry, tertiary educational institutions can play a key role in promoting and advocating for these positive uses of deepfakes.

This review reveals both beneficial and adverse impacts of deepfakes relevant to higher education. However, the lack of research in these contexts signifies a gap, which our research agenda intends to address. While comprehensive and confirming similar results to that of Godulla et al. (2021) and Gil et al. (2023), our review is limited by the lack of formal quality assessment of included studies. Nonetheless, it provides an account of current literature on deepfakes and synthetic media in education. The creation of low-cost, on-demand instructional materials, including the capacity to generate multilingual resources, is perhaps one of the greatest potential benefits in education. Similarly, the use of synthetic media to create simulations, scenarios, or synthetic datasets could facilitate course creations. Finally, in a fashion similar to the extant use of synthetic media for entertainment purposes, education providers could use these technologies to make engaging videos featuring university leadership or thought leaders in industries and fields, making attractive resources for students.

The aim of our review and exploration of deepfakes as a topic is to relate it to higher education. However, while potential use cases with beneficial applications have been detailed above, at the time of writing such potential benefits have yet to be realised. On the other hand, there are multiple cases reported on a frequent basis regarding malicious applications of deepfake technology in educational contexts (e.g. Morgan and Hales, 2024). Consequently, in the below we focus on describing more fully the risks to tertiary education of deepfake technology, and potential counterstrategies that may mediate their effects.

Translating the Risks of Deepfakes to a Tertiary Educational Context

Cyberbullying of students or faculty

Cyberbullying is prevalent in universities, with one study finding that 38% of North American students (n = 439) knew a victim of cyberbullying (MacDonald & Roberts-Pittman, 2010). Although the definition of cyberbullying may vary cross-culturally (Akbulut & Eristi, 2011), core elements

include online harassment, threats, circulating embarrassing or sensitive information, and dissemination of visual media, such as videos of bullying instances, all of which can have serious consequences on mental health (Kowalski et al., 2014). The advent of deepfake technology has introduced a new and potentially more damaging form of cyberbullying.

Our literature review revealed that fraud and non-consensual pornography are prominent impacts of deepfake technology, and deepfakes may be used in the context of cyberbullying (Langguth et al., 2021). This may include creating videos of an individual making a claim which is controversial or damaging via audio or video (Associated Press, 2024), or through sexual bullying by creating non-consensual pornography, as noted in the Kenyan university context (Mayoyo, 2023).

A widely reported incident in April 2024 at Pikesville High School in Maryland, USA exemplifies this threat. A high school teacher was arrested for allegedly using AI to create a deepfake audio recording of the school principal making racist comments. This false audio led to the principal being placed on leave and receiving death threats, demonstrating the severe real-world consequences of deepfake-enabled cyberbullying (Looker, 2024). While this example occurred in a high school setting, it underscores the potential for similar incidents in higher education, where the stakes for professional reputations and career implications may be even higher. In universities, cyberbullying may have immediate career implications for students upon graduation (Cunningham et al., 2015), making it an especially prominent risk. As accessibility and barriers to the entry of deepfake technology continue to decrease, and the convincing nature of the technology increases, cyberbullying deepfakes become an ever-more likely scenario, meaning that institutions must proactively find ways to address it.

Academic Dishonesty

Initial concerns about AI and academic dishonesty focused on detecting AI-generated text and using GenAI tools to violate academic integrity (Cotton et al., 2023; Perkins, 2023; Zhou et al., 2024). However, deepfakes pose an equal risk. While low-tech methods of manipulating research data have existed for centuries, AI-based image fraud may be harder to detect (Sundar et al., 2021). Deepfakes could create fake data, manipulate findings, or distribute false material to sabotage other students, and this risk is amplified by increasing pressures in competitive educational systems (Roe, 2022). For example, a researcher may use deepfake technology to generate audio of interviews which specific individuals which never occurred or provide evidence of experiments that did not take place. In health sciences specifically, it may be possible to generate synthetic media of medical imaging (e.g. X-rays) to use as research data. In the realm of academic references, unscrupulous applicants for grants, jobs, or other competitive processes may be able to use deepfake technology to impersonate high-profile referees, thus giving an unfair advantage. Deepfake technology could even be used to show an individual attending a graduation ceremony from a school that they did not attend, or receiving an award that was given to someone else.

Among a variety of education stakeholders (including undergraduate and graduate students and educators), Doss et al. (2023) found that between 27% and 50% of participants could not identify deepfake videos, meaning that believable, fraudulent videos, or other deepfake materials could impact academic honesty and integrity. Deepfakes can even extend to admission fraud, being used to create fake application material, audio or video testimonials, or recommendations, thus impacting equity and inclusion in higher education. Diploma mills which mimic established universities continue to threaten the validity and integrity of qualifications (Roe & Perkins, 2023), and deepfakes could conceivably be used to create false video testimonials or endorsements from high-profile academics or institutions.

Finally, existing security measures in online proctoring systems (OPS) may be at risk from deepfake technologies. Since OPS may rely on biometric data such as facial and voice recognition (Slusky, 2020), deepfake technologies which can accurately reproduce both images and voice may be used through contract cheating or other forms of academic misconduct. OPS which rely on surveillance and monitoring technologies requiring students to be visible through video software, may be vulnerable to convincing avatars which can be added to livestreaming, a technology which is already commercially available through applications such as HeyGen and Synthesia.

Decaying Institutional Trust and Reputation

Deepfakes can propagate disinformation and thus lead to the acquisition of false beliefs (Fallis, 2021). In a higher education context, deepfakes could be produced regarding university leadership, staff, or students, in which they endorse or espouse controversial or untrue beliefs as has already occurred in a K-12 context (Associated Press, 2024). Furthermore, younger students are vulnerable to misinformation (Doss et al., 2023) and may be more impressionable because of social factors (Gwon & Jeong, 2018), making them targets for disinformation campaigns (Ali et al., 2021). Consequently, the risk for higher education students to be affected by deepfakes in a way that damages the reputation of the institution or community spirit may be larger than in the general population.

In a highly marketised competitive landscape, reputation plays a vital role in universities (Angliss 2022). As the main risks of deepfakes to businesses are reputation and trustworthiness (Mustak et al., 2023), such risks may also affect the operations, aspirations, and financial success of higher education institutions (HEIs). Such reputational damage may be irreparable (Moerschell & Novak, 2020) and should be treated seriously by university leadership.

Pataranutaporn (2024) also identifies critical ethical issues through various studies, particularly focusing on AI-generated characters and their potential for mis-portrayal and disinformation. For instance, in the analysis of AI-generated characters as virtual instructors, the risk of these systems disseminating false or misleading information is underscored, especially when these characters are perceived as authoritative figures (Pataranutaporn, 2024, p.76). The study on deceptive AI explanations demonstrates how AI systems can generate explanations that are more persuasive than truthful ones, posing a significant risk in educational contexts where deepfakes could be used to manipulate academic discourse or fabricate evidence (Pataranutaporn, 2024, p. 156). These findings suggest the ethical use of deepfakes in higher education requires stringent oversight and the development of policies that safeguard against these risks, ensuring that synthetic media is used to enhance, rather than compromise, academic integrity.

Strategies for Countering the Risks of Deepfakes in Higher Education

In addressing our second aim, we contend that the greatest potential threats to tertiary education stakeholders regarding deepfakes are cyberbullying, reputational damage, and academic dishonesty. As educational institutions do not yet have effective ways of preparing students to tackle the issues of deepfakes and disinformation (Naffi et al., 2023), we propose three strategies based on the extant literature in the fields of cyberbullying and crisis management for institutions to begin a proactive approach to countering the potential threat of deepfakes.

In explaining the first counterstrategy, it is important to note that many university codes of conduct do not explicitly address cyberbehaviors (Faucher et al., 2015). Anti-cyberbullying policies should specifically refer to behaviours like creating non-consensual deepfakes or spreading disinformation, while avoiding overbreadth that infringes on students' rights (O'Connor et al., 2018). No universal approach exists, as institutions operate in diverse cultural and legal contexts. For example, the U.K. has criminalised non-consensual sexually explicit deepfakes (Milligan,

2024), enabling aligned institutional reporting policies. Globally, a preventative, context-specific approach is necessary (O'Connor et al., 2018), with anonymous reporting to reduce retaliation risks (Cunningham et al., 2015). Given the gendered nature of deepfakes and higher rates of cyberbullying affecting university women (Faucher et al., 2015), gender-specific policies should be considered. This is one of the most pressing issues regarding deepfakes in education, with numerous examples of nonconsensual pornography appearing. Indeed, in the Australian context, a 'crisis' of deepfake pornography is occurring in K-12 education (Schmidt, 2024), while in Korea, encrypted chatgroups to create and spread deepfake pornography have been linked to specific schools and universities (Mackenzie & Marsh, 2024).

Table 2

Risks of Deepfakes in Tertiary Education and Potential Counterstrategies

Risk Name	Risk Description	Counterstrategy
Cyberbullying	The creation of deepfakes which depict teachers or students in explicit or controversial situations. This has implications for cyberbullying and the associated mental health and trauma that accompanies bullying incidents.	Universities must not wait for legal frameworks to catch up to technology, rather they should proactively create policy frameworks for dealing with such cases, including gender-specific policies if necessary.
Deepfake Illiteracy	Students, teachers, and other stakeholders may be unaware of deepfakes and may not have received adequate training in identifying misinformation or manipulated media. This may lead to a higher level of belief in deepfake incidents and less critical analysis of deepfake content.	Digital and media literacy efforts for staff, students, and stakeholders must begin to incorporate identification of deepfake disinformation to mediate the impact of a deepfake incident.
Ineffective Response to a Deepfake Incident	Deepfake incidents are a new form of AI threat. Universities must proactively plan for an incident. An inefficient response to a deepfake incident may lead to worse outcomes for those affected.	A crisis management plan which details a communication strategy for a deepfake incident should be implemented. This can be part of a broader plan to deal with advanced technological threats to a university's operation and ensures that impact from a deepfake incident is minimised.

In relation to the second counterstrategy, educational interventions may be undertaken to inform about the risks of deepfakes and synthetic media. Deepfakes may be specifically included as a component in anti-cyberbullying awareness interventions, given that education and training can be effective in reducing the prevalence of cyberbullying (Li, 2007). However, educating stakeholders about deepfakes in general may also be beneficial, as it can be part of a broader media literacy effort in an age of growing disinformation. In one of the few studies in this area, Ali et al. (2021) described a series of activities to educate middle-school students about how deepfakes are created and identified as a method of fostering critical literacy. Such educational programs could also be valuable for university students and academic staff.

By undertaking deepfake education, students and faculty can be taught what deepfakes are and how to suspect them. As Bearman et al. (2024) argue, building students' evaluative judgment and media literacy capabilities is essential in a world where misleading deepfakes are increasingly prevalent, and this approach can help to 'pre-bunk' or 'inoculate' students and staff to be sensitive and aware of deepfakes as a misinformation strategy (Horvitz, 2022). Sharing some of the more relevant research findings on the psychology of deepfake impacts, such as the fact that deepfakes may even alter memories of events (Murphy et al., 2023), may also be beneficial in highlighting the potential consequences of deepfake disinformation.

Targeted educational interventions and training programs can help higher education stakeholders develop resilience against deepfake technologies and learn to guard against such threats (Hancock & Bailenson, 2021). These interventions combat educational fraud, promote trust and integrity, and reduce academic misconduct (Perkins et al., 2020). Familiarising students with the potential harms of deepfakes may reduce misuse.

The third counterstrategy relates more broadly to managing crises in institutions of higher education. Higher education institutions often treat crises as rare occurrences, making them ill-equipped to respond (Booker Jr., 2014). In reference to a deepfake 'crisis' in which there is an immediate and high level of risk to the integrity of the institution, developing a specific management and communication plan is necessary. For example, a deepfake crisis may include the production of fabricated content of an academic leader saying something inflammatory or insensitive, or show them doing something that would cause reputational damage. Such disinformation may spread quickly via social media, and a coordinated plan to counter it may focus on media strategies aimed at limiting and repairing reputational damages. Such plans can also be proactive, including the development of communication strategies to inform people who are at risk of being targeted as well as other stakeholders and the media (Moerschell & Novak, 2020).

Building on Moerschell and Novak (2020), several examples of crisis management planning can be adapted to deal with deepfake incidents. First, practical considerations can involve developing a 'dark site' to be published only in the event of a crisis to communicate with stakeholders, which may enable institutions to combat any disinformation spread by a deepfake and provide accurate information to the recipients. A social media page can also be developed to manage crises, including deepfake crises, act against conflicting social media information, and release information quickly and strategically (Moerschell & Novak, 2020). Fundamentally, crisis management plans for deepfakes depend on the information contained in the falsified content. However, proactive communication planning and crisis management may mitigate reputational damage to institutions, students, and staff involved. Given that the effectiveness of organisational responses to disinformation and fake news varies depending on several factors (Vafeiadis et al., 2019), formulating a set of protocols for deepfake crisis management is an important research task.

Deepfake Education Research Agenda

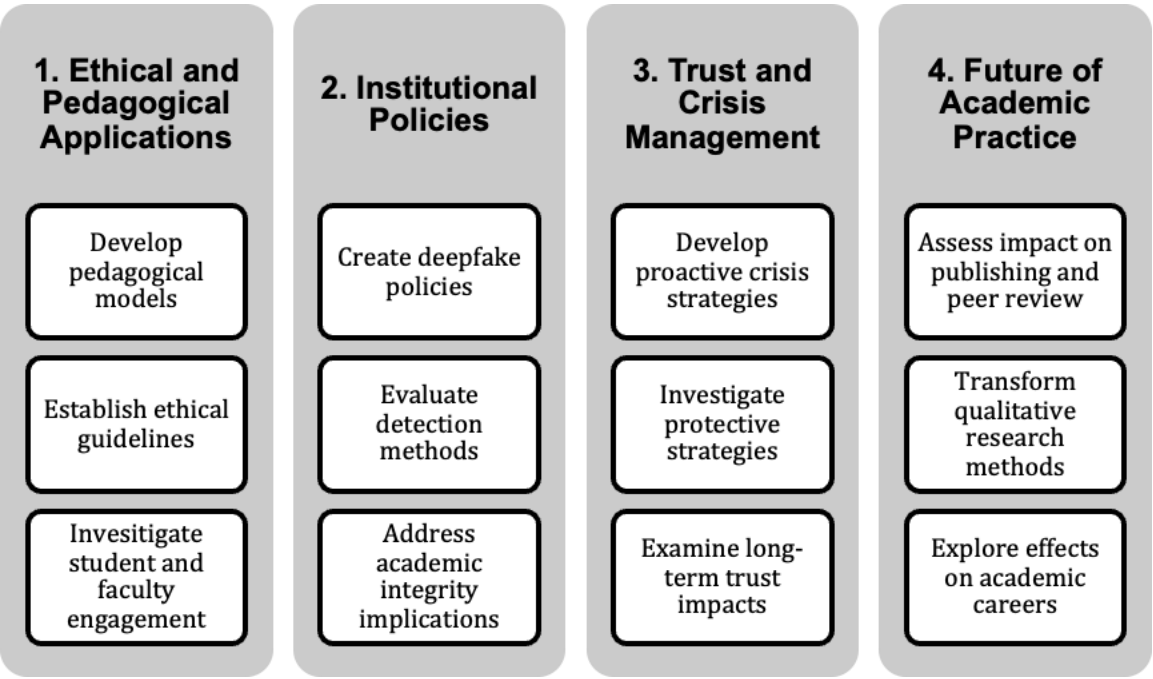
Academic research is vital in advancing deepfake technology (Farid, 2022), with a dramatic increase in journal articles in recent years (Gil et al., 2023). However, much of this research focuses on detection, creation, and political dimensions, with less emphasis on educational value or safeguarding students, institutions, and faculty. The lack of existing specific research on deepfakes in higher education contexts presents a significant challenge. Without dedicated studies in this area, we cannot confidently assume that findings from other industries or contexts directly apply to tertiary education. The unique dynamics of academic environments, including the emphasis on intellectual property, the nature of student-faculty relationships, and the critical role

of institutional reputation, may result in different implications and challenges compared to those observed in other sectors.

Our review has revealed significant gaps in research specifically addressing deepfakes and synthetic media in higher education contexts. While studies have explored technological solutions for deepfake detection (Giudice et al., 2021; He et al., 2021), potential societal impacts (Citron & Chesney, 2019; Harris, 2021), and ethical concerns (de Ruiter, 2021), there is a dearth of research examining how these technologies intersect with teaching, learning, and institutional practices in higher education. This gap is particularly concerning given the rapid advancement of deepfake technology and its potential to both enhance and disrupt educational processes (Caporusso, 2021; Westerlund, 2019). Consequently, we formulate a research agenda for deepfakes in tertiary educational contexts as below, leaning on four pillars.

Figure 1

A Deepfake Research Agenda for Tertiary Education



1. Ethical and Pedagogical Applications of Synthetic Media in Learning

Our literature review highlighted potential positive uses of deepfakes and synthetic media in education (Caporusso, 2021; Westerlund, 2019) but revealed a significant gap in understanding how to effectively integrate these technologies into higher education curricula. The proposed research priorities in this theme directly address this gap by calling for the development of pedagogical models and ethical guidelines specific to higher education contexts. This theme also responds to the lack of research on student and faculty engagement with synthetic media in learning environments, an area only tangentially explored in studies with younger students (Ali et al., 2021; Blankenship, 2021). Building on the work of Pataranutaporn (2024) and investigating the potential benefits of AI-generated advisors and tutors, the potential for misinformation and disinformation will be of primary concern when developing pedagogical applications of synthetic media technologies in education.

We recommend a focus on the following areas exploring the ethical and pedagogical applications of synthetic media:

a. Develop pedagogical models for integrating synthetic media into curricula without compromising authenticity.
b. Establish ethical guidelines and disclosure practices for using deepfakes in education.
c. Investigate student and faculty engagement with synthetic media, including its potential to enhance inclusivity and accessibility in education.

2. Institutional Policies for Deepfake Management

The literature review identified a clear gap in institution-specific policies and frameworks for managing deepfakes, with most existing research focusing on broad societal implications (Citron & Chesney, 2019). The research priorities under this theme directly address this gap by proposing the development of comprehensive policy frameworks specific to higher education. Additionally, while the literature review found numerous studies on deepfake detection (e.g., Giudice et al., 2021; He et al., 2021), there was little exploration of how these methods could be applied in academic settings. Although evidence suggests that HEIs have been slow to adapt to the rapid growth of GenAI tools (Perkins & Roe, 2023), given the increasing institutional knowledge bases in these areas, institutions may now be more capable of producing effective deepfake policy, but require evidence to base this on. Ongoing work, such as the R.E.A.L framework for managing deepfake risks (Kietzmann et al., 2020), offers potential starting points for this.

While some efforts are being made to educate younger students about deepfakes (Ali et al., 2021; Blankenship, 2021), empirical studies should test and compare specific pedagogical strategies for building university students' and staff critical evaluation skills. This is particularly important given the broader challenges faced by academia regarding the detection of text-based GenAI content (Chaka, 2023, 2024; Perkins et al., 2023, 2024; Weber-Wulff et al., 2023). This theme's focus on investigating detection and authentication methods in academic contexts fills this critical gap. The priorities in this area are:

a. Develop comprehensive policy frameworks addressing deepfake challenges in higher education.
b. Investigate effectiveness of deepfake detection and content authentication in academic settings.
c. Examine policy implications for academic integrity and assessment.

3. Impact on Institutional Trust and Crisis Management

Our review revealed significant concerns about the potential of deepfakes to erode institutional trust and create crises within universities but found little research on how higher education institutions can pre-emptively manage these risks. The research priorities in this theme directly address this gap by proposing investigations into proactive crisis management strategies and efforts specific to the higher education context. We echo the point made by Karasavva and Noorbhai (2021) that research on how to best support victims of deepfake pornography and other specific issues is warranted given that deepfakes are likely to become commonplace (Hancock & Bailenson, 2021) and continue to grow over time (Kietzmann et al., 2020). The focus on long-term trust restoration processes also addresses a gap in understanding the prolonged effects of deepfake incidents on academic communities. We suggest the following priorities to address the impact of deepfakes on institutional trust:

a. Develop proactive crisis management strategies specific to deepfake incidents in higher education.
b. Investigate strategies to protect academic communities against deepfake disinformation.
c. Examine long-term trust restoration processes and the impact of deepfakes on student trust in educational institutions.

4. Deepfakes and the Future of Academic Practice

The literature review found no substantial research on how deepfakes might transform core academic practices such as publishing, peer review, and research methods. Additionally, while some literature touched on the general impact of deepfakes on identity and self-presentation (de Seta, 2021), there was a clear gap in understanding these effects within the specific context of academia. The priority focusing on long-term implications for academic careers and identity formation addresses this gap. This final theme is ambitious but necessary for understanding the long-term impacts of deepfake technologies in higher education. The research priorities in this theme directly address this gap by proposing investigations into these areas:

a. Investigate impacts on academic publishing and peer review processes.
b. Examine transformation of qualitative research methods and develop ethical guidelines for deepfake use in research.
c. Explore long-term implications on academic career progression, expertise recognition, and identity formation in academic contexts.

Conclusion

Deepfakes are not the first form of digital manipulation and will not be the last (Farid, 2022). However, we believe that the findings of this study should be cause for significant alarm for universities and those who work and study at them. We note that while deepfakes are a dual-use technology with potential positive applications (Gamage et al., 2022) and that the technology may offer some beneficial use-cases for teaching and learning specifically (e.g. virtual lectures) many of the examples we noted relate to more nefarious uses. Such uses may disrupt teaching and learning activities, impact the mental health of those affected, and significantly damage the reputation of institutions. Furthermore, the underlying credibility of research, researchers, and data may be less secure than ever before in a world where deepfake technology is common.

Specifically, we note that while deepfake technologies are still new and have yet to gain widespread awareness, examples in the USA, Korea, and Australia demonstrate that explicit, nonconsensual deepfakes are being spread in educational contexts at a rapid rate, and that these cases have a gendered dimension. As deepfake technology matures and it becomes harder to tell fact from fiction, the potential impacts of malicious deepfakes grows even more concerning. Aside from the counterstrategies that we have posited, there is also a pressing and urgent need for a greater, interdisciplinary research effort. Consequently, we call for research that addresses four areas of a broader agenda in higher education, including:

1. Ethical and Pedagogical Applications
2. Institutional Policy Frameworks
3. Institutional Trust and Crisis Management
4. Future of Academic Practice

By pursuing this research agenda, the academic community can aim to remain open to the potential benefits to teaching and learning that new technologies may afford, while equally forming a coherent and robust policy response to the concerning rise of deepfakes.

Acknowledgements

The authors declare no conflict of interest. This research received no external funding. JR: Conceptualisation, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. MP: Conceptualisation, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. LF: Investigation, Writing - Review & Editing. This study used Generative AI tools (ChatGPT-4 and Claude 3 Opus) for draft text creation, revision, and editorial purposes throughout the production of the manuscript. The authors reviewed, edited, and take responsibility for all outputs of the tools used in this study.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (n.d.). *Protecting World Leaders Against Deep Fakes*.

Ahmed, S. (2023). Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media & Society*, *25*(5), 1108–1129. https://doi.org/10.1177/14614448211019198

AI Voice Generator & Text to Speech. (2024). ElevenLabs. https://elevenlabs.io

Akbulut, Y., & Eristi, B. (2011). Cyberbullying and victimisation among Turkish university students. *Australasian Journal of Educational Technology*, *27*(7), Article 7. https://doi.org/10.14742/ajet.910

Akhtar, Z. (2023). Deepfakes Generation and Detection: A Short Survey. *Journal of Imaging*, *9*(1), Article 1. https://doi.org/10.3390/jimaging9010018

Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence*, *2*, 100040. https://doi.org/10.1016/j.caeai.2021.100040

Anderson, N., Belavy, D. L., Perle, S. M., Hendricks, S., Hespanhol, L., Verhagen, E., & Memon, A. R. (2023). AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sport & Exercise Medicine*, *9*(1), e001568. https://doi.org/10.1136/bmjsem-2023-001568

Angliss, K. (2022). An Alternative Approach to Measuring University Reputation. *Corporate Reputation Review*, *25*(1), 33–49. https://doi.org/10.1057/s41299-021-00110-y

Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, *27*(4), zmac008. https://doi.org/10.1093/jcmc/zmac008

Arhiptsev, I. N., Николаевич, А. И., Aleksandrov, A. N., Николаевич, А. А., Maksimenko, A. V., Владимирович, М. А., Ozerov, К. І., & Игоревич, О. К. (2021). Pornographic deepfake: Fiction or virtual reality? *Sociopolitical Sciences*, *11*(1), Article 1. https://doi.org/10.33693/2223-0093-2021-11-1-69-74

Associated Press. (2024, April 25). US teacher charged with using AI to frame principal with racist audio. *The Guardian*. https://www.theguardian.com/us-news/2024/apr/25/maryland-teacher-ai-principal

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 1–13. https://doi.org/10.1080/02602938.2024.2335321

Blankenship, R. J. (2021). Educational Responsibility in the Deepfake Era: A Primer for TPACK Reform. In *Deep Fakes, Fake News, and Misinformation in Online Teaching and Learning Technologies* (pp. 1–23). IGI Global. https://doi.org/10.4018/978-1-7998-6474-5.ch001

Bode, L., Lees, D., & Golding, D. (2021). The Digital Face and Deepfakes on Screen. *Convergence*, *27*(4), 849–854. https://doi.org/10.1177/13548565211034044

Booker Jr., L. (2014, January 1). *Crisis Management: Changing Times for Colleges.* | *Journal of College Admission* | *EBSCOhost*. https://openurl.ebsco.com/contentitem/gcd:97394361?sid=ebsco:plink:crawler&id=ebsco:gcd:97394361

Burkell, J., & Gosse, C. (2019). Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday*. https://doi.org/10.5210/fm.v24i12.10287

Busacca, A., & Monaca, M. A. (2023). Deepfake: Creation, Purpose, Risks. In D. Marino & M. A. Monaca (Eds.), *Innovations and Economic and Social Changes due to Artificial Intelligence: The State of the Art* (pp. 55–68). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33461-0_6

Campbell, C., Plangger, K., Sands, S., & Kietzmann, J. (2022). Preparing for an Era of Deepfakes and AI-Generated Ads: A Framework for Understanding Responses to Manipulated Advertising. *Journal of Advertising*, *51*(1), 22–38. https://doi.org/10.1080/00913367.2021.1909515

Caporusso, N. (2021). Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology. In T. Ahram (Ed.), *Advances in Artificial Intelligence, Software and Systems Engineering* (Vol. 1213, pp. 235–241). Springer International Publishing. https://doi.org/10.1007/978-3-030-51328-3_33

Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021). Deepfake: An Overview. In P. K. Singh, S. T. Wierzchoń, S. Tanwar, M. Ganzha, & J. J. P. C. Rodrigues (Eds.), *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security* (pp. 557–566). Springer. https://doi.org/10.1007/978-981-16-0733-2_39

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*, *6*(2), Article 2. https://doi.org/10.37074/jalt.2023.6.2.12

Chaka, C. (2024). Accuracy pecking order – How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. *Journal of Applied Learning and Teaching*, *7*(1), Article 1. https://doi.org/10.37074/jalt.2024.7.1.33

Cheres, I., & Groza, A. (2023). The profile: Unleashing your deepfake self. *Multimedia Tools and Applications*, *82*(20), 31839–31854. https://doi.org/10.1007/s11042-023-14568-x

Citron, D., & Chesney, R. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*(6), 1753.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1–12. https://doi.org/10.1080/14703297.2023.2190148

Cunningham, C. E., Chen, Y., Vaillancourt, T., Rimas, H., Deal, K., Cunningham, L. J., & Ratcliffe, J. (2015). Modeling the anti-cyberbullying preferences of university students: Adaptive choice-based conjoint analysis. *Aggressive Behavior*, *41*(4), 369–385. https://doi.org/10.1002/ab.21560

de Rancourt-Raymond, A., & Smaili, N. (2022). The unethical use of deepfakes. *Journal of Financial Crime*, *30*(4), 1066–1077. https://doi.org/10.1108/JFC-04-2022-0090

de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology*, *34*(4), 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

de Seta, G. (2021). Huanlian, or changing faces: Deepfakes on Chinese digital media platforms. *Convergence*, *27*(4), 935–953. https://doi.org/10.1177/13548565211030185

Delfino, R. A. (2019). Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Fordham Law Review*, *88*, 887.

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*(7), 2072–2098. https://doi.org/10.1177/1461444820925811

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics*, *26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., Bush, L., & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports*, *13*(1), 13429. https://doi.org/10.1038/s41598-023-39944-3

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, *34*(4), 623–643. https://doi.org/10.1007/s13347-020-00419-2

Farid, H. (2022). Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety*, *1*(4), Article 4. https://doi.org/10.54501/jots.v1i4.56

Faucher, C., Jackson, M., & Cassidy, W. (2015). When Online Exchanges Byte: An Examination of the Policy Environment Governing Cyberbullying at the University Level. *Canadian Journal of Higher Education*, *45*(1), 102–121.

Feher, K. (n.d.). Exploring AI media. Definitions, conceptual model, research agenda. *Journal of Media Business Studies*, 1–24. https://doi.org/10.1080/16522354.2024.2340419

Franks, M. A., & Waldman, A. E. (2018). Sex, lies, and videotape: Deep fakes and free speech delusions. *Md. L. Rev.*, *78*, 892.

Gamage, D., Ghasiya, P., Bonagiri, V., Whiting, M. E., & Sasahara, K. (2022). Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517446

Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review*, *57*(3), 64. https://doi.org/10.1007/s10462-023-10679-x

Gil, R., Virgili-Gomà, J., López-Gil, J.-M., & García, R. (2023). Deepfakes: Evolution and trends. *Soft Computing*, *27*(16), 11295–11318. https://doi.org/10.1007/s00500-023-08605-y

Giudice, O., Guarnera, L., & Battiato, S. (2021). Fighting Deepfakes by Detecting GAN DCT Anomalies. *Journal of Imaging*, *7*(8), Article 8. https://doi.org/10.3390/jimaging7080128

Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with deepfakes – an interdisciplinary examination of the state of research and implications for communication studies. *Studies in Communication and Media*, *10*(1), 72–96. https://doi.org/10.5771/2192-4007-2021-1-72

Gwon, S. H., & Jeong, S. (2018). Concept analysis of impressionability among adolescents and young adults. *Nursing Open*, *5*(4), 601–610. https://doi.org/10.1002/nop2.170

Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2022). You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media. *Social Media + Society*, *8*(3), 20563051221116346. https://doi.org/10.1177/20563051221116346

Hancock, J. T., & Bailenson, J. N. (2021). The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 149–152. https://doi.org/10.1089/cyber.2021.29208.jth

Harris, K. R. (2021). Video on demand: What deepfakes do and how they harm. *Synthese*, *199*(5), 13373–13391. https://doi.org/10.1007/s11229-021-03379-y

He, Y., Yu, N., Keuper, M., & Fritz, M. (2021). *Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis* (arXiv:2105.14376). arXiv. https://doi.org/10.48550/arXiv.2105.14376

Helmus, T. C. (2022). *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*. RAND Corporation. https://www.jstor.org/stable/resrep42027

Hern, A. (2024, March 17). Labour considers 'nudification' ban and cross-party pledge on AI deepfakes. *The Guardian*. https://www.theguardian.com/politics/2024/mar/17/ai-deepfakes-misinformation-nudification-labour

Hernandez-Ortega, J., Tolosana, R., Fierrez, J., & Morales, A. (2020). *DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation* (arXiv:2010.00400). arXiv. https://doi.org/10.48550/arXiv.2010.00400

HeyGen—AI Spokesperson Video Creator. (2024). HeyGen. https://app.heygen.com

Horvitz, E. (2022). On the Horizon: Interactive and Compositional Deepfakes. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 653–661. https://doi.org/10.1145/3536221.3558175

Huang, Y., Juefei-Xu, F., Wang, R., Guo, Q., Ma, L., Xie, X., Li, J., Miao, W., Liu, Y., & Pu, G. (2020). FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction. *Proceedings of the 28th ACM International Conference on Multimedia*, 1217–1226. https://doi.org/10.1145/3394171.3413732

Jacobsen, B. N., & Simpson, J. (2023). The tensions of deepfakes. *Information, Communication & Society*, *0*(0), 1–15. https://doi.org/10.1080/1369118X.2023.2234980

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access*, *8*, 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

Karasavva, V., & Noorbhai, A. (2021). The Real Threat of Deepfake Pornography: A Review of Canadian Policy. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 203–209. https://doi.org/10.1089/cyber.2020.0272

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

Kleine, F. (2022). *Perception of Deepfake Technology—The Influence of the Recipients' Affinity for Technology on the Perception of Deepfakes*.

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618

Kwok, A. O. J., & Koh, S. G. M. (2021). Deepfake: A social construction of technology perspective. *Current Issues in Tourism*, *24*(13), 1798–1802. https://doi.org/10.1080/13683500.2020.1738357

Langa, J. (2021). Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes. *Boston University Law Review*, *101*, 761.

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.632317

Li, Q. (2007). Bullying in the new playground: Research into cyberbullying and cyber victimisation. *Australasian Journal of Educational Technology*, *23*(4), Article 4. https://doi.org/10.14742/ajet.1245

Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. *Australasian Journal of Educational Technology*, *39*(1), Article 1. https://doi.org/10.14742/ajet.8695

Looker, R. (2024, April 26). Baltimore high school teacher arrested over deepfake racist audio of principal. BBC. https://www.bbc.com/news/world-us-canada-68907895

MacDonald, C. D., & Roberts-Pittman, B. (2010). Cyberbullying among college students: Prevalence and demographic differences. *Procedia - Social and Behavioral Sciences*, *9*, 2003–2009. https://doi.org/10.1016/j.sbspro.2010.12.436

Mackenzie, J., & Marsh, N. (2024, September 28) South Korea faces deepfake porn emergency. BBC. https://www.bbc.com/news/articles/cg4yerrg451o

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE*, *18*(8), e0285333. https://doi.org/10.1371/journal.pone.0285333

Mak, S., & Thomas, A. (2022). Steps for Conducting a Scoping Review. *Journal of Graduate Medical Education*, *14*(5), 565–567. https://doi.org/10.4300/JGME-D-22-00621.1

Mayoyo, N. (2023). The Influence of Social Media Use in the Wake of Deepfakes on Kenyan Female University Students' Perceptions on Sexism, Their Body Image and Participation in Politics. In K. Langmia (Ed.), *Black Communication in the Age of Disinformation: DeepFakes and Synthetic Media* (pp. 89–103). Springer International Publishing. https://doi.org/10.1007/978-3-031-27696-5_5

Meskys, E., Kalpokiene, J., Jurcys, P., & Liaudanskas, A. (2019). *Regulating Deep Fakes: Legal and Ethical Considerations* (SSRN Scholarly Paper 3497144). https://papers.ssrn.com/abstract=3497144

Milligan, E. (2024, April 16). *U.K. Criminalizes Creating Sexually Explicit Deepfake Images*. TIME. https://time.com/6967243/uk-criminalize-sexual-explicit-deepfake-images-ai/

Moerschell, L., & Novak, S. S. (2020). Managing crisis in a university setting: The challenge of alignment. *Journal of Contingencies and Crisis Management*, *28*(1), 30–40. https://doi.org/10.1111/1468-5973.12266

Morgan, C. & Hales, H. (2024, June 12). Student AI Deepfake Images Reflective of Porn Crisis. Australian Associated Press. https://www.aap.com.au/news/student-deepfakes-reflective-of-school-porn-crisis/

Murillo-Ligorred, V., Ramos-Vallecillo, N., Covaleda, I., & Fayos, L. (2023). Knowledge, Integration and Scope of Deepfakes in Arts Education: The Development of Critical Thinking in Postgraduate Students in Primary Education and Master's Degree in Secondary Education. *Education Sciences*, *13*(11), Article 11. https://doi.org/10.3390/educsci13111073

Murphy, G., Ching, D., Twomey, J., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *PLOS ONE*, *18*(7), e0287503. https://doi.org/10.1371/journal.pone.0287503

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, *154*, 113368. https://doi.org/10.1016/j.jbusres.2022.113368

Naffi, N., Charest, M., Danis, S., Pique, L., Davidson, A.-L., Brault, N., Bernard, M.-C., & Barma, S. (2023). Empowering Youth to Combat Malicious Deepfakes and Disinformation: An Experiential and Reflective Learning Experience Informed by Personal Construct Theory. *Journal of Constructivist Psychology*, *0*(0), 1–22. https://doi.org/10.1080/10720537.2023.2294314

Nema, P. (2021). Understanding copyright issues entailing deepfakes in India. *International Journal of Law and Information Technology*, *29*(3), 241–254. https://doi.org/10.1093/ijlit/eaab007

O'Connor, K., Drouin, M., Davis, J., & Thompson, H. (2018). Cyberbullying, revenge porn and the mid-sized university: Victim characteristics, prevalence and students' knowledge of university policy and reporting procedures. *Higher Education Quarterly*, *72*(4), 344–359. https://doi.org/10.1111/hequ.12171

Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, *22*(2), 133–140. https://doi.org/10.1007/s10676-019-09522-1

Online Safety Act, Pub. L. No. c. 50 (2023). https://www.legislation.gov.uk/ukpga/2023/50

OpenAI. (2024). *Navigating the Challenges and Opportunities of Synthetic Voices*. https://openai.com/blog/navigating-the-challenges-and-opportunities-of-synthetic-voices

Pandey, C. K., Mishra, V. K., & Tiwari, N. K. (2021). Deepfakes: When to Use It. *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, 80–84. https://doi.org/10.1109/SMART52563.2021.9676297

Pataranutaporn, P. (2024). Cyborg psychology: The art & science of designing human-AI systems that support human flourishing (Doctoral dissertation, Massachusetts Institute of Technology).

Pashentsev, E. (2023). The Malicious Use of Deepfakes Against Psychological Security and Political Stability. In E. Pashentsev (Ed.), *The Palgrave Handbook of Malicious Use of AI and Psychological Security* (pp. 47–80). Springer International Publishing. https://doi.org/10.1007/978-3-031-22552-9_3

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, *20*(2). https://doi.org/10.53761/1.20.02.07

Perkins, M., Gezgin, U. B., & Roe, J. (2020). Reducing plagiarism through academic misconduct education. *International Journal for Educational Integrity*, *16*(1), 3. https://doi.org/10.1007/s40979-020-00052-8

Perkins, M., & Roe, J. (2023). Decoding Academic Integrity Policies: A Corpus Linguistics Investigation of AI and Other Technological Threats. *Higher Education Policy*. https://doi.org/10.1057/s41307-023-00323-2

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023). Detection of GPT-4 Generated Text in Higher Education: Combining Academic Judgement and Software to Identify Generative AI Tool Misuse. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-023-09492-6

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). *GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education* (arXiv:2403.19148). arXiv. http://arxiv.org/abs/2403.19148

Porcile, G. J. A., Gindi, J., Mundra, S., Verbus, J. R., & Farid, H. (2024). *Finding AI-Generated Faces in the Wild* (arXiv:2311.08577). arXiv. https://doi.org/10.48550/arXiv.2311.08577

Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., & Zhao, J. (2020). DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. *Proceedings of the 28th ACM International Conference on Multimedia*, 4318–4327. https://doi.org/10.1145/3394171.3413707

Roe, J. (2022). Reconceptualizing academic dishonesty as a struggle for intersubjective recognition: A new theoretical model. *Humanities and Social Sciences Communications*, *9*(1), Article 1. https://doi.org/10.1057/s41599-022-01182-9

Roe, J., & Perkins, M. (2022). What are Automated Paraphrasing Tools and how do we address them? A review of a growing threat to academic integrity. *International Journal for Educational Integrity*, *18*(1), Article 1. https://doi.org/10.1007/s40979-022-00109-w

Roe, J., & Perkins, M. (2023). Welcome to the University of life, can I take your order? Investigating Life Experience Degree Offerings in Diploma mills. *International Journal for Educational Integrity*, *19*(1), Article 1. https://doi.org/10.1007/s40979-023-00138-z

Samoilenko, S. A., & Suvorova, I. (2023). Artificial Intelligence and Deepfakes in Strategic Deception Campaigns: The U.S. and Russian Experiences. In E. Pashentsev (Ed.), *The Palgrave Handbook of Malicious Use of AI and Psychological Security* (pp. 507–529). Springer International Publishing. https://doi.org/10.1007/978-3-031-22552-9_19

Saner, E. (2024, January 31). Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box.' *The Guardian*. https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box

Schick, N. (2020). *Deepfakes: The Coming Infocalypse*. Hachette UK.

Schmidt, N. (2024, June 12). Australian Schools Battling With AI Driven Deepfake Crisis. News.com.au. https://www.news.com.au/technology/online/proposed-deepfake-bill-not-enough-to-combat-unprecedented-crisis/news-story/ffb8e5f9a661ef9270b222c5ff4e0dec

Shahzad, H. F., Rustam, F., Flores, E. S., Luís Vidal Mazón, J., de la Torre Diez, I., & Ashraf, I. (2022). A Review of Image Processing Techniques for Deepfakes. *Sensors*, *22*(12), Article 12. https://doi.org/10.3390/s22124556

Silva, S. H., Bethany, M., Votto, A. M., Scarff, I. H., Beebe, N., & Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, *4*, 100217. https://doi.org/10.1016/j.fsisyn.2022.100217

Slusky, L. (2020). Cybersecurity of Online Proctoring Systems. *Journal of International Technology and Information Management*, *29*(1), COV3+. Gale Academic OneFile.

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps? *Journal of Computer-Mediated Communication*, *26*(6), 301–319. https://doi.org/10.1093/jcmc/zmab010

Suno. (2024). *My account | Suno*. https://accounts.suno.com/sign-in?redirect_url=https%3A%2F%2Fsuno.com%2Fcreate

Synthesia—Create studio-quality AI videos from text. (2024). https://www.synthesia.io/

Thompson, K., Corrin, L., & Lodge, J. M. (2023). AI in tertiary education: Progress on research and practice. *Australasian Journal of Educational Technology*, *39*(5), Article 5. https://doi.org/10.14742/ajet.9251

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, *64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

Udio. (2024). *Udio | AI Music Generator—Official Website*. Udio. https://udio.com

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1), 2056305120903408. https://doi.org/10.1177/2056305120903408

Vafeiadis, M., Bortree, D. S., Buckley, C., Diddi, P., & Xiao, A. (2019). Refuting fake news on social media: Nonprofits, crisis response strategies and issue involvement. *Journal of Product & Brand Management*, *29*(2), 209–222. https://doi.org/10.1108/JPBM-12-2018-2146

Vougioukas, K., Petridis, S., & Pantic, M. (2019). *Realistic Speech-Driven Facial Animation with GANs* (arXiv:1906.06337). arXiv. https://doi.org/10.48550/arXiv.1906.06337

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), Article 1. https://doi.org/10.1007/s40979-023-00146-z

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*. https://api.semanticscholar.org/CorpusID:214014129

Whittaker, L., Letheren, K., & Mulcahy, R. (2021). The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing. *Australasian Marketing Journal*, *29*(3), Article 3. https://doi.org/10.1177/1839334921999479

Xu, S., Chen, G., Guo, Y.-X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., & Guo, B. (2024). *VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time* (arXiv:2404.10667). arXiv. https://doi.org/10.48550/arXiv.2404.10667

Zhou, X., Zhang, J., & Chan, C. (2024). Unveiling Students' Experiences and Perceptions of Artificial Intelligence Usage in Higher Education. *Journal of University Teaching and Learning Practice*, *21*(06), Article 06. https://doi.org/10.53761/xzjprb23