This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology

John Twomey, Didier Ching, Matthew Peter Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy

***Abstract*—Deepfakes are a form of synthetic media that uses deep-learning technology to create fake images, video, and audio. The emergence of this technology has inspired much commentary and speculation from academics across a range of disciplines, who have contributed expert opinions regarding the implications of deepfake proliferation on fields such as law, politics, and entertainment. A systematic scoping review was carried out to identify, assemble, and critically analyze those academic narratives. The aim is to build on and critique previous attempts at defining the technology and categorizing the harms and benefits of deepfake technology. A range of databases were searched for relevant articles from 2017 to 2023, resulting in a large multi-disciplinary dataset of 102 papers, 181,659 words long, which were analyzed qualitatively through thematic analysis. Implications for future research include questioning the lack of research evidence for the supposed positives of deepfakes, recognizing the role that identity plays in deepfake technology, challenging the perceived accessibility/ believability of deepfakes, and proposing a more nuanced approach to the dichotomous "positive and negatives" of deepfakes. Furthermore, we show how definitional issues around what a deepfake is versus other forms of fake media feeds confusion around the novelty and impacts of deepfakes.**

***Index Terms*—Deepfakes, social implications of technology, media.**

I. Introduction

DEEPFAKES are a form of synthetic media that uses deep-learning technology to create fake images, video, and audio, which are then usually inserted into real footage. Early academic and media discourse around deepfakes has been characterized by conceptual vagueness. For example, the term "deepfake" has been used to refer to a range of different technologies and types of content [1]. Early deepfake literature is also characterized by a tendency from authors to predict a wide range of harms upon future proliferation of the technology. Due to the conceptual vagueness of early literature, it is often difficult to understand which of these predicted harms are new or unique to deepfake technology and content. Through reviewing the academic discourse about deepfakes across a wide range of disciplines, this paper aims to provide the reader with a stronger understanding of the many ways that experts see deepfakes impacting upon individuals, organizations, and society.

A. Critical Qualitative Work on Deepfake Narratives

Previous qualitative research into the harms of deepfakes has focused on understanding news media narratives and how they focus on highly fearful hypotheticals as opposed to practical existing cases. For example, an analysis carried out by Gosse and Burkell [2] of news media articles found that news media overly focuses on the hypothetical harms of deepfakes as the "ultimate threat to a democratic society". They also found that news media narratives also tend to speculate on how the average person may potentially be misled by future deepfakes, instead of discussing the existing harms of real deepfakes. The Gosse and Burkell study highlights the benefit of using qualitative approaches to critically analyze the emerging narratives around this technology. Yet little qualitative research has applied a similar critical lens specifically to academic narratives across disciplines. Most previous review work seeking to understand what deepfakes are and how they cause harm have often drawn from quantitative analysis or an attempt to synthesize a definition for deepfakes [3], [4].

B. Our Scoping Review Approach

A scoping review was chosen as the review methodology as it is appropriate for work which aims to provide a critical qualitative synthesis of academic articles across disciplines. While previous reviews have attempted to synthesize both the impacts and definitions of deepfakes, our research instead critically challenges a broad range of existing academic narratives across multiple disciplines using qualitative analysis. Our research also intends to challenge the definitional issues of

Received 22 April 2024; revised 6 September 2024 and 4 November 2024; accepted 4 November 2024. This work was supported in part by the Science Foundation Ireland under Grant 13/RC/2094_2, and in part by the European Regional Development Fund through the Southern and Eastern Regional Operational Programme to Lero—The Science Foundation Ireland Research Centre for Software (www.lero.ie). *(Corresponding author: John Twomey.)*

John Twomey, Didier Ching, Conor Linehan, and Gillian Murphy are with the School of Applied Psychology, University College Cork, Cork, T12 CY82 Ireland, and also with Lero—The Science Foundation Ireland Research Centre for Software, Limerick, V94 NYD3 Ireland (e-mail: John.Twomey@ucc.ie; 116378223@umail.ucc.ie; conor.linehan@ucc.ie; Gillian.murphy@ucc.ie).

Matthew Peter Aylett is with the Mathematics and Computer Science Department, Heriot Watt University, EH14 4AS Edinburgh, U.K., and also with CereProc Ltd., EH3 9DR Edinburgh, U.K. (e-mail: m.aylett@hw.ac.uk).

Michael Quayle is with the Centre for Social Issues Research and the Department of Psychology, University of Limerick, Limerick, V94 T9PX Ireland, and also with the Department of Psychology, School of Applied Human Sciences, University of KwaZulu-Natal, Pietermaritzburg 3201, South Africa (e-mail: mike.quayle@ul.ie).

This article has supplementary downloadable material available at https://doi.org/10.1109/TTS.2024.3493465, provided by the authors.

Digital Object Identifier 10.1109/TTS.2024.3493465

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

deepfake literature. Our review examines what researchers feel is novel about deepfake technology and provides implications as to what current deepfake researchers can learn from the technologies that have preceded deepfakes. We also seek to understand the "positive" and "negative" uses of deepfakes through a more critical multi-disciplinary lens.

II. Background

A. The Origins of Deepfake Technology and Definitional Confusion

The term deepfakes is a portmanteau of 'deep learning' (AI technology) and 'fake' which originates from a member of the Internet forum Reddit who went under the name "deepfakes", and the community that he founded, 'r/deepfakes' [5]. This individual took a previous Internet practice of photoshopping female celebrities into pornographic photographs ("r/celebfakes") and used modern artificial intelligence technologies to create fake pornographic videos. A typical face-swap deepfake video from the original online community, involves three processes; the extraction of facial data from a source and destination video; training the software to generate a face swap using encoder networks and refining this using generative adversarial networks; and finally, the insertion of the original face from the source video into a destination video [6].

While the term deepfake originally was used to describe generative AI enabled fake pornography videos, the term has since been used more broadly. It has been described as "*digital manipulation of sound, images, or video to impersonate someone or make it appear that a person did something—and to do so in a manner that is increasingly realistic, to the point that the unaided observer cannot detect the fake*" [7]. Indeed, while the general understanding of deepfakes includes audio-visual media modified using deep learning, there have been a number of other forms of media referred to by academics as deepfakes such as AI generated images, text and digital models [8]. Vasist and Krishnan [5], in a review of 68 deepfake-related journal articles and their definitions found that apart from the role of sophisticated generative AI techniques in their creation, there is little agreement as to what constitutes a deepfake and urged future review work to develop a precise conceptualization of deepfakes.

B. Further Review Work on Deepfake Definitions

Subsequent efforts to synthesize a definition in review work and theoretical work have just added to the confusion. In a meta-review by Altuncu and colleagues [3] of 12 deepfake papers, a much broader definition of the technology was proposed: deepfakes are highly realistic/ natural media of any kind generated by a sophisticated automated method (usually AI). This would include media such as generative AI images of geography and medical scans made by programs such as OpenAI's Dall-E 2 [9]. Whittaker and colleagues [4] synthesized a narrower definition for deepfakes from 80 research articles published between January 2017 and June 2021, defining deepfakes as "synthetic media generated using artificial intelligence and deep learning technology which produce realistic yet fake representations of people undertaking actions or saying words in the form of video, image, or audio content".

However, review work which focuses on synthesizing a definition does not critically reflect on the nature and impact of differing disciplinary perspectives on deepfakes. It is also important to question what the existing frameworks offer in the context of differing disciplinary perspectives. Because of these gaps in existing review work it is necessary to carry out our scoping review, which seeks to critically analyze how people are defining deepfakes across academic discipline and criticize the current definitional inconsistencies and confusion.

C. Defining Deepfakes in Relation to Other Technologies

The definitional issues around deepfake technology are inextricably linked to how it is understood relative to other technologies. In attempting to understand how previous research defines deepfakes, it is also necessary to understand how they are being compared to the technologies which preceded them. In Vasist and Krishnan's [5] review of deepfake definitions, they also put forward that deepfakes may not constitute a completely novel phenomenon, but rather constitute latest in a long history of media editing technologies. If so, it becomes vital to understand how existing responses to deepfake misinformation (false information) and disinformation (false information spread maliciously) can be broadened to include those of similar current and future technologies (we use the term misinformation throughout the paper as it is the broader term, except when referring to deepfake information spread maliciously).

D. How do Deepfakes Compare to Other Forms of Edited Media?

Deepfakes are often described as "photoshop on steroids" [10], and current attitudes around deepfakes are linked to the fears which have surrounded earlier image editing technologies. As well as relating to the definitional crisis of deepfakes, the novelty of the technology is important when considering current narratives around the impacts of deepfakes. Some critics have pointed out that most novel technologies have been associated with similar fearmongering at their onset [11] and early discourse tends to focus on fearful possibilities as opposed to any evidenced harms [12].

Many of the other harmful uses of deepfakes can also be seen as continuations of traditional forms of video and image editing. For example, the original deepfake community replicated the already-existing use of photoshop towards female celebrities and extended them to video, while reducing the technological/ artistic ability required. The question of what makes deepfakes different from previous media editing technologies such as photoshop has not been interrogated in detail in previous review work.

Our scoping review seeks to tackle the lack of existing literature and review work considering the novelty of deepfake technology. It is essential for research and policy on deepfakes to consider whether deepfakes are a novel phenomenon or a

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

symptom of larger problems with novel media editing technologies and society, especially with advances in generative AI. Otherwise, one risks designing solutions for a technology and not for the broader harms of media editing technology wholesale. If the harms of deepfakes are not novel then we need to focus on placing deepfakes within existing work and taxonomies of harms such as privacy, election interference and cybersecurity.

E. Narratives Around the Current and Future Uses of Deepfakes

The original deepfake community was banned on Reddit as a result of public outrage brought about by news coverage of the practice [13]. However, forums dedicated to the making and sharing of deepfakes still exist today [14]. As a result, both news media and academic organizations have expressed concerns about the potential harms of deepfake technology. In their seminal paper, Chesney and Citron [7] put forward their view of the key potentials for harm, and benefits, from deepfake technology (summarized in Table I). The non-consensual pornographic related harms of deepfakes are the most immediate and widespread, with pornographic content representing approximately 96% of deepfake videos currently online [15]. Outside of pornographic content, deepfake technologies have demonstrated a host of uses, both benign and malicious [16].

Fueled by the fear that deepfakes make more believable fake news than previous technologies, journalistic discourses around deepfakes often involve conjecture around the future use of deepfakes by "resource-rich bad actors" [17]. Notably, some commentators have characterized the typical academic, institutional, and journalistic responses to deepfake technology as "techno-paranoid narratives" [11] and as a "moral panic" [18]. This stems from the asymmetrical coverage of deepfakes in news media which focuses on the potential harms as opposed to the current uses of deepfakes [19]. Indeed, academic discourse around deepfakes tends to focus on the negative impacts of technology (note, for example, the disparity between positive and negative examples in Figure 1).

F. Critical Review Work on the Uses of Deepfake Technology

A systematic review using network analysis carried out by Gamage and colleagues [20] on the societal implications of deepfake technology highlighted the diversity of possible harms and threats to society in the literature. They found that security-related harms were dominant of these, with psychological, legal, and political harms also emerging. Similarly, research into the media coverage of the technology has also highlighted the dystopian and utopian narratives around the technology replicating previous hysterias over the emergence of technologies such as radio and television [11]. The media focus on deepfakes has also tended to center on the implications for knowledge and truth, despite pornographic deepfakes making up the majority of deepfake related content. By relegating deepfake pornography to the history of deepfakes, and by not prioritizing it as the most pertinent of deepfake harms, news coverage of deepfakes devalues the current harms of deepfake pornography [2].

TABLE I
THE KEY BENEFICIAL/ HARMFUL USES OF DEEPFAKE TECHNOLOGY ACCORDING TO CHESNEY AND CITRON [7]

Uses of Deepfakes		Explanation
Positive Uses of Deepfakes		
1., Education		Educational deepfake videos for the classroom
2. Art		Creative uses often involving the resurrecting of actors.
3. Autonomy		Digital self-expression – particularly for people suffering with paralysis or other restrictive health conditions
Negative Uses of Deepfakes		
1. Individual and organizational harms		
	Exploitation	Theft, abuse, and blackmail facilitated by deepfakes
	Sabotage	Sabotaging the reputation of others using defamatory deepfakes
2. Societal harms		
	Distortion of democratic discourse	Deepfake misinformation may harm political discourse
	Manipulation of elections	Use of deepfake videos to attack election candidates
	Eroding trust in institutions	Prevalence of deepfake videos will damage trust in institutions
	Exacerbating social divisions	Deepfake misinformation might encourage and stoke social issues, dividing and polarizing communities.
	Undermining public safety	Deepfakes could encourage public panic
	Undermining diplomacy	Deepfakes may inflame relations between countries
	Jeopardizing national security	Deepfakes could damage the reputations of militaries
	Undermining journalism	News organizations may be challenged to verify deepfake misinformation/ disinformation
	The liar's dividend	Deepfakes could make it easier for liars to deny truth

Many technical review papers for deepfakes have produced summaries of the positive and negative uses of deepfakes [1], [21], [22], [23], [24], generally focusing on the effects of deepfakes on entertainment, fake news, and the value of video evidence. However, there have been few reviews which have critically examined the contents of those arguments using a multi-disciplinary and qualitative approach. As well

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

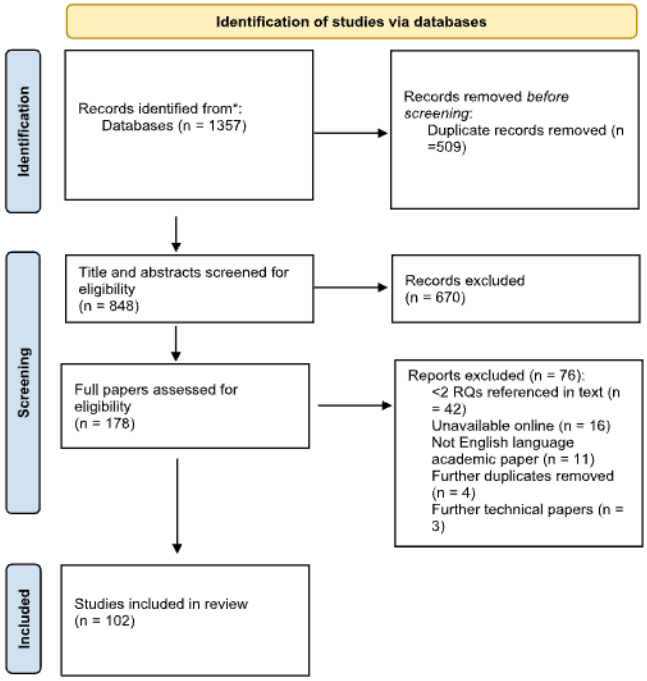


Fig. 1. Prisma flow diagram for search results.

as highlighting the definitional issues of deepfakes in their previously mentioned review, Vasist and Krishnan [5] also found what they call "divergent perspectives" across different academic disciplines. These were differences in opinion as whether or not deepfakes are harmful and the accuracy of overly fearful journalistic concerns. They found that some disciplines were split between discussing both the positive and negative impacts of deepfakes on their discipline (e.g., advertising) and they found some positive opinions around the technology in research fields which were at risk of the harms of deepfakes.

Vasist and Krishnan also called for and subsequently carried out a review of existing qualitative deepfake literature [25] and developed a conceptual framework of deepfake engagement using social shaping of technology theory. Their framework provides a comprehensive understanding of the socio-technical elements which underpin the fabrication of the videos as well as their dissemination/ sharing. However, their research is also predicated on deepfakes being classified into either positive or malicious deepfakes. There seems to be little work done in questioning this dichotomous classification. The current scoping review intends to build on existing review work which state the pros/ cons of deepfakes across multiple disciplines and to engage more critically with the impacts of deepfakes and to question what exactly is so beneficial/ harmful about them.

G. Aims and Research Questions

The emergence of deepfake technology has prompted responses from a wide range of academic disciplines including computer science/ information systems, humanities, social sciences (such as psychology and sociology) politics, and law [5]. There is little consistency between how these papers define deepfakes, in what ways they describe the technology as novel or different, and how they envisage the potential impacts of deepfakes. While the early response to deepfakes typically tended towards sensationalism, as with the introduction of each generation of new media [26], we are at a point where it is important to clarify whether there is any good reason why we should consider deepfakes as uniquely harmful or worrying. We have chosen to carry out a scoping review over a systematic review as we seek to study a large wealth of papers across disciplines with differing or no study design [27]. Our scoping review seeks to collate and review the specific concerns and predictions across disciplines to understand whether there is something qualitatively different about deepfakes and their harms from previous media editing technology.

The goal of this review is to answer three questions.

1) How do researchers define deepfakes?
2) In what ways do commentators describe deepfakes as being different to other fake media?
3) What current and future uses and implications of this technology are currently observed/ predicted?

III. Methods

We chose to use a scoping review for this research as it is useful in reviewing academic literature more broadly than other methodologies. Specifically, we focus not on the quality of specific empirical papers but instead to summarize and disseminate broader research findings, to more broadly examine the themes of academic literature and to identify where existing research, reviews and theory are falling short [27]. The search was carried out on the 11th of January 2023 across seven databases, limited to English language journal articles published between the 1st of January 2018 and the date of data extraction.

The search terms used were (deepfake* or deep-fake*) within the categories of title, abstract or keywords. These search terms were kept intentionally broad to include the largest possible amount of deepfake articles. Similarly, the databases were chosen to gain a wide, multi-disciplinary dataset across the relevant research domains. The databases chosen were Web of Science. SCOPUS, Hein Online Legal Databases and three EBSCO academic databases (Academic Search Complete, Business Source Complete, APA Psychinfo). The protocol of this scoping review was pre-registered on the Open Science Framework (OSF) website (https://osf.io/xc7mk) and the findings are reported using the PRISMA-SCR checklist.

A. Screening Process

To be included in the scoping reviews, papers had to come from a relevant academic discipline such as Information Systems, Computer Science, Law, Sociology, Philosophy, Media, or Psychology. We excluded approximately 261 papers that were solely concerned with deepfake detection/ mitigation or production, as this has already been carried out comprehensively by previous deepfake review papers (e.g., [4]). We did

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

not search for grey literature on this topic due to our review's specific focus on academic narratives. Papers published before 2017 were also excluded as part of the search as they predate the emergence of deepfakes.

At this stage, most of the duplicates were identified based on title and removed. The Rayyan open-source software for systematic reviews was used to store the data and two reviewers used the software to screen the papers. Papers were first excluded where the abstracts and/ or title did not mention at least one of the three extracted topics (definitions of deepfakes, comparisons with other technologies, and potential effects). Papers were then excluded if they did not contain two or more of the three variables to be extracted in their full text.

B. Data Extraction

There were three qualitative variables extracted for the scoping review: definitions of deepfakes, novelty of deepfakes/ comparisons with other technologies, and potential effects of deepfakes. To be included in our dataset as a definition, papers had to mention deepfakes as a form of media/ content or a technology which produces media/ content which reflects both the original 2017 deepfake content and definitions for the technology which have been established by previous reviews [3], [4]. For text to be included in the analysis of novelty, the paper had to mention deepfakes as novel, position the technology in comparison with other technologies/ forms of media, or talk about other technologies in the text in the context of deepfakes. Finally, impacts had to be concerned with what deepfakes have, can, should or will do. The final dataset consists of 102 papers and is 181,659 words long (a full list of papers included can be found on the project's OSF page https://osf.io/bxnds).

C. Data Analysis

For our analysis we broadly follow the steps outlined by Braun and Clarke [28] for conducting and reporting thematic analysis. We coded each extracted textual unit or paragraph, as opposed to coding the text line by line due to the size of the dataset, and used these codes to generate our themes. This method was derived from how other methodologies (such as applied thematic analysis [29]) describe coding larger textual blocks. While time-intensive, this method was invaluable as a way of qualitatively synthesizing a broad range of articles. As for similarly intensive qualitative review methodologies [30], this meant that we were not able to include more recent work, though we did keep abreast of relevant review work published during the analysis. We then related the resulting themes back to the data corpus as well as the existing theoretical frameworks, examining differences and common factors. While entire paragraphs were coded, when reporting the data in the analysis we went back over the codes to gather illustrative examples of the codes for the reader, similarly we also gathered word counts from the different themes as a way to use corpus data to further illustrate for the reader how often certain concepts emerged.

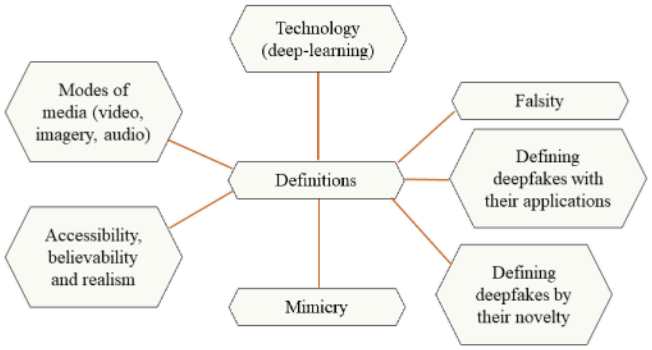


Fig. 2. Themes drawn from deepfake definitions.

IV. How Are Deepfakes Defined in the Literature

A. Deepfake Definitions Across the Dataset

We found 86 separate definitions of deepfakes in our dataset. Of those definitions, some were cited repeatedly (Supplementary Materials). Specifically, two definitions were cited over five times in the dataset:

1) Westerlund [24]: "*A combination of "deep learning" and "fake", deepfakes are hyper-realistic videos digitally manipulated to depict people saying and doing things that never actually happened*"
2) Chesney and Citron [7]: "*It [Deep-fake technology] leverages machine-learning algorithms to insert faces and voices into video and audio recordings of actual people and enables the creation of realistic impersonations out of digital whole cloth. The end result is realistic-looking video or audio making it appear that someone said or did something*"

One essential characteristic that both these definitions have in common is in how deepfakes can be used to represent someone doing or saying something which never actually happened.

The capacity of deepfakes to mimic and appropriate identity was an integral part of how the impacts of the technology related to the definition in our analysis. Similarly, there was an assumption of untruthfulness around deepfake technology, and the technology was seen as manipulating media for the purposes of falsification and fabrication. Deepfake technology was often defined by its ability to blend the fake into the real, through interpolation, as in face-swap videos.

B. Common Factors of Definitions

Our analysis identified a set of common factors across deepfake definitions (Fig. 2). As expected, we found that deepfakes were generally defined by their use of deep-learning technology, a subset of machine learning and artificial intelligence (e.g., *"a technique for human image synthesis that employs artificial intelligence and deep learning technology to produce false yet deceptively realistic images and videos"*), though definitions shared little detail on the technical process underlying the technology [31]. Some papers limited the modality of deepfakes to videos, in line with the original deepfakes published in 2017: "*At the most basic level, a deepfake is created by taking an input source of videos or*

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

other images of an individual and outputting a manipulated video of that individual's face" [32]. However, deepfakes were more broadly viewed as consisting of video, image or audio and often multi-modal, involving audio and video.

The broader definition was justified as being necessary in light of developments in deepfake technology to capture "*the full range of hyper-realistic digital falsification of images, video, and audio*" [7]. There were references to deepfake text, referring to AI generated text such as that created by large language models (e.g., [33]), but this was not common. Considerations of the modality of deepfakes ended up drawing from comparisons of the technology to previous media editing technologies, connecting these two research questions, Deepfakes were often presented as one of a host of synthetic media technologies (e.g., "*Deepfakes are a new form of synthetic media that broke upon the world in 2017*" [15]).

C. Points of Confusion Around Definitions

The need for greater clarity and precision in deepfake definitions became apparent when analyzing the impacts of the technology. The dataset included a mix of real-world examples of the impacts of deepfakes and projections of potential future impacts. Confusingly, some examples given in the dataset regarding the real-world impacts of deepfakes referred to traditionally edited media instead of, or along with, incidences of deepfakes. "*In 2017, researchers successfully created deepfake stills of former President Barack Obama saying things that he never actually said but were voiced by actor, Jordan Peele.* (deepfake). *The same technology can create false videos inciting plans to carry out political assassinations, seemingly private conversations featuring elected leaders, or Speaker Nancy Pelosi slurring her speech* (not a deepfake, just slowed down video)" [34]. Using examples of the harms of photoshop and rudimentary video editing technologies occasionally made sense, as these examples were used to illustrate the current harms that deepfakes evolve upon, however in some cases the traditional examples were used interchangeably with deepfake examples. Worryingly, some papers used a combination of real examples of deepfakes and traditional CGI/ photo editing without clarifying the differences between the technologies. Switching between examples of deepfakes and other false media has the effect of exacerbating the definitional issues of deepfakes and highlights the need to clarify what exactly defines deepfakes in comparison with other technologies.

V. The Novelty of Deepfakes

A. Accessibility and Believability: The Two "Mores" of Deepfakes

Two concepts were consistently observed in our data set, where authors discussed the novel nature of deepfake technology. Many viewed deepfakes as being alike to previous technologies but "more" in a certain way, particularly around the technology's accessibility (e.g., "*In the last four years, deepfakes have developed from an academic project to a rapidly evolving and accessible form of technology*" [35]) and believability (e.g., "*Concerns that the audiovisual nature of deepfakes makes them inherently more believable than previous fake news forms leads to worries over the impending weaponization of deepfakes by resource-rich 'bad actors*"' [17].

In discussing the accessibility of deepfake technology some authors emphasized how comparatively easier it is to produce deepfake videos than equivalent methods of fake-video production. These narratives often revolved around the cost-free availability of open source deepfake technology to the tech-savvy Internet users and the availability of consumer deepfake applications. "*While Hollywood has for many years been able to generate synthetic content, it has been very cost, time and resource intensive. However, the technology associated with deepfakes requires a lower threshold of skills and time, and the resources used to create such materials are readily available and accessible to individuals globally*" [36]. The focus on the average person having access to deepfake technology highlights fears around the democratization of harmful AI technologies. While the general view of deepfakes is that they are comparatively easier to make than other forms of edited videos, it must also be noted that realistic deepfake videos still require specialized technical skills.

B. Broader Impacts of Accessibility and Believability

The realism and believability of the technology exacerbates both the social harms of the technology (through misinformation) and the potential for reputational damage (as most often reflected in the harms of deepfake pornography). These characteristics were seen as integral to the technology's unique capacity for harm. The harms were frequently described either as novel or, less often, as an evolution of the similar harms of previous technologies. Deepfakes harms are novel in the way they create realistic faked faces with less effort than previous technologies: "*most everyday manipulation or remixing fell short at generating moving images and sound that could convince a relatively discerning viewer that it was recorded footage*" [18]. Deepfakes have the capacity to manipulate both audio and visual data and as a result was seen as "better" than previous techniques such as CGI and Photoshop which are visual editing technologies. However, similar technologies such as photoshop have been commonly used for defamatory content, pornography, and misinformation for two decades prior, so deepfakes mostly serve as "*the latest iteration of a longstanding practice*" [37].

The positive uses of deepfake were similarly discussed as being facilitated by the accessibility of the technology. While talking about the entertainment usage of digital recreations and recasting, papers were quick to point out how time-consuming and poor-quality previous manual attempts have been. "*The deepfake technology available to consumers superficially resembles Hollywood special effects, the sort where dead actors are revived anew onscreen. But these professional techniques are extremely expensive and time consuming. What makes deepfakes remarkable is that they provide an approximation of the same effect, much more cheaply and quickly*" [38]. There is also a temporal element to the availability and realism of the technology, as the years since early deepfake research have seen an increasing amount

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

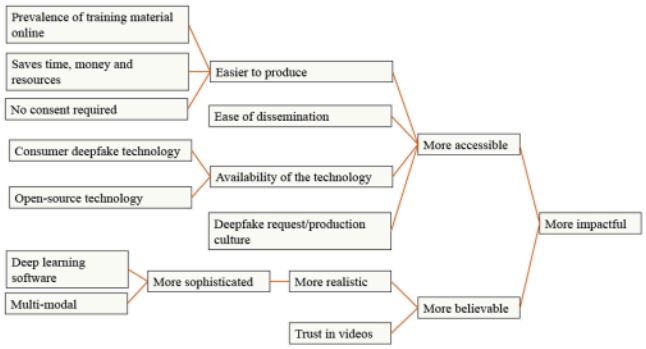


Fig. 3. Themes and subthemes relating to the accessibility and believability of deepfakes.

of cheap or free deepfake applications. Few papers considered that being more accessible and believable than previous video editing technologies does not mean that deepfakes themselves are very realistic or very accessible. Notably no papers presented any evidence that deepfake images (not videos) are more realistic than traditionally manipulated images.

C. Comparisons With Other Technologies

There were three different ways in which papers attempted to relate deepfakes to other media technologies; the deepfake – shallowfake dichotomy, the cheapfake-deepfake spectrum, and as part of a taxonomy of synthetic media (Table II). The deepfake-shallowfake dichotomy [39] is a means of classifying media editing based on whether deep-learning technology is involved. "*Deepfakes should be distinguished from shallowfakes, which are also manipulated media, but manipulated through human intervention rather than artificial intelligence*" [40]. The deepfake-cheapfake spectrum [41] organizes media technology on a spectrum based on the sophistication of the technology used to create the fake media. "*At the other end of the deepfake spectrum, less sophisticated actors can create "cheap fakes," or lower-quality deepfakes, using consumer-grade software or simple video-editing techniques*" [42]. The terms "cheapfake" and "shallowfake" are notable, as many forms of traditional editing can be quite complex and expensive and is neither cheap nor shallow.

It is also interesting to note how these two categorizations interact with the ideas of deepfakes being more believable and accessible. The view of previous technologies as "shallowfakes" lines up with the increased sophistication of deepfake technology (and hence realism and thus believability). In contrast describing previous technology as cheap is inconsistent with the view of deepfake technology being a more accessible and cheaper way to manipulate videos.

These definitions/ comparisons serve an important purpose; delineating deepfakes from the spectrum of prior media editing technologies and thus establishing its novelty. However, these two terms often muddy the waters of deepfake definitions, especially as they refer to forms of image manipulation that have been defined for decades prior. Finally, deepfakes were often characterized in relation to other so-called "synthetic media" [43] (as opposed to recorded media). Deepfakes are a means of producing a type of audio-visual synthetic media using deep learning technologies.

TABLE II
APPROACHES TO ORGANIZING DEEPFAKES IN COMPARISON TO OTHER MEDIA

Typology	Year	Explanation
Deepfake – shallowfake dichotomy [39]	2019	Deepfakes are made with sophisticated AI, shallowfakes are not. Shallowfakes often consist of videos which have been relabelled and re-uploaded with a different context. If shallowfakes are manipulated, it usually involves simple edits, or a new audio track.
Cheapfake-deepfake spectrum [41]	2019	The authors propose a spectrum charting audio-visual manipulation techniques based on their technical sophistication, barriers to entry, and techniques. In their most complex and expensive form, deepfakes include completely virtual and realistic performances. "Cheaper" techniques include the typical lip-synching and face-swapping deepfakes. At the other end of the spectrum, "cheapfakes" at their least complex consist of recontextualized videos. At their most complex they involve rotoscoped face-swapping and sped-up/ slowed down videos.
Deepfakes as part of a taxonomy of synthetic media [43], [44]	2021, 2022	Synthetic media can be partially or totally synthetic. Partially synthetic media involves the modification of archival media with new media created using a generative component. Totally synthetic audio-visual media is not produced by modifying pre-existing archival media but instead involves generating new audio or visual media. Deepfakes originally referred to face-swapping videos, but now refer to a wide range of deep-learning based audio-visual manipulations which can be partially or totally synthetic. Limiting deepfakes to these taxonomical distinctions misses the effect deepfake technology may have across all synthetic media

VI. THE EXISTING AND FUTURE USES OF DEEPFAKES

A. Potential Uses of Deepfakes

We identified a number of potential harms and uses raised by experts across the disciplines in our dataset. Through our analysis, we classified these into three chief types of harms:

1) Personal harms of defamatory pornographic deepfakes, particularly harassment and objectification.
2) Societal harms of defamatory political deepfakes, enacted through social manipulation.
3) The epistemic harms of deepfakes to video (harms to truth and evidence).

The dominance of these three types of related harms is apparent in the corpus (Table III), showing a focus on terms referring to the pornographic deepfakes (e.g., pornography, sexual, women), political deepfakes (e.g., political, election, disinformation) and the epistemic harms of deepfakes (e.g., trust, evidence, epistemic). In contrast, a small minority of papers suggested that there may also be political and

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
WORDCOUNTS OF DEEPFAKE-RELATED WORDS IN THE IMPACTS SECTION OF OUR DATASET

Deepfakes	1987	Epistemic	107
Deepfake	1345	Impact	106
Video	877	Victims	104
Technology	750	Voice	103
Videos	576	Data	103
Media	487	Users	100
Fake	483	Recordings	98
Public	353	Uses	98
Social	346	Intelligence	98
Content	342	Security	97
Political	307	Problem	97
Deep	303	Nonconsensual	97
People	288	Footage	96
News	265	Spread	96
Potential	252	Citron	96
Information	251	Individual	95
Fakes	242	Internet	95
Evidence	235	Possible	94
Online	216	Effects	93
Pornography	212	Legal	92
Digital	208	Threats	91
Real	195	Speech	90
Harm	193	concerns	90
Individuals	190	Chesney	90
Images	183	Fact	90
Threat	163	Journalists	90
Person	158	Abuse	89
Pornographic	157	Virtual	89
Face	156	Reality	87
Trust	154	Future	87
Porn	147	Misinformation	87
False	147	Manipulated	85
Audio	144	Artificial	85
Sexual	140	Revenge	85
Women	130	Sex	84
Image	129	Elections	83
Technologies	127	Undermine	83
Disinformation	126	Truth	82
Ai	125	Malicious	82
Actors	119	Manipulation	81
Society	116	Realistic	81
Election	115	Different	80
Time	115	Opportunities	79
Harms	114	Negative	79

societal benefits. For example, improved skepticism in light of misinformation deepfakes and more trust in established and reputable politicians/ media [32]. Some papers expressed that deepfakes would empower political activists through satire [45] and noted the potential use of deepfakes in advocacy videos by charities [15]. However, these papers were in the minority. While some papers in the dataset discussed the positive uses of deepfakes, this was not common, as reflected in the corpus data. This illustrates either (or both) the dominance of dystopian narratives around deepfake technology, or the lack of many real-world examples of the potentially positive uses unique to the technology.

Narratives around the current and potential use of deepfakes were generally characterized by warnings that the technology has caused or will cause more harm than good. "*One might well leave this discussion with a lingering sense that the potential for more damage is enormous and already rather concrete, whereas the benefits are much more uncertain*" [15]. Other papers proposed that reasonable responses to deepfake technology involved balancing the positive and negatives. A minority of papers expressed that the technology was not as bad as people believed "*The reality is that as the technology has progressed, a different picture of how deepfakes will be used has emerged. They may not be such a problem at the highest levels of security and defence. Instead, the problems that deepfakes pose lie at the lowest levels*" [46].

There were also papers which expressed the opinion that political and news media concerns around deepfakes consisted of a moral panic and commented on the extreme harms of deepfakes as being a mostly hypothetical and unevidenced. *"Past experience suggests that "the panic around" deepfakes will turn out to be a "false alarm," as one commentator recently concluded. As far as the courts are concerned, I predict that this panic will prove largely unfounded. The nation's courts are robust institutions that have shown themselves capable of handling each new variant of the age-old problem of fakery*" [47].

The victims of deepfake harms in the text were split between individuals and organizations/ institutions. Public figures emerged as a significant target of deepfakes. In line with the general focus on political and pornographic harms of the technology, as well as gendered narratives around the technology, these public figures tended to be male politicians and female celebrities (though there were some exceptions to both).

B. Personal Harms of Defamatory Deepfakes

Defamatory deepfakes were associated with a series of personal harms. They undermine the consent of the parties in the videos and are then used to harass and defame the targeted individuals. This was clearly exemplified by the use of deepfakes to generate pornographic videos of female celebrities. In many ways, this harm was seen as an evolution of previous forms of image-based sexual abuse, now being able to target anybody [2]. It was also seen as part of a larger pattern of technology being used to invade female celebrities' privacy and to silence activists [48]. The use of pornographic deepfakes to intimidate, harass, objectify, and humiliate targets was identified commonly as one of the biggest threats of deepfakes.

These defamatory harms of deepfake technology were associated with significant adverse personal responses. These include 1) the loss of control over how one's identity is visually represented as well as 2) the disempowering potential of the technology as part of the social and individual consequences of deepfakes. "*Much like the painful accuracy of non-consensually disclosed pornography, the misrepresentation of deepfake pornography impacts one's ability to control their sexual identity*" [49]. In many papers that were themselves unrelated to pornography, it was still referred to as part of the origin of deepfake technology. "*While deepfakes originated in the pornography industry, there are growing concerns about how they could interfere with politics and elections and threaten individual privacy*" [13]. The reputational damage of deepfakes can potentially cause severe financial consequences and career damage if the fake content reaches employers.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Defamatory deepfakes can also harm an individual's self-image, forcing them to doubt and question themselves in the face of convincing visual evidence.

C. Socially Manipulation Using Defamatory Deepfakes

The other prevalent harm involved the use of deepfakes in social manipulation, particularly the use of defamatory political deepfakes to spread lies and misinformation. Many papers argued that deepfakes give people an increasing number of tools for imitating public figures and tampering with existing videos. These concerns often focused on the harms to governments, online platforms, journalism, and news media and business organizations. "*Deepfakes can be used to spread fake news, influence elections, introduce highly realistic fake evidence in courts and make fake porno movies. Each of these applications potentially has a big impact on society, social relationships, democracy and the rule of law*" [45]. Many papers also considered the cognitive effects of socially manipulative deepfakes, discussing the use of deepfakes in misinformation and how they may encourage confusion, reinforce previous distrust and conspiratorial thinking, perpetuate biases, influence individuals' ideologies, and induce false beliefs/ memories [16], [32], [50].

Other concerns focused on hypothetical scenarios of companies' CEOs being defamed, leading to their market value plunging [51]. There was also a concern around existing financial crimes such as blackmail, identity theft and fraud which could be augmented using deepfake technology. AI generated fraud phone calls emerged as a distinct harm of deepfakes in this field due to a small number of real-life examples where fake video calls were purportedly used to extort millions from companies [52]. Most papers, particularly law and politics papers, focused on institutional harms and particularly the destabilizing effect of deepfakes on politics, international relations and democracy. Deepfakes may spread disinformation through its capacity for novel forgery and identity theft, augmenting text-based disinformation with fake profile pictures and deepfake videos. Or, with the advancement of real-time deepfake technology, there is a possibility of deepfake video calls being used to spread disinformation and commit fraud [33].

D. Socially Manipulative Deepfakes in Conflict

There were fears that socially manipulative deepfakes would be utilized in future conflicts as a weapon for sabotage or as propaganda. Some papers also expressed a fear that deepfakes would be used to garner supports for terrorists or to be used for terrorism in the form of cyberattacks. The most extreme securitization fear around the technology involved the hypothetical use of deepfakes to start a war through deepfake misinformation. "*In our present age of misinformation, society will soon have to deal with deepfakes that can threaten national security. Consider a deepfake of President Trump announcing impending nuclear missile attack on North Korea* "[53]. Often these worries were characterized by fears of "bad actors", usually political agents or in many cases so called "foreign actors".

One paper alluded to the military benefits of deepfake technology in military engagements as a means of creating propaganda and fake news [32]: "*In the military and national security context, the same Al technology used to create deepfakes could be repurposed to assist intelligence agencies in surveilling hostile areas by detecting human threats with facial recognition software, thereby countering guerilla warfare and insurgency. The U.S. military could also use Al technology to achieve a strategic advantage over enemies with lesser capabilities*". Similarly, some papers viewed Deepfakes as giving people a new means of political expression through deepfake political satire. This highlights the perspectival element of deepfake harms and benefits where the positives of deepfakes are often seen as negatives from different points of view.

E. Epistemic Harms of Deepfake Videos

A third, broader harm in the dataset involved the epistemic harms of deepfakes. Deepfakes were seen as a threat to the general public's trust in video evidence. Some papers expressed that video records are anomalous in human history as previous forms of novel media have all been shortly followed by the ability to be manipulated and forged. These epistemic concerns related to the broader societal and philosophical impact of deepfakes, particularly how deepfakes threaten truth by undermining faith in videos. "*While videos and audio remain powerful pieces of evidence, their authenticity will be hard to gauge in the era of deepfakes*" [54]. In turn the damaged trust in videos undermines trust in documentary records, evidence, institutions, and governments [42].

Similarly, deepfakes were seen to challenge the freedom of activists and protestors, especially through undermining low-quality videos and thus challenging the validity of civilian journalism. "*As deepfakes become easier to generate, and more common, media outlets may be well-advised not to take for granted the accuracy of video footage from unknown sources. In this way, some video footage will be unsuitable as evidence*" [55].

F. Epistemic Harms of Deepfake Narratives

The epistemic harm of deepfakes has similar implications to the political harms of deepfakes, as they both draw from the undermining of trust in politicians and institutions/ organizations. However, it should be noted that much of the epistemic harms of deepfakes drew from the harms of the narratives around deepfakes as opposed to the actual use of deepfakes. Legal papers tended to focus on this harm, expressing worry about the damage of this lack of trust to courtrooms, by devaluing video evidence and forcing legal teams to have to use expensive computer forensics to decide whether a technology is a deepfake or not [56].

The so-called 'liar's dividend' emerged from legal discourse as a unique harm of deepfakes [7]. This is the effect of epistemic doubt around videos as providing liars a new form of deniability. With the epistemic damage of deepfakes, future politicians may have to regularly deny deepfake misinformation. Many papers highlighted the role of deepfakes in facilitating lies and liars. However, there were fewer empirical

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

papers which explored the extent to which deepfakes may exacerbate or cause epistemic damage. While the harm of deepfakes to truth and trust fuel academic narratives around the negative aspects of deepfakes, there is little real-world evidence given for these harms.

G. Artistic, Educational and Self-Expressive Deepfakes

The positive uses of deepfakes, expressed across disciplines, focused on creative and artistic uses or educational uses. Deepfakes allow the creation of media that depicts things that did not really occur, or were at least not filmed, and while this has a significant impact on truth and misinformation, it will also increase what people are able to creatively express. The use of deepfakes allows users more creative expression in how they can present humorous content, as well as resurrecting dead celebrities, gender swapping and exploring alternate realities. For example, a deepfake audio of U.S. president John F. Kennedy of the speech he was planning to make before his assassination or a deepfake of Nixon presenting a speech he would have made had the Apollo 11 astronauts died [51], [57]. Deepfakes were seen as having the potential to influence entertainment and art, either through the use of deepfake technology in online content and memes or to improve fictional films and TV.

In papers which focused on media/ entertainment, Deepfakes were seen as encouraging new practices in Hollywood [58], where films and media could be changed significantly by using deepfakes to recast actors and to digitally resurrect deceased actors [15]. "*Created by a number of online video artists, these convincing 'mash-ups' playfully rewrite film history by retrofitting canonical cinema with new star performers*" [58]. There were still harms noted in these cases, particularly concerning the ethics of appropriating the identity of dead actors to feature them in movies without their consent.

This was particularly relevant for deepfaked celebrity advertisement, where there was also the potential for false endorsements and the disrespect of using dead celebrities to market products. Further criticisms highlighted how the positive entertainment uses of deepfake tended to target less marginalized individuals versus the harms. "*The many creative involvements with Cage's star image and his emergence as a regular target for Deepfakes technology has further surfaced how the cyberstars of such 'take two' cinephilic activities are typically cisgender white males"* [58].

H. Economic Benefits of Educational Deepfakes

The positive economic effects of deepfakes in tourism/ marketing literature drew from the belief that educational historic deepfakes were one of the main positive uses of deepfakes. The widespread incorporation of these were felt to potentially increase visitor engagement and thus tourism revenue for nations, towns, and businesses [59]. Deepfakes were generally seen as having educational benefits including the use of deepfakes in museums or schools to make educational videos more immersive. Some papers problematized this however, stating "*Some perceive the idea that Napoleon would give a history lesson in secondary schools a nice prospect, others find it problematic because children would be confronted at an early age with blurred lines between fact and fiction*" [45]. No paper questioned the extent to which the educational role of deepfakes could be similarly accomplished with impersonators in costumes and there was no empirical evidence given for the education benefits of deepfakes.

Positive educational narratives also focused on the use of deepfakes to create easy lip-synched multilingual videos [60]. This has the potential to make video more accessible, reducing the cost of having to remake videos in different languages and providing more support for lip-reading than traditionally dubbed media. One positive element of the ability of deepfakes to erase the identity of the source was its use in anonymization, particularly for documentary movies. "*Deepfakes can be used to empower the marginalized to allow them to share their stories using authentic emotion without fear of identification. Welcome to Chechnya, a documentary which explores the persecution of LGBTQ people in Chechnya, uses deepfakes to anonymize the subjects of the documentary by transplanting the faces of New York-based activists onto the faces of the Chechen documentary subjects*" [61]. In this way, deepfakes could protect anonymous whistle-blowers from harassment and reprisal. However, it has not been established if this form of anonymization improves over existing practices of anonymization besides subjectively making content more "expressive".

I. Deepfakes as a "Double-Edged Sword"

The positive uses of deepfakes were seen by some papers (mostly legal papers) as balancing out the existing harms of the technology. Narratives around balancing the impacts of the technology described deepfakes as a two-sided coin [37] or a double-edged sword [62] as a way of equivocating between the positives and negatives of the technology. This falsely implies parity between the positive and negative uses of the technology, which does not reflect the incidence rates of both [63]. While discussion of non-consensual pornographic related harms was prevalent in the dataset, the academic discussion does not reflect the dominance of pornographic deepfakes online. While many papers suggested correctly that pornographic deepfakes are the current dominant use of the technology, some papers played down their harms in comparison to the more hypothetical and apocalyptical political impacts. "*But what happens when deepfakes stop being the fare of redditors with free time and a fixation on Game of Thrones actresses and become a weapon for state actors interested in destabilization of governments like the United States*?" [64].

VII. Discussion

Our research builds on and extends the work of previous systematic and literature reviews across our three research questions. The scoping review methodology [27] allows us to collate a broad range of research and theory to give us insight into the existing narratives around deepfake technology. Regarding existing work on deepfake definitions, our

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

research finds similar themes to the synthesized definition established by Whittakker and colleagues [4]. However, in place of Whittaker and colleagues' theme of "saying" (which they defined as the generation of "*spoken words which are seemingly verbally communicated from the person depicted in the deepfake*"), our critical analysis found the importance of identity and mimicry in how deepfakes are defined (instead of the generation of spoken words). This is more consistent with the popular established definitions we identified which describe the deepfakes as mimicking people "*saying*" or doing things which they never actually *said* or *did* [7], [25], and even implicitly within the definition they provide themselves.

This work also advances the work of existing categorizations of deepfakes by identifying problems with the use of existing terms like "shallowfake" and "cheapfake". These terms seemed to add unnecessary confusion as to what constitutes a deepfake. In contrast, we emphasize Millière's [44] philosophical paper which categorizes deepfake technology alongside other synthetic media (defining deepfakes as being audiovisual media modified through AI technology and not purely AI generated images or text). This categorization is not only fundamental for researchers outside of philosophy to understand how deepfakes are different to other media, but it is also a useful guideline as to what constitutes a deepfakes.

Further theoretical contributions of the work include identifying numerous unsubstantiated claims in the literature regarding the novelty, accessibility, believability, and harms/ benefits of deepfakes. We also discuss implications for how researchers approach deepfakes, outlining potentials for approaches which consider the role of identity and consent. The review also encourages the potential of a more holistic approach to deepfakes. Our review problematizes the commonly assumed dichotomy of positive and negative deepfakes which is implicit in most review work on the topic (e.g., [1], [21], [22], [23], [24]). Instead our scoping review takes a critical approach to existing theoretical and review taxonomies on the harms and benefits of deepfakes and highlight how the harms and benefits are often two sides of the same coin, and the positive uses of deepfakes may have unintended harms.

A. Deepfakes are Inconsistently Defined

The deepfake literature is mired in the definitional issues which plague news media narratives around deepfake technology. The definitions for deepfake technology were inconsistent throughout the literature, with 16 papers not containing a definition for deepfakes at all. Traditional methods of generating synthetic media such as CGI and video editing were reframed as either "cheapfakes" [41] or "shallowfakes" [39]. This review highlights how this confusion leads non-deepfakes being used to describe the potential harms of the technology. While this may have been important at the very advent of deepfake technology, there are now a multitude of examples of real cases of deepfake harms to cite. Our work encourages future work which seeks to reference other types of fake videos/ images should use terms that are already associated with traditional video and image editing (CGI, photoshopped, photogrammetry, and fake videos (in cases where it is clear that a video is fake, but the type of fakery is unclear)).

B. The Importance of Mimicked Identity to Deepfakes

Our analysis highlights how the importance of mimicry in deepfake definition is understated in previous systematic and scoping reviews on deepfake definitions. Across disciplines, the primary harms of the technology draw from the technologies ability to defame through false visual representations of individuals. While most of the legal literature concerned with deepfakes has touched upon this damage, it is rarely recognized that most deepfakes inherently challenge representations of identity. In particular, pornographic deepfakes were seen as causing defamation and irreparable harm to individuals self-image and a loss of control over their identity: "*the misrepresentation of deepfake pornography impacts one's ability to control their sexual identity*" [49]. The importance of identities' role in deepfake impacts has been understated by past research and it may be important to consider many harms related to deepfakes as drawing from the capacity of the technology to control or appropriate the identity of its target and to erase or disguise the identity of its source. Our review identifies the need for novel approaches to deepfakes which consider the role of manipulated identity as fundamental to the personal harms of deepfakes.

Our review highlights a research gap in the study of identity and deepfakes which has broader implications for work focusing on challenging deepfake perpetration. Not only do deepfake communities trade in identities, but the specific identities of deepfake targets reflect their values. The defamatory harms in our analysis were viewed as currently and potentially targeting women as well as racial, gender, and sexual minorities and a means of silencing targeted activists. Existing work has highlighted the need to consider how deepfakes further social injustice and the role of intersectional approaches in responding to the damage of deepfakes [19]. Future research from the social and political sciences into the personal and privacy harms of deepfakes needs to privilege the concept of identity in their approach to deepfakes, focusing on how the technology serves as a vehicle to appropriate and disguise visual representations of marginalized groups. Similarly, policymakers and legal scholars who seek to restrict the harms of technology must recognize deepfakes as part of a larger pattern of technologies which facilitate the theft and appropriation of identity.

C. Deepfakes Are a New Way of Causing Old Problems

Our analysis questions what exactly academic researchers have been classifying as novel about deepfakes. We found that deepfake impacts are an evolution of the harms of previous technologies, and reflective of the online culture in which deepfakes emerged. The pornographic harms of deepfakes are an evolution of a previous practice using photoshop, the use of deepfakes in mis/ disinformation evokes the use of photo editing in fake news. This once again ties into the accessibility of the technology, as while photoshop was not the first means

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

of image manipulation, it made image-editing more accessible to the average person [65].

It is interesting to consider the prevalence of these fears in the political/ legal literature around deepfakes in contrast to lack of academic literature on traditional photo-editing technologies and their harms in misinformation [66]. If the foundational harms of deepfake are exemplified by existing examples of non-deepfake content (e.g., slowed down videos being used to defame someone as drunk), researchers must consider whether it is worthy in studying deepfakes or simply the impacts or harms themselves. Unless deepfakes can be empirically proven to be significantly more convincing or uniquely harmful, studying the impact of all fake media on defamatory content will have broader impact, particularly in light of future advancements in generative AI which may have an impact in its own right on society as well as potentially increasing the accessibility of deepfake technology [67].

D. The Novel Epistemic Harms of Deepfakes

We call for further research, (particularly from social/ political sciences) questioning how the prevalence and accessibility of these technologies will affect online spaces, whether it be through the increase in AI-generated "spam" content online, or the increase in personalized defamatory deepfake content. In particular, the epistemic harms of deepfakes may be most affected by the prevalence and accessibility of deepfake videos. An increase in the accessibility of deepfake video production would lead to an increase in the prevalence of deepfake videos online and thus lead to increased damage to the value of real media in our information environment.

In this way, we propose that deepfakes should be considered a novel method for "epistemic pollution", false information which can worsen the quality of our information environment [68]. If the number of fake videos increases online due to deepfakes and other kinds of false videos, our current online information environments, online forums, social media etc. could become significantly less usable as users seek to navigate increasing amounts of manipulated information. Of particular concern to legal scholars is the way deepfake footage damages the epistemic value of video evidence. Particularly how deepfakes allow individuals to falsify or incorrectly challenge CCTV evidence [69]. Mitigating the amount and damage of generative AI misinformation online is an essential goal for policymakers, media companies and researchers in political science and across disciplines.

E. Deepfake Accessibility and Believability: Cause for Concern, Strategies for Solution and Need for Empirical Evidence

Our scoping review highlights how the accessibility and believability of deepfakes is central to both the impacts and novelty of the technology. Deepfake technology, while more complex than other forms of media editing, is often hidden behind user friendly tools [19]. Existing work has studied how the accessibility of deepfakes is a byproduct of the open-source movement. The lack of oversight of open-source technology fuels negative uses of the technology [70]. We propose that the extent to which deepfakes are both believable and accessible are constantly in flux, depending on people's awareness of the technology and the prevalence, not only of deepfake technology but the expert/ hobbyist communities dedicated to their production. Our work calls for increasing empirical and theoretical work which seeks to understand the role of accessibility and believability in the potential uses of deepfakes.

The multi-disciplinary nature of this review highlights how the design of solutions to deepfakes and future research into the technology across disciplines need to consider the function of accessibility and believability. For example, legal and political efforts to mitigate the harms of deepfakes should consider controlling the prevalence and accessibility of particularly harmful versions of these technologies and challenging the content available in consumer deepfake applications. Similarly, the perceived believability is often used to argue for how dangerous political deepfakes can be, but there is little evidence of this.

Research on the believability of deepfakes should interrogate whether hyper-realistic deepfakes effect people's attitudes more than lower quality deepfakes and if, without the obvious novelty of the technology, they provide anything useful for advertisers or other "positive" users. Similarly, research on the educational uses of deepfakes should question if "believable" educational historical deepfakes are more exciting to students than the use of costumes and actors. Research should question whether educational content must be hyper-realistic and believable. Other supposedly positive uses of deepfakes such as multi-lingual media do pose some advantages over using traditional dubbing, notably that it may help people who rely on lip-reading, but these positives need to be tested empirically by future research.

F. Academic Narratives Around the Positive and Negative Uses of Deepfakes

Our findings build upon early legal scholarship done by Chesney and Citron [7] which categorized positive and negative deepfake applications. While their taxonomy is still foundational to understanding the impact of deepfakes, our analysis takes from multiple disciplines over the first five years of the technology and highlights the complexities and problems behind categorizing deepfakes as purely positive and negative and bound within either personal/ organizational or societal domains. As shown in our data, there were more references to real world incidences of deepfake harms than of their benefits.

Often, the discussion of the supposedly positive uses of deepfakes in disciplines such as entertainment (e.g., resurrecting dead actors or recasting movies) were problematized by legal criticisms that it violates intellectual property, takes away roles from actors and runs into the same issues with consent that were more dominant in discussion around the pornographic use of the technology. Existing philosophical work on deepfake technology shows how lack of consent often underlies the harm of deepfake videos [71]. Our findings

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

encourage a more critical and consent-based approach to deepfakes, which is particularly relevant to researchers studying the supposedly positive uses of deepfakes. These researchers need to consider how even positive uses of deepfakes may undermine the consent of individuals involved and may damage people's trust in video media. They should also question if the novelty and strangeness of the content draws attention away from the positive value of deepfake content.

G. Reconsidering the Categorization of Deepfakes Uses

A significant contribution of the review to academic theory around deepfakes is the need to apply a more holistic lens to deepfakes (i.e., to take an approach to deepfakes which more accurately conveys the complexity behind positive and negative uses). Much of the literature, particularly law and politics literature viewed defamatory pornographic deepfakes as being one of the main harms of deepfakes and focused on the technology as having personal (individual) harms. The distinction between these harms is fundamental to much of the established literature on deepfakes including Chesney and Citron's work [7]. Yet, what also emerged in the analysis is that the defamatory harms of deepfakes are in themselves political (and thus societal). They often target specific groups such as politicians, celebrities and activists and are used as part of campaigns to silence and harass their mostly female targets.

It should also be noted that while in general the political harms generally exist on a social level and the pornographic harms exist on an individual level, both types of harms have the capacity to exist both socially and individually. The social harms of deepfakes will have effects on individuals and the reputational effects of deepfakes will have broad effects on society in general. Building on this, one should consider that every harm of deepfake has both a personal, inter-personal and a societal element and most positive uses of deepfake have a corresponding (and often significantly more impactful) harm. While taking care to avoid overly fearful conjecture [5], researchers should consider the uses of deepfakes as being mostly negative (or morally suspect as proposed by Rini [72]). Our work has implications for disciplines participating in the "beneficial" uses of the technology, where they should consider any associated or possible harms across individual and societal domains as well as within and outside of their own discipline.

H. Further Implications for a Holistic Approach to Deepfakes

Our work calls for deepfake literature to focus less on potential future harms and to take a more holistic view on deepfakes. We provide two contributions of a holistic approach to future research to interrogate the gaps in the literature exposed by this review. First is the necessity of more empirical research on the present impacts of deepfakes. This should specifically focus on the lived experiences of the targets of deepfakes. This would provide insight into how deepfakes harm and whether these harms are unique or an evolution of existing harms.

Secondly, the fixation on the future of deepfake technology provokes an alarmism that needs to be studied in more detail. We have found similar evidence of alarmism in academic literature to alarmism in the media narratives around deepfakes analyzed by existing research [17]. The harms of deepfakes to trust in video evidence often rely on the ability of deepfake awareness [7] to undermine faith in real video evidence [73]. Existing review work has previously critiqued the perceived fears and stigma around the use of deepfakes in academic literature [5]. Perhaps the fearful speculations around deepfake technology identified in our review of academic literature is a further indicator of the harms of broader social narratives around deepfakes. Academics across disciplines need to consider how awareness existence of this potentially harmful technology evokes overt skepticism and fearfulness of the future, detracting from the ability to focus on the present values and harms of the technology.

VIII. Limitations

Deepfakes are an emerging form of technology and any attempt to survey current attitudes and narratives around the technology must be disclaimed with the fact that this field is constantly evolving. As such, our dataset provides a broad approximation of narratives from within the first four years of academic research across stakeholder domains. We highlight the necessity of understanding divergent viewpoints across disciplines on deepfakes. Though future work making empirical claims as to differences between disciplines is necessary, it will only be valid when more research is published as outside from law and IT, as there is a comparative dearth of deepfake research outside these disciplines.

While our findings benefitted from the breadth of our review, it impacted how we carried out our qualitative analysis of our dataset. The dataset contained over 180000 words, and while the resulting document provided a wealth of multidisciplinary discourse around deepfake technology, its size hindered how reflexive the qualitative analysis could be. Qualitative analysis is influenced by the quality and quantity of the data it studies [74]. As a result, our thematic analysis was more focused on identifying broader ideas and narratives within the corpus as opposed to drawing out smaller themes within each use case.

Furthermore, while a second reviewer was involved in deciding on which papers were included/ excluded and read through the full text of each, producing inter-rater reliability would not have been practical given the size of data required to study and in light of broader theoretical concerns as to the legitimacy of inter-rater reliability as a construct [75]. Similarly, we could not include every single paper on deepfakes in our analysis - while our inclusion criteria was broad, it also had to be somewhat economical to avoid including papers which only tangentially mentioned one of the themes leading to the dataset becoming too large to analyze. This practicality also led us to focus distinctly on journal articles as many of the conference proceeding focused specifically on deepfake generation and production which were outside this reviews scope. However, we recognize the importance of

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

conference proceedings to academic work especially in AI related research.

IX. Conclusion

Our scoping review challenges current academic narratives around the definitions, novelty and impacts of deepfakes using a large multi-disciplinary dataset. There are three contributions of the review to academic theory around deepfakes and to current and future deepfake research.

1. Deepfakes are inconsistently defined but there is some general consensus as to what constitutes a deepfake.
2. Deepfakes are an evolution of previous technologies, it is better to understand how deepfakes uniquely (often by being more accessible/ believable) augment and challenge existing practices and solutions.
3. There is a conceptual issue with the assumed dichotomous positives and negatives impacts of deepfakes across distinct domains.

We would like to elaborate further on each of these points. Firstly, our research has highlighted the evolving definitional issues around deepfakes. We highlight how they have caused confusion in discussions of the novelty and impacts of deepfakes. Yet also, on a more positive note, we have shown that academics are arriving at some commonality between the use of deepfakes with more recent papers more likely to include a definition or cite notable previously established definitions. The way deepfakes have been defined has evolved during the early years of the technology to reflect academia's updated understanding of the phenomenon and the requirements of different disciplines. Our work expands and further refines previous review work on the definitions of deepfakes. Particularly we update existing theory on deepfake definitions [4] by highlighting that mimicked identity is integral to deepfakes.

This review questions the purpose of research that focuses specifically on novel technologies as having unique hypothetical harms. Research in this field should explore the harms in themselves and understanding what novel technologies change and do not change about existing solutions. Furthermore, the lenses of believability and accessibility must be considered when exploring the impact of deepfake technology. Deepfakes can be seen as one of many evolutions to synthetic media in the last decade which have furthered both the accessibility and believability of fake information online. We call for more empirical work on the existing harms of deepfakes which engages and tests the factors of accessibility and believability of the technology.

We further problematize the current attitudes towards deepfakes as having a clearly established positive and negative across distinct personal, organizational, and societal levels. There is an over-focus on dystopian and utopian potentials of deepfakes in comparison with the ethical issues with the technology itself. Clearly, broader critical issues around consent, copyright, and identity underly even the positive uses of deepfake technologies. These are vital factors to consider in the study of deepfakes. We also highlight the overlap between the personal, organizational, and societal levels of the technology. We critique existing models of deepfakes uses which divide uses into personal and societal harms/ benefits.

References

[1] M. Albahar and J. Almalki, "DeepFakes: Threats and countermeasures systematic review," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 22, pp. 3242–3250, 2019.

[2] C. Gosse and J. Burkell, "Politics and porn: How news media characterizes problems presented by deepfakes," *Crit. Stud. Media Commun.*, vol. 37, no. 5, pp. 497–511, Dec. 2020.

[3] E. Altuncu, V. N. Franqueira, and S. Li, "DeepFake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review," 2022, *arXiv:2208.10913*.

[4] L. Whittaker, R. Mulcahy, K. Letheren, J. Kietzmann, and R. Russell-Bennett, "Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda," *Technovation*, vol. 125, Jul. 2023, Art. no. 102784.

[5] P. N. Vasist and S. Krishnan, "DeepFakes: An integrative review of the literature and an agenda for future research," *Commun. Assoc. Inf. Syst.*, vol. 51, no. 1, p. 14, 2022.

[6] I. Perov et al., "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," 2020, *arXiv:2005.05535*.

[7] B. Chesney and D. Citron, "Deep Fakes: A looming challenge for privacy, democracy, and national security," *California Law Rev.*, vol. 107, no. 6, pp. 1753–1819, Dec. 2019.

[8] N.-M. Aliman and L. Kester, "VR, deepfakes and epistemic security," presented at the 2022 IEEE Int. Conf. Artificial Intelligence and Virtual Reality (AIVR), 2022, pp. 93–98.

[9] M. C. Williams, S. E. Williams, and D. E. Newby, "Artificial intelligence–based text-to-image generation of cardiac CT," *Radiol., Cardiothor. Imag.*, vol. 5, no. 2, 2023, Art. no. e220297.

[10] J. H. B. Dixon Jr., "DeepFakes: More frightening than photoshop on steroids," *Judges' J.*, vol. 58, no. 3, pp. 35–37, 2019.

[11] A. Broinowski, "Deepfake nightmares, synthetic dreams: A review of dystopian and utopian discourses around deepfakes, and why the collapse of reality may not be imminent—Yet," *J. Asia-Pac. Pop Culture*, vol. 7, no. 1, pp. 109–139, 2022.

[12] B. C. Taylor, "Defending the state from digital deceit: The reflexive securitization of deepfake," *Crit. Stud. Media Commun.*, vol. 38, no. 1, pp. 1–17, 2021.

[13] L. Wilkerson, "Still waters run deep(fakes): The rising concerns of 'deepfake' technology and its influence on democracy and the first amendment," *Missouri Law Rev.*, vol. 86, no. 1, pp. 407–432, 2021.

[14] B. Timmerman et al., "Studying the online deepfake community," *J. Online Trust Safety*, vol. 2, no. 1, pp. 1–28, 2023.

[15] C. Kerner and M. Risse, "Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds," *Moral Philos. Polit.*, vol. 8, no. 1, pp. 81–108, 2021.

[16] N. Liv and D. Greenbaum, "Deep fakes and memory malleability: False memories in the service of fake news," *AJOB Neurosci.*, vol. 11, no. 2, pp. 96–104, 2020.

[17] K. Wahl-Jorgensen and M. Carlson, "Conjecturing fearful futures: Journalistic discourses on deepfakes," *J. Pract.*, vol. 15, no. 6, pp. 803–820, 2021.

[18] R. Cover, "Deepfake culture: The emergence of audio-video deception as an object of social anxiety and regulation," *Continuum*, vol. 36, no. 4, pp. 609–621, May 2022, doi: 10.1080/10304312.2022.2084039.

[19] S. Karnouskos, "Artificial intelligence in digital media: The era of deepfakes," *IEEE Trans. Technol. Soc.*, vol. 1, no. 3, pp. 138–147, Sep. 2020, doi: 10.1109/tts.2020.3001312.

[20] D. Gamage, K. Sasahara, and J. Chen, "The emergence of deepfakes and its societal implications: A systematic review," in *Proc. TTO*, 2021, pp. 28–39.

[21] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: Generation, detection, and applications," *Int. J. Multimedia Inf. Retrieval*, vol. 11, no. 3, pp. 219–289, 2022, doi: 10.1007/s13735-022-00241-w.

[22] B. U. Mahmud and A. Sharmin, "Deep insights of deepfake technology: A review," 2021, *arXiv:2105.00192*.

[23] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[24] M. Westerlund, "The emergence of deepfake technology: A review," *Technol. Innovat. Manag. Rev.*, vol. 9, no. 11, pp. 39–52, 2019.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[25] P. N. Vasist and S. Krishnan, "Engaging with DeepFakes: A meta-synthesis from the perspective of social shaping of technology theory," *Internet Res.*, vol. 33, no. 5, pp. 1670–1726, 2022, doi: 10.1108/INTR-06-2022-0465.

[26] E. Pashentsev, "The malicious use of deepfakes against psychological security and political stability," in *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Cham, Switzerland: Springer, 2023, pp. 47–80.

[27] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *Int. J. Soc. Res. Methodol., Theory Pract.*, vol. 8, no. 1, pp. 19–32, 2005. [Online]. Available: https://doi.org/10.1080/1364557032000119616

[28] V. Braun and V. Clarke, *Thematic Analysis*. Washington, DC, USA: Am. Psychol. Assoc., 2012.

[29] G. Guest, K. M. MacQueen, and E. E. Namey, *Applied Thematic Analysis*. Thousand Oaks, CA, USA: SAGE Publ., 2011.

[30] G. Stokes, K. Sutcliffe, and J. Thomas, "Is a one-size-fits-all '12-month rule' appropriate when it comes to the last search date in systematic reviews?" *BMJ Evid.-Based Med.*, vol. 28, no. 6, pp. 359–363, 2023.

[31] P. Conte, "Mockumentality: From hyperfaces to deepfakes," *World Lit. Stud.*, vol. 11, no. 4, pp. 11–25, 2019.

[32] J. Langa, "Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes," *Boston Univ. Law Rev.*, vol. 101, no. 2, pp. 761–801, Mar. 2021.

[33] R. L. Anna, "DeepFakes: What are they, and why are they dangerous?" *South Carolina Lawyer*, vol. 33, pp. 34–39, 2021.

[34] N. Nour and J. Gelfand, "DeepFakes: A digital transformation leads to misinformation," in *Proc. Int. Conf. Grey Lit.*, Jan. 2022, pp. 55–65.

[35] D. Krishna, "Deepfakes, Online platforms, and a novel proposal for transparency, collaboration, and education," *Richmond J. Law Technol.*, vol. 27, no. 3, pp. 1–74, 2020.

[36] C. Cross, "Using Artificial Intelligence (AI) and deepfakes to deceive victims: The need to rethink current romance fraud prevention messaging," *Crime Prev. Commun. Saf.*, vol. 24, no. 1, pp. 30–41, 2022, doi: 10.1057/s41300-021-00134-w.

[37] S. E. Molina, *Lying Beneath the Surface: The Impacts of Deepfake Technology on the Privacy and Safety of the LGBTQ+ Community Articles and Surveys*, vol. 46, Nova Law Rev., Fort Lauderdale, FL, USA, 2021, pp. 251–276.

[38] R. Rini and L. Cohen, "Deepfakes, deep harms," *J. Ethics & Soc. Phil.*, vol. 22, no. 2, pp. 143–161, 2022.

[39] S. Gregory, "Deepfakes and synthetic media: What should we fear? What can we do?" Witness. 2028. [Online]. Available: https://blog.witness.org/2018/07/deepfakes

[40] S. Sylvester, *Don't Let Them Fake You out: How Artificially Mastered Videos Are Becoming the Newest Threat in the Disinformation War and What Social Media Platforms Should Do about It*, vol. 72, Federal Commun. Law J., Washington, DC, USA, 2020, pp. 369–392.

[41] B. Paris and J. Donovan, *Deepfakes and Cheap Fakes*, Data & Soc., New York, NY, USA, 2019.

[42] K. Geddes, "Ocularcentrism and deepfakes: Should seeing be believing?" *Fordham Intellect. Prop., Media Entertain. Law J.*, vol. 31, no. 4, pp. 1042–1083, 2020.

[43] C. Barnes and T. Barraclough, "Deepfakes and synthetic media," in *Emerging Technologies And International Security: Machines, the State, and War*. Abingdon, U.K.: Routledge, 2021.

[44] R. Millière, "Deep learning and synthetic media," *Synthese*, vol. 200, no. 3, p. 231, 2022, doi: 10.1007/s11229-022-03739-2.

[45] B. Sloot and Y. Wagensveld, "DeepFakes: Regulatory challenges for the synthetic society," *Comput. Law Secur. Rev.*, vol. 46, Sep. 2022, Art. no. 105716, doi: 10.1016/j.clsr.2022.105716.

[46] T. Ascott, "Microfake: How small-scale deepfakes can undermine society," *J. Digit. Media Policy*, vol. 11, no. 2, pp. 215–222, 2020, doi: 10.1386/jdmp_00018_1.

[47] R. Pfefferkorn, *'DeepFakes' in the Courtroom*, vol. 29, Boston Univ. Public Interest Law J., Boston, MA, USA, 2019, pp. 245–276.

[48] S. Maddocks, "'A deepfake porn plot intended to silence me': Exploring continuities between pornographic and 'political' deep fakes," *Porn Stud.*, vol. 7, no. 4, pp. 415–423, 2020.

[49] M. B. Kugler and C. Pace, "Deepfake privacy: Attitudes and regulation," *Northwestern Univ. Law Rev.*, vol. 116, no. 3, pp. 611–680, 2021.

[50] G. Murphy and E. Flynn, "Deepfake false memories," *Memory*, vol. 30, no. 4, pp. 480–492, 2021.

[51] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "DeepFakes: Deceptions, mitigations, and opportunities," *J. Bus. Res.*, vol. 154, Jan. 2023, Art. no. 113368, doi: 10.1016/j.jbusres.2022.113368.

[52] A. de Rancourt-Raymond and N. Smaili, "The unethical use of deepfakes," *J. Financ. Crime*, vol. 30, no. 4, pp. 1066–1077, 2022, doi: 10.1108/JFC-04-2022-0090.

[53] D. Harris, "DeepFakes: False pornography is here and the law cannot protect you," *Duke Law Technol. Rev.*, vol. 17, no. 1, pp. 99–128, 2018.

[54] A. McPeak, "The threat of deepfakes in litigation: Raising the authentication bar to combat falsehood," *Vanderbilt J. Entertain. Technol. Law*, vol. 23, no. 2, pp. 433–450, 2021.

[55] K. R. Harris, "Video on demand: What deepfakes do and how they harm," *Synthese*, vol. 199, no. 5, pp. 13373–13391, Sep. 2021.

[56] R. Pfefferkorn, "Too good to be true," NWLawyer, vol. 73, pp. 22–25, Aug. 2019.

[57] R. M. Reynolds, "When seeing is no longer believing: Deep fakes, the right of publicity, and CDA 230 reform," *Am. Univ. Intellect. Prop. Brief*, vol. 12, pp. 1–16, 2020.

[58] C. Holliday, "Rewriting the stars: Surface tensions and gender troubles in the online media production of digital deepfakes," *Convergence*, vol. 27, no. 4, pp. 899–918, 2021.

[59] S. Y. Shin and J. Lee, "The effect of deepfake video on news credibility and corrective influence of cost-based knowledge about deepfakes," *Digit. J.*, vol. 10, no. 3, pp. 412–432, 2022, doi: 10.1080/21670811.2022.2026797.

[60] N. Diakopoulos and D. Johnson, "Anticipating and addressing the ethical implications of deepfakes in the context of elections," *New Media Soc.*, vol. 23, no. 7, pp. 2072–2098, 2021.

[61] L. Whittaker, K. Letheren, and R. Mulcahy, "The rise of DeepFakes: A conceptual framework and research agenda for marketing," *Aust. Mark. J.*, vol. 29, no. 3, pp. 204–214, 2021, doi: 10.1177/1839334921999479.

[62] P. Nema, "Understanding copyright issues entailing deepfakes in India," *Int. J. Law Inf. Technol.*, vol. 29, no. 3, pp. 241–254, 2021, doi: 10.1093/ijlit/eaab007.

[63] H. Ajder, F. Patrini, and L. Cullen, *The State of DeepFakes: Landscape, Threats, and Impact*, Deeptrace Labs., Milan, Italy, 2019.

[64] A. Yadlin-Segal and Y. Oppenheim, "Whose dystopia is it anyway? Deepfakes and social media regulation," *Convergence*, vol. 27, no. 1, pp. 36–51, 2021.

[65] M. Fineman, *Faking it: Manipulated Photography Before Photoshop*. New York, NY, USA: Metropolitan Museum of Art, 2012.

[66] T. Weikmann and S. Lecheler, "Visual disinformation in a digital age: A literature synthesis and research agenda," *New Media Soc.*, vol. 25, no. 12, pp. 3696–3713, Dec. 2022.

[67] M. R. Shoaib, Z. Wang, M. T. Ahvanooey, and J. Zhao, "Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models," in *Proc. Int. Conf. Comput. Appl. (ICCA)*, 2023, pp. 1–7, doi: 10.1109/ICCA59364.2023.10401723.

[68] S. Ryan, "Epistemic environmentalism," *J. Philosoph. Res.*, vol. 43, pp. 97–112, Jan. 2018.

[69] Y. Apolo and K. Michael, "Beyond a reasonable doubt? Audiovisual evidence, AI manipulation, deepfakes, and the law," *IEEE Trans. Technol. Soc.*, vol. 5, no. 2, pp. 156–168, Jun. 2024.

[70] D. G. Widder, D. Nafus, L. Dabbish, and J. Herbsleb, "Limits and possibilities for 'ethical AI' in open source: A study of deepfakes," in *Proc. ACM Conf. Fairness, Accountabil., Transp.*, Jun. 2022, pp. 2035–2046, doi: 10.1145/3531146.3533779.

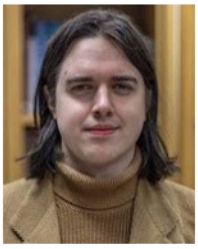
[71] A. de Ruiter, "The distinct wrong of deepfakes," *Philos. Technol.*, vol. 34, no. 4, pp. 1311–1332, 2021, doi: 10.1007/s13347-021-00459-2.

[72] R. Rini, "Deepfakes and the epistemic backstop," *Philosoph. Imprint*, vol. 20, no. 24, pp. 1–16, 2020.

[73] R. Chesney and D. Citron, "Deepfakes and the new disinformation war the coming age of post-truth geopolitics," *Foreign Affairs*, vol. 98, no. 1, pp. 147–155, 2019.

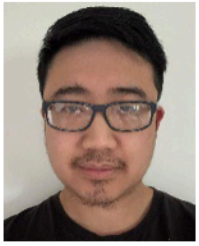
[74] J. M. Morse, "Determining sample size," *Qualitat. Health Res.*, vol. 10, no. 1, pp. 3–5, 2000.

[75] V. Braun and V. Clarke, "One size fits all? What counts as quality practice in (reflexive) thematic analysis?" *Qualitat. Res. Psychol.*, vol. 18, no. 3, pp. 328–352, 2021.



John Twomey received the B.A. degree in applied psychology from University College Cork. He is currently pursuing the Ph.D. degree with Lero—The SFI Research Centre for Software, and the School of Applied Psychology, University College Cork. He studies deepfakes through a human-computer interaction approach.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



Didier Ching received the B.A. degree in applied psychology from University College Cork and the M.A. degree in clinical neuropsychology from Leiden University. He is currently pursuing the Ph.D. degree in cognitive psychology with Lero—The SFI Research Centre for Software, and the School of Applied Psychology, University College Cork.

Matthew Peter Aylett received the M.Sc. and Ph.D. degrees in speech and language technology from the University of Edinburgh. He is an Associate Professor of Applied AI with Heriot-Watt University and is the CSO with Cereproc Ltd. He studies human-robot interaction, conversational systems, and conversational interaction.

Michael Quayle recieved the Ph.D. degree in psychology from the University of KwaZulu-Natal in 2011. He is a Professor of Psychology with the University of Limerick and an Honorary Senior Lecturer with the School of Applied Human Sciences, University of KwaZulu-Natale. He is a Funded Investigator with Lero—The SFI Research Centre for Software. He studies networks of identity and the psychological impacts of active identity production.

Conor Linehan received the B.A. and Ph.D. degrees in psychology from Maynooth University. He is a Senior Lecturer of Applied Psychology with University College Cork. He is a Funded Investigator with Lero—The SFI Research Centre for Software. His research area is human computer interaction, focusing on the design and evaluation of technology and games to support education and health behaviors.

Gillian Murphy received the B.A. and Ph.D. degrees in applied psychology from University College Cork, where she is a Senior Lecturer. She is a Funded Investigator with Lero—The SFI Research Centre for Software. Her research explores cognition in applied settings, including misinformation, memory distortion, and attention failures.