Communications of the Association for Information Systems

Volume 51 Article 14

2022

Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research

Pramukh Nanjundaswamy Vasist
Indian Institute of Management Kozhikode

Satish Krishnan
Indian Institute of Management Kozhikode

Recommended Citation

Vasist, P., & Krishnan, S. (2022). Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. Communications of the Association for Information Systems, 51, pp-pp. Retrieved from https://aisel.aisnet.org/cais/vol51/iss1/14

Communications of the Association for Information Systems

Accepted Manuscript

Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research

Pramukh Nanjundaswamy Vasist
Information Systems Area
Indian Institute of Management Kozhikode

Satish Krishnan
Information Systems Area
Indian Institute of Management Kozhikode

Please cite this article as: Vasist, Pramukh Nanjundaswamy; Krishnan, Satish: Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research, *Communications of the Association for Information Systems* (forthcoming), In Press.

Communications of the Association for Information Systems

Research Paper ISSN: 1529-3181

Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research

Pramukh Nanjundaswamy Vasist
Information Systems Area
Indian Institute of Management Kozhikode

Satish Krishnan
Information Systems Area
Indian Institute of Management Kozhikode

Abstract:

We are witnessing a growing concern around the impact of hyper-realistic synthetic media and its dissemination in what is widely known as "deepfakes." However, the phenomenon's relative newness and the fragmented nature of existing research on it across several disciplines leave a lot to be desired. In addition, empirical research is scarce, and deepfake literature is moving forward in a variety of directions without a strong theoretical foundation. The fragmented nature of extant literature on the phenomenon calls for consolidation in order to produce a thorough and current summary of deepfake research to date and map its present intellectual boundaries. We offer an integrative overview of the existing corpus of research on deepfakes in this paper, noting the wide range of domains, samples, and approaches used. We point out various gaps in deepfake narratives, including definitional concerns, a lack of comprehensive demographic and cross-geographic coverage, a lack of theoretical underpinning, thematic tensions, and imbalances in the extant literature on deepfakes. In the last section of the paper, we propose future research directions, which include a set of themes and research questions and a theoretical framework to guide future research on the topic.

Keywords: Deepfakes, Fake News, Integrative Review, Misinformation, Synthetic Media.

[Department statements, if appropriate, will be added by the editors. Teaching cases and panel reports will have a statement, which is also added by the editors.]

[Note: this page has no footnotes.]

This manuscript underwent [editorial/peer] review. It was received xx/xx/20xx and was with the authors for XX months for XX revisions. [firstname lastname] served as Associate Editor.] **or** The Associate Editor chose to remain anonymous.]

1 Introduction

We are living in a post-truth era where actual facts are supplanted by alternative facts, and shared fiction pervades our lives across various facets such as news, sports, politics, and other areas (Harari, 2018; McIntyre, 2018). In this new age, manipulated images and videos of people are now increasingly finding their way online, leading to the rise and proliferation of a new phenomenon called 'deepfakes' (Vaccari & Chadwick, 2020). Deepfakes refer to hyper-realistic synthetic media where an individual's face in a photo or a video is swapped with that of another person (Öhman, 2020; Somers, 2020). Software applications leverage machine learning (ML) algorithms to create these fake videos and continually improve them by mimicking the individual's expressions and voice modulations which end up making these 'deepfakes' all the more realistic and indiscernible from the authentic videos (Maras & Alexandrou, 2019).

It has never been easier to generate fake photos and videos to deceive the eye than it is now, thanks to easy access to various tools that allow for digital media manipulation (Maras & Alexandrou, 2019). Breakthroughs in the field of artificial intelligence (AI) imply that deepfake technologies are rapidly improving, with deepfakes generated online growing at an alarming rate of nearly nine hundred percent year on year since 2019 (Tammekänd et al., 2020). However, the negative impact of deepfakes is expected to far outweigh its positive aspects, with deepfakes being ranked the most serious crime threat posed by AI with potentially serious ramifications (UCL, 2020). Corporates and governments are starting to take notice of this rise in deepfake content, with the U.S government issuing over ten initiatives and acts in 2019 to tackle deepfakes (Tammekänd et al., 2020), while technology giants like Facebook, Microsoft, and Amazon have launched deepfake detection challenges worth over ten million U.S dollars (Facebook, 2019; Tammekänd et al., 2020). Academic research on deepfakes is slowing but steadily gaining traction, with research on the technologies that underpin their creation and detection dominating the research agenda (Hancock & Bailenson, 2021), while social sciences research on deepfakes is starting to explore the potential effects and threats posed by this new form of fake news (Carvajal & Iliadis, 2020).

We contend that the increased attention garnered by deepfakes in academic and mainstream discussions which span disparate disciplines, including computer science, humanities, social sciences, legal and political studies is fragmented and calls for a need to integrate the different perspectives under discussion and the mechanisms by which deepfakes are created, consumed, and disseminated. To gain a holistic understanding of deepfakes, it is critical to synthesize and identify the intellectual structure of existing literature on the subject. We address this need by conducting an integrative review of literature on deepfakes.

The integrative literature review is a type of research that analyzes, critically examines, and synthesizes representative literature on a topic in order to generate new frameworks and perspectives on the subject (Torraco, 2005; Webster & Watson, 2002). This form of review is especially appropriate when the extant research on a topic is scattered across disparate areas and has not been systematically integrated and analyzed by scholars (Scully-Russ & Torraco, 2020). Such is the case with literature on deepfakes. In the past few years since the emergence of deepfakes as a phenomenon, few scholars have conducted literature reviews on the state of deepfake research. However, these reviews have been highly restrictive in limiting themselves to either a survey of preliminary research on the topic (Carvajal & Iliadis, 2020) or an analysis of online news articles on deepfakes (Westerlund, 2019), or a review focused on literature related to deepfake creation (Albahar & Almalki, 2019) and detection techniques (Albahar & Almalki, 2019; Botha & Pieterse, 2020; Verdoliva, 2020) or one aimed at highlighting implications in a specific domain such as communication studies (Godulla et al., 2021). We contend that all these studies have adopted a narrow approach that does not offer a comprehensive overview of past literature on deepfakes. Furthermore, despite the growing literature on deepfakes in recent years, empirical research on deepfakes remains scarce (Ahmed, 2021b), and the literature is moving forward without a strong theoretical framework to guide its progress.

Our research aims to transcend the gaps in these previous literature reviews by using an integrative approach that contributes to the literature and advances research in the domain. We take a broader view of deepfakes and incorporate literature that is not confined by certain fields, which allows us to provide contemporary and holistic insights on deepfakes. In doing so, we offer a robust research profile and major themes discussed in extant research on the phenomenon. Without a clear understanding of extant research on deepfakes, scholars researching deepfakes may be discouraged from conducting research in this domain, and practitioners may not be able to incorporate the accumulated knowledge in dealing with deepfakes. Hence, in line with our research purpose, the research questions (**RQ**s) of our study are:

RQ1: What is the status of the research profile on existing literature on deepfakes?

RQ2: What are the research gaps, limitations, and recommendations for scholars and practitioners in the context of future research on deepfakes?

To answer these research questions, this integrative review on deepfakes aims to review all empirical and theoretical research on deepfakes through an analysis of articles published so far on the topic in the past two decades since 2001. To identify the scholarly contribution of deepfakes to literature, we conducted a literature search of articles across leading databases for peer-reviewed literature. Although deepfakes as a term traces its origin to 2017 (Albahar & Almalki, 2019), we ensured that the literature survey accounted for publication lags, if any, and aimed to cover all articles since 2001. In the first part of this review, as a response to RQ1, we identify, synthesize, and present a current profile of extant literature on deepfakes, including annual publication trends, geographic coverage of studies and frameworks, variables, and measures related to the characterization and evolution of deepfakes. In response to RQ2, the second part of the review focuses on delineating the accumulated body of research on deepfakes since its inception to the present day and describes research findings published in literature over the past two decades since 2001. In doing so, we detail out the findings from each of the selected articles on deepfakes published in peer-reviewed journals during the timeframe and highlight gaps and limitations in extant literature. We then present potential avenues for future research, as well as a state-of-the-art framework for deepfake research based on extant literature and insights gained from this integrative review.

The remainder of the paper is laid out as follows: A brief background of deepfakes is presented in the second section. The third section profiles the existing literature on deepfakes and describes the methodological processes used to carry out this integrative review. Section four enlists the key aspects discussed thus far in deepfake literature, while section five reviews the empirical literature on deepfakes. This synthesis of extant literature segues into the identification of gaps and limitations in section six and forms the basis for future research themes. Section seven contains recommendations as well as a research framework for future researchers to address existing knowledge gaps. Finally, the paper concludes with a discussion of the study's theoretical and practical implications, as well as its limitations.

2 Background on the Topic

This section draws attention towards a brief background of Deepfakes as a topic. Given that the vast majority of existing research on deepfakes focuses on its application for deception, notably as an amplifying factor for fake news, it is critical to understand the origins of the phenomenon. Hence, prior to delving into a discussion of deepfakes, it's critical to situate the phenomenon within the context of fake news. The term 'fake news' has been so pervasive (Kalpokas & Kalpokiene, 2022a) and misused that it has lost all meaning (Jankowicz, 2020). Nonetheless, it must be acknowledged that the contemporary information environment, enabled by “automation and algorithmization” (Kalpokas & Kalpokiene, 2022a, p. 7), tends to foster the propagation of intentionally created false information, which is a critical characteristic that determines the usage patterns of deepfakes (Kalpokas & Kalpokiene, 2022a). This also explains the motivation for deepfakes as a choice over other methods for creating and disseminating false information. In this context, while fake news focuses on the content of messages, “deepfakes create a simulation of the speaker” (Maddalena & Gili, 2020, p. 16), further destabilizing information: not only the content, but also the source or subject which can be fabricated (Kalpokas & Kalpokiene, 2022b). In doing so, deepfakes elevate the complexity of fake news to a new level by making it even more difficult to decipher authentic information (Breen, 2021). Additionally, while image fabrication is not a new phenomenon, deepfaked media is entirely modified or generated by AI (Schick, 2020), wherein the deepfaked victim is falsely accused of saying or doing something through the deepfake (Kalpokas & Kalpokiene, 2022c). Furthermore, three characteristics of deepfakes make them particularly concerning: (a) the low entry barriers for deepfakes as a result of their ability to be generated with minimal skills and resources, (b) the ease with which content can be shared on social media networks, and (c) the ever-growing amount of digital material featuring a sizable proportion of the global population that can be used as training data (Whittaker et al., 2021), which can make deepfakes even more indiscernible to the naked eye. These characteristics of synthetic content pose a severe threat to the public's perception of events (Breen, 2021) and will further exacerbate the fake news crisis (Whittaker et al., 2021). However, remedial measures such as collaboration between media houses and platform players (Vizoso et al., 2021), calls for implementing ethical guidelines (Deloitte, 2019) and regulating AI content (Oxford Analytica, 2022), as well as increased awareness of deepfakes, may help mitigate the negative effects of deepfakes. Additionally, we are witnessing a favorable trend in the introduction of positive use cases for deepfakes, which are also referred to as synthetic data in this context (e.g., Chen et al., 2021; Shin et al., 2018).

The term ‘Deepfake’ finds its origins in 2017 in a Reddit community user by the name ‘deepfakes’ who bragged about the technological advancements which allowed face swapping in adult videos using the faces of celebrities with open-source ML tools (Cole, 2017). Early definitions of deepfake in research articles refer to it as “a portmanteau of deep learning AI and faked imagery” (Wagner & Blewer, 2019, p.33). Afchar et al. (2018) define deepfakes as a technique aimed at swapping the face of a targeted person with another one in videos, while Yang et al. (2018) define deepfakes as AI-generated fake images or videos. Öhman (2020) terms deepfakes as videos that are hyper-realistic and generated using deep learning techniques which superimpose a person’s face on top of another. A broader definition for deepfakes may be noted in the research article by Nguyen et al. (2021), who define deepfakes as content that is synthesized with the aid of AI and categorized as either face-swaps, lip-sync, or puppet-masters. While face-swaps involve superimposing the images of a target on the source, lip-sync involves videos that alter lip movements to align with an audio clipping, and puppet-masters include videos of a target person on ‘puppet’ and facial expressions and head and eye movements of another person or the ‘master’ which are used to animate the video of the ‘puppet (Nguyen et al., 2021). A noteworthy mention in this context is the recent extension of the construct to include manipulated media beyond the faces of individuals, wherein Zhao et al. (2021) incorporate the term in cartographic studies to discuss manipulation in geospatial imagery.

In the past few years of its evolution as a topic, many studies have been published on deepfakes. However, the phenomenon is still new, and extant literature on deepfakes is fragmented across several disciplines. Further, empirical research on deepfakes is scarce, and literature is advancing without strong theoretical underpinnings. To help guide future empirical studies and theoretical development in the field, a comprehensive review of deepfake literature and a mapping of its present intellectual boundaries may prove useful to researchers. Hence, we offer a thorough overview of the state of deepfake literature in this paper.

3 Methodology

Integrative reviews offer an immense opportunity to capture the state of knowledge on a particular topic to date and act as a catalyst for future research on it (Torraco, 2016). Furthermore, the review allows for the generation of new perspectives previously unexplored in literature and “can be an influential force in shaping practice and the future directions of the field” (Torraco, 2016, p. 67).

3.1 Review Process

Deepfakes have garnered significant attention in computer science (Güera & Delp, 2018), while interest in the topic has been steadily rising in the social sciences domain with deepfake related implications in the areas such as politics (Dobber et al., 2021), journalism (Yadlin-Segal & Oppenheim, 2021) and ethics (Öhman, 2020). In line with the multidisciplinary nature of the topic, we imposed strict selection criteria on articles to be considered as part of the sample for the study. We searched for articles using two major databases, namely Scopus and Web of Science. The reasons for our choice of these two databases were twofold. Firstly, the creation and consumption of deepfakes pertain to psychological implications associated with manipulated media, and secondly, the concept of deepfakes presents multifaceted connotations within the technological field. Hence, Scopus and Web of Science (WoS) are chosen for their leading status and exhaustive coverage of peer-reviewed literature (Fahimnia et al., 2015; Minola et al., 2014; Sigerson & Cheng, 2018).

While our initial choice of keywords was primarily informed by prior literature reviews by Carvajal & Iliadis (2020) and Westerlund (2019) that aimed to be comprehensive in coverage, we complemented this search with a follow-up of the databases using additional search terms to reduce the element of subjectivity in sample selection (David & Han, 2004). For this purpose, we executed a search on google scholar for the keyword “deepfakes.” Based on the search results, the first 50 results were sorted based on relevance, reviewed, and variant terms “face manipulation,” “fake video,” “video manipulation,” “audio manipulation,” “fake audio,” and “voice manipulation” were also added to the list of search terms and the search was executed on Scopus and Web of Science databases for these keywords on the title and abstracts of articles. Furthermore, in line with meeting the objective of studying the evolution of deepfakes even prior to the formal origins of the term, studies published between 2001 to 2022 were included in the study.

We determined the inclusion and exclusion criteria for the study based on past research (e.g., Khan et al., 2021; Schilpzand et al., 2016). In the context of inclusion criteria, we included those articles that had a substantiative component of discussion on deepfakes. Empirical studies on deepfakes are still scattered and few in number (Ahmed, 2021b). Hence, we expanded our selection to include both empirical and theoretical papers, including discussion and opinion papers on deepfakes. The studies were limited to peer-reviewed journal articles published in the English language any time between 2001 to 2022. We excluded articles that were solely concerned with the technical aspects of deepfake production or detection.

The keywords-enabled search conducted in the month of November 2021 yielded 319 articles comprising 212 and 107 studies from peer-reviewed journals in English from Scopus and WoS respectively. The articles were screened for duplicates, thereby yielding a total of 236 studies. Both authors then examined the titles and abstracts of each of these articles and marked them with a check (if it looked promising), a question mark (if we were unsure of its inclusion) or left unmarked (if eliminated). The articles that had inter-rater disagreement were reread in more detail to make a final decision for inclusion. In doing so, 150 articles were excluded and 86 were retained. Next, we obtained the full text of all articles which had received a check or question mark and reviewed them in the same manner and determined their eligibility for inclusion in the final sample. In doing so, we excluded 35 studies and retained 51 as part of the final sample. We also performed a forward and backward citation chaining and reviewed any other potential articles which may be considered for the review (Webster & Watson, 2002). As a result, we included 6 articles in the final sample. We reran the search twice, once in December 2021 and again in April 2022 to account for additional journal articles published during the time lag. Our rerun of the search in December 2021 led us to add 8 additional studies to the final sample. Finally, we ran the search again in April 2022 to account for additional studies that may have been published in the timeframe since January 2022 and included 3 additional studies in the final sample. In total, we identified 68 journal articles that were accessible in English, published in a peer-reviewed journal or in press, and were from the subject area of social sciences, arts, psychology, business, and management. The key findings from empirical and non-empirical papers are summarized in Appendix A.

3.2 Research Profile

The profiling of the research articles indicates that research on deepfakes has gained considerable traction in the last three years since 2019 (see Figure 1a). Next, we performed an analysis of the first author affiliations (by geographic location) to analyze the geolocation and study context of literature on deepfakes. The analysis by first author affiliation suggests that the USA (n=27), U.K (n=8), Australia (n=7), and Canada (n=6) collectively account for more than 70 percent of all reviewed articles in the final sample (see Figure 1b). Empirical studies on deepfakes are only a handful to date (Ahmed, 2021b), with just over 40 percent (29 studies) out of the final sample (see Figure 1d).

The quantitative studies have employed samples predominantly from the USA (e.g., Barari et al., 2021; Shin & Lee, 2022) and the U.K (e.g., Fido et al., 2022; Köbis et al., 2021; Vaccari & Chadwick, 2020) (see Figure 1c) indicating that deepfake literature is largely restricted to a few geographic areas although the potential for sophisticated video manipulation techniques and the use of deepfakes to spread and reinforce disinformation is rapidly growing (Paterson & Hanley, 2020) and warrants its analysis as a phenomenon that occurs worldwide with far-reaching ramifications. Lastly, among the empirical studies on deepfakes, a bulk of them (n=17) have adopted a qualitative methodology (see Figure 1d).

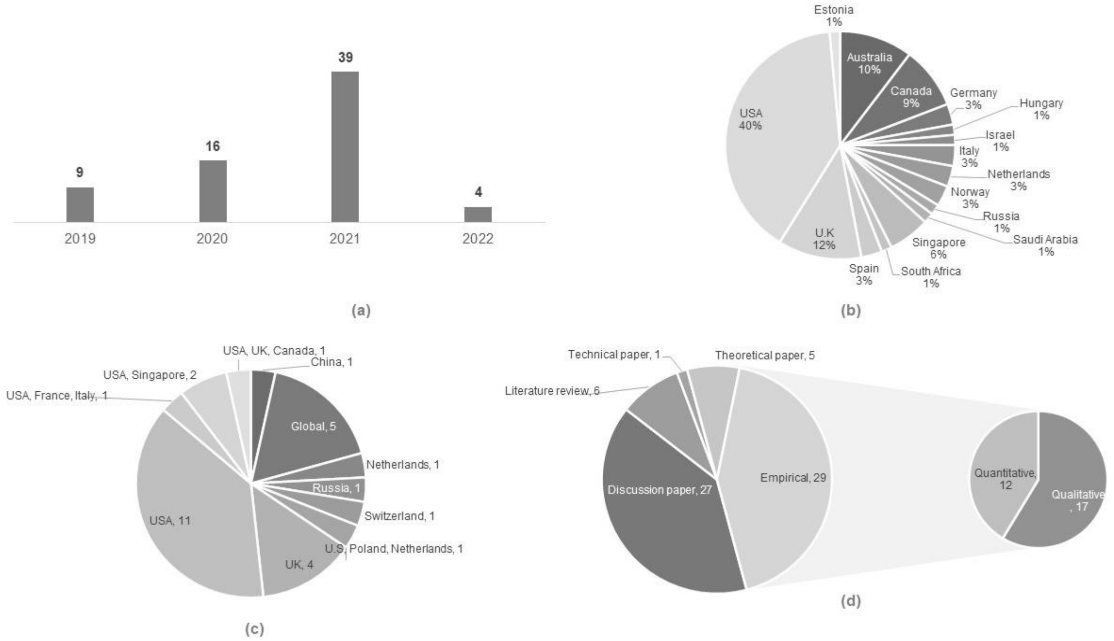

Figure 1. Distribution of Studies by (a) Year of Publication (b) Country based on First Author Affiliation (c) Geographic Focus of Samples and (d) Type of Article with Research Design

4 Synthesis of a Growing Body of Deepfake Literature

The past few years have witnessed a growing body of literature on deepfakes. We can safely say that the topic is gaining momentum in scholarly communities. It comprises a diverse body of work across several domains and professions and has arrived at a logical point where it can be reviewed as a whole and integrated to provide new paths for scholars to take it forward. In this section, we synthesize extant literature on deepfakes and elaborate on various facets through which it has been studied thus far.

4.1 Past Literature Reviews on Deepfakes

A synthesis of this body of deepfake literature must begin by acknowledging the contribution through prior reviews and building on it to inform research. One of the earliest reviews of the literature on deepfakes may be found in the article by Albahar & Almalki (2019), who trace the history and origins of deepfake technology and analyze the ways in which deepfake photos and videos are created. The review article briefly discusses certain applications of the technology, its impacts, and the ethics around it. However, a significant focus of the article is mainly on deepfake detection methods. In a similar vein, the literature review by Botha & Pieterse (2020) discusses deepfake creation and detection methods as a prominent theme of its literature review but also discusses detection methods for fake news as a broader topic.

Carvajal and Iliadis (2020) carry out an analysis of scholarly literature on deepfakes, but their analysis is limited to a classification of articles into several themes with statistical charts to depict the distribution across themes and relationships of the data. Their review suggests that most articles on deepfakes are related to the science, technology, engineering, and mathematics (STEM) area, while a growing interest is seen in the social aspects of the technology and its impacts. Godulla et al. (2021) examine deepfake literature and briefly discuss opportunities and risks, but a key focus of their review is the implications of deepfakes in the field of communication studies. Verdoliva's (2020) review deals with media forensics with a special focus on deepfakes. However, the review mainly discusses methods related to deepfake detection, datasets for training algorithms, and upcoming challenges which these must address. Westerlund (2019) attempts a more holistic review in examining the deepfake phenomenon, its creation, benefits and harms, and ways to combat them. However, the review focuses on news articles from media outlets such as CNN, CBS News, CNET, financial times, and so on. A summary of literature reviews discussed in this section is provided in Table 1 below.

We discuss three key shortcomings in these reviews. First, deepfake detection has been a key focus of most reviews which flows from the technological advancements aimed at deepfake creation and detection, while other aspects of deepfake engagement, which include the motivations behind deepfake creation, the accuracy judgments of the deepfake viewer, and deepfake dissemination patterns have been largely ignored. Second, most studies have been restrictive in their approach by examining the phenomenon in the context of a specific domain or industry, such as media forensics or communication studies. Third, most studies do not offer a theoretical perspective or research questions to guide future research in the deepfake domain. Based on these shortcomings, we contend that literature reviews on deepfakes so far have adopted a narrow approach and do not cover the phenomenon holistically in their reviews. At the same time, the fast-evolving nature of deepfakes and their potential ramifications calls for an immediate need to integrate research on the phenomenon and draw out its intellectual boundaries to guide theoretical development and future empirical studies in the field.

Table 1. Prior Literature Reviews on Deepfakes

Source	**Key findings**
Albahar & Almalki (2019)	Examines how deepfakes are created and discusses techniques for deepfake detection
Botha & Pieterse (2020)	Conducts a review of creation and detection techniques in the area of fake news including deepfakes
Carvajal & Iliadis (2020)	Conducts a preliminary literature review of academic works in the area of deepfakes
Godulla et al. (2021)	Focuses on a literature review of deepfake articles in the domain of communication studies
Verdoliva (2020)	Provides a review of methods used for detection of manipulated images and videos with specific focus on deepfakes
Westerlund (2019)	Performs a literature review of deepfakes with specific focus on online news articles

4.2 Deepfake Engagement Process

Deepfake engagement may be viewed as a three-tier process starting with deepfake creation, deepfake dissemination, and deepfake detection. We elaborate on each of these process steps below.

4.2.1 Creation of Deepfakes

Deepfake creation is now easier than ever before, thanks to the increased sophistication possible through deep neural networks and realistic content offered through GANs (Whittaker et al., 2021). These video manipulations have been witnessing an alarming rise in numbers, with over 85 thousand harmful deepfake videos detected up to December 2020, with the numbers doubling every six months (Petkauskas, 2021; Sensity, 2020). Our review of the literature on deepfakes points to three factors that may be contributing to the worrying surge in the number of deepfakes being developed. First, the growing sophistication of the GAN approach has meant that it is now possible to create increasingly convincing deepfakes which could go undetected by the untrained eye (Maras & Alexandrou, 2019). Second, digital forensics has largely focused on detecting low-level alterations in images, while research on the detection of face manipulation is growing but still sparse (Maras & Alexandrou, 2019). This, in turn, would mean that it would take years before deepfakes are detected reliably by the systems (Porter, 2020). Third, the widespread availability of deepfake creation technologies has meant that it is easy to produce deepfakes without the need for expert intervention (Gosse & Burkell, 2020)

A fourth factor in the form of network effects may also have a role in the creation of deepakes, while it is unclear whether this factor is exclusive to a specific region. In this context, de Seta (2021) discusses the dynamics specific to the phenomenon in China, where deepfakes are referred to as ‘Huanlian.’ The author highlights the peculiarities of China’s technology industry and its contribution to the global advancement of ML alongside the contribution of local players in popularizing synthetic media. In that context, the author points to network effects in accelerating technical know-how wherein ‘huanlian’ creator communities contribute to the exchange of knowledge on technical expertise and showcase their deepfake creations.

4.2.2 Dissemination of Deepfakes

The dissemination of deepfakes has garnered significant attention in the literature. Chesney and Citron (2018) point to three reasons for the spread of deepfakes. First, they attribute the spread to an ‘information cascade’ where people stop paying attention to what they receive, assume it to be true, and share it further. Second, individuals are more receptive to negative information and exhibit intentions to share it. Third, filter bubbles insulate us against information contrary to our beliefs and act as reinforcements which eventually lead us to share deepfakes. On a related note, prior research finds that an individual's specific interests and cognitive abilities may be reasons why they mistakenly share deepfakes (Ahmed, 2021c), while a lack of information cues and an impression of correctness may also contribute to people propagating deepfakes (Ahmed, 2021a)

Echo chambers, like filter bubbles, work to reinforce our cognitive processes through the heuristic of confirmation bias (Sample et al., 2020). While these processes apply in the case of fake news dissemination in general, scholars point out that the likelihood of sharing a fake news story increases with the inclusion of an image in the story (Fenn et al., 2019), which largely explains the potential of deepfakes to outperform other fake news items in terms of dissemination. Furthermore, in the context of the spread of deepfakes, Dasilva et al. (2021) leverage social network analysis to analyze the actors who control the sharing of deepfakes on Twitter. Their analysis reveals that although adult content dominates deepfakes, public attention is predominantly focused on political deepfakes. This finds support in prior research by Maddocks (2020).

Lastly, literature highlights specialized efforts aimed at spreading disinformation. O'Donnell (2021) discusses 'disinformation-for-hire' companies which offer a variety of services, including posting and commenting on articles on social networking sites (SNSes) with efforts aimed at targeting western businesses. In mentioning this, it is not difficult to imagine 'deepfake-for-hire' services on similar lines (O'Donnell, 2021).

4.2.3 Detection of Deepfakes

We now discuss aspects related to deepfake detection. Research on face manipulation detection is gaining traction as deepfake proliferation grows but so is its dependence on the availability of datasets that could be used to assist with face manipulations in videos (Maras & Alexandrou, 2019). In this context, it's worth noting that the attention garnered by political deepfakes may be due not only to the fact that these individuals are public figures but also to the large number of images and videos available of these individuals, making it easier to train AI deepfake systems to manipulate videos (Dasilva et al., 2021; Westerlund, 2019).

Several public and private organizations have started to realize the importance of this technological advancement by launching bespoke initiatives aimed at deepfake detection. For example, the United States Defense Advanced Research Projects Agency (DARPA) has launched a research initiative in the field of media forensics to develop technological prowess aimed at assessing the integrity of digital media (Maras & Alexandrou, 2019). In a similar vein, Gregory (2021) discusses the role of a specific human rights and civic journalism network in helping fight media manipulation and efforts aimed at providing access to reliable information. Kerner and Risse (2021) discuss legislations passed by several states in the USA to combat deepfakes. For example, California passed two bills recognizing the threat of deepfakes to elections and digital forgeries of adult content, while Virginia and Texas have passed similar legislations (Kerner & Risse, 2021). In a similar vein, the European Commission highlighted the need for governments to invest in research aimed at fighting misinformation campaigns and the necessity of holding social media companies accountable for content on their platforms (European Commission, 2018).

Several platform players have taken notice of this trend and are devising strategies to combat it. Facebook is funding initiatives aimed at creating a corpus of videos that can aid researchers in combating deepfakes through precise detection mechanisms (Vizoso et al., 2021). Twitter is taking a similar approach through an intricate set of rules which go about identifying tweets carrying manipulated content and warning users about them alongside carrying the authentic source wherever possible and eliminating the doctored content (Vizoso et al., 2021).

Deepfakes, as highlighted earlier, have impacted the media industry by questioning journalistic credibility and creating a sense of mistrust (Yadlin-Segal & Oppenheim, 2021). These exacerbated perils of deepfake technology have led several media houses to launch initiatives to detect and debunk deepfake misinformation campaigns. Vizoso et al. (2021) discuss initiatives by The Wall Street Journal, The Washington Post, and Reuters while they also point to the increased collaboration between media outlets and platform companies (e.g., Reuters' collaboration with Facebook) to combat fake news in general. Furthermore, they recommend marking fraudulent content as fake rather than deleting it (as it may have already been viewed), as this may help raise awareness about misinformation and safeguard future users from engaging with similar news items. Langa (2021) expresses a similar view in mentioning that an outright ban on deepfakes may be ill-advised, while media houses and democratic institutions may use this deepfake threat to drive more transparency and mitigate potential harms.

Scholarly discourse on deepfake detection comprises the ability of humans to detect deepfakes and the technological capabilities towards the same. In the context of human abilities, Köbis et al. (2021) undertook an experimental study with U.K citizens to measure their ability to detect deepfakes. The outcomes of this study revealed that humans overestimated their ability to detect deepfakes but were not able to do so even when awareness of the same and financial incentives was introduced. Köbis et al. (2021) point out that humans are more biased toward mistaking deepfake videos as authentic instead of the contrary, while concluding that deepfake detection was more of an inability to do so rather than a lack of motivation. The lack of accurate detection may also be attributable to the growing sophistication of deepfake technology, which Maras & Alexandrou (2019) say would eventually go undetected by the naked eye.

Deepfakes demand new tools to validate the authenticity of information online (French et al., 2021). In the context of technologies related to deepfake detection, Sample et al. (2020) point toward a need to combine data science and linguistics, which may offer tools for the rapid detection of deepfake propaganda. The authors further put forth information trajectory modeling as a counter tactic in line with a similar suggestion by Cybenko et al. (2002). However, unlike the linguistics approach, this method lacks data on the spread of authentic data, which makes comparison with fake news spread challenging (Sample et al., 2020).

An excessive focus on deepfake detection may hinder us from progressing technology linked to the detection of similar but subtly distinct phenomena. In this context, Yankoski et al. (2021) draw attention to shallow fakes, which could be original images or videos that have been subtly edited to change the context through examples such as a relabeling exercise or a slowing of video frames (Denham, 2020). While shallow fakes may not always be aimed at pushing disinformation campaigns, they certainly drive specific individual behaviors (e.g., anti-vaccination memes), which are worrisome and have deeper ramifications as compared to deepfakes (Yankoski et al., 2021). Hence, the authors call for the advancement of AI techniques that are fast improving in detecting deepfakes but lack the sophistication to detect shallow fakes as the two technologies are isolated from one another (Yankoski et al., 2021). They also point to semantic analysis and algorithms which not only detect a fake news item but also identify the concerted multimodal disinformation campaigns across several platforms. However, deepfake detection mechanisms certainly need to maintain and gain more traction to keep pace with technologies being leveraged for deepfake creation (Ring, 2021).

Sample et al. (2020) also discuss the possibility of using game theory to combat disinformation. In this context, they discuss the topic of attitudinal inoculation wherein users are preemptively warned, and false narratives are preemptively debunked, a combination of which leads to users discarding information when exposed to deceptive data such as deepfakes. Furthermore, they also discuss the possibility of using archival reputation analysis to combat disinformation wherein prior reputational data about the reporter and publisher may be leveraged to assess the credibility of news items. Chesney and Citron (2018) discuss the creation of immutable life logs, which will help in building an authentication trail using blockchain technologies through which the privacy of individuals may be preserved and help deal with deepfakes effectively. However, such approaches may lead to the access and use of personal data in the absence of appropriate security standards, but overall, the benefits of such an approach outweigh the harms (Chesney & Citron, 2018).

4.3 The Promise and Peril of Deepfakes

The engagement process with deepfakes broadly involves four key stakeholders: the deepfaked person, the deepfake maker, the deepfake viewer, and the deepfake disseminator (Pavis, 2021), with each one of them playing a role in the influence of deepfakes. In this context, the literature on deepfakes lists several positives, but the risks of this technology outweigh the benefits. The harms inflicted by deepfakes may broadly be classified under three areas, namely: harms to viewers, harms to subjects, and harms to social institutions (Diakopoulos & Johnson, 2021). Chesney and Citron (2018) attempt a similar classification in that the harms inflicted by deepfakes are segregated into several categories, namely: individual, organizational, or societal levels. Deepfake related harms at the individual level and organizational level could be in the form of humiliation, instigation of violence, exploitation through blackmailing, causing psychological damage, or sabotage through reputational harm (Chesney & Citron, 2018). At the societal level, deepfake related harms could mean distortion of democratic discourse, manipulation of electoral outcomes, erosion of trust in both public and private institutions, undermining of public safety, disruption of diplomatic relations and national security, and undermining of the journalistic profession (Chesney & Citron, 2018). Here, we elaborate on the key harms and benefits of deepfake technology.

4.3.1 The Epistemic Threat to Viewers

In the context of harm to viewers, Porter (2020) discusses the role of deepfakes in creating an ‘epistemological pluralism’ where people would not just question reality but resort to accommodating the plurality through several ways of acquiring knowledge. According to Fallis (2020), this threat is so profound that deepfakes would inflict epistemic harm on us wherein we may fail to acquire true beliefs as a result of this technology. Further, it may not just cause epistemic harm but also lead us to an ‘epistemic wrong’ where deepfake technology would lead us to move from what merely existed in our minds to actually viewing it in reality through doctored videos (Kerner & Risse, 2021). In the process, people will succumb to epistemic helplessness where they would give up on critically examining information and conform to their own worldview in what would eventually lead to a "worrisome epistemic future” (Rini, 2020, p. 13). Ziegler (2021) elaborates on this in suggesting that the worry is twofold: first, in people ceasing to be critical, and second, in being unable to detect manipulation, which in combination leads to ‘epistemically unhygienic minds’ (Ziegler, 2021, p. 3). Harris (2021), however, suggests that concerns over deepfakes bringing about an epistemic catastrophe are exaggerated, but they would indeed be psychologically impactful wherein people would challenge the most authentic video for authenticity if delivered by a dubious source. Köbis et al. (2021) also express similar concerns over the epistemic threat of deepfakes, while Chesney & Citron (2018) suggest that this, in turn, may benefit guilty individuals in what may be termed as ‘liar’s dividend’, which helps them evade accountability by challenging the veracity of the media which might in fact be true.

4.3.2 The Reputational Injury to Subjects

In the context of harm to subjects who are part of the video, Harris (2021) suggests that associating a celebrity personality with a deepfake may interfere with the individual’s ability to construct a desired persona around self. Similarly, a politician's appearance in a deepfaked video may have an adverse effect on the candidate's reputation, with voters refusing to take the candidate seriously (Harris, 2021). de Ruiter (2021) argues that the technology's repercussions are far greater when the person is deepfaked as part of adult video content, resulting in extreme embarrassment and distress, which Franks & Waldman (2018) identify as a central wrong that causes “reputational injury” (Franks & Waldman, 2018, p. 893). Furthermore, Diakopoulos & Johnson (2021) assert that the fundamental misattribution motif prevalent in deepfake videos violates the subject's ownership rights as the victim in the doctored video, while Mullen (2022) discusses potential litigation options for deepfake victims, including invoking misappropriation doctrines or seeking an injunction against the deepfake, with the goal of removing the content from the internet and limiting its undesired effects.

4.3.3 The Distortion of Credibility

Deepfakes' impact on social institutions has been seen most strongly in the political arena, where the technology has been used to deceive voters, create political tensions, and attempts have been made to sway the outcome of democratic electoral processes (Chesney & Citron, 2018; Diakopoulos & Johnson, 2021; Paterson & Hanley, 2020). In a similar vein, Holliday (2021) discusses several instances of deepfakes of politicians which have circulated in the media during the U.S electoral process, while Chesney & Citron (2018) allude to its potential to impact diplomatic relations between countries. Beyond the political landscape, deepfakes could have severe ramifications on the news media industry as well, where journalists look upon this technology to impact journalistic credibility, accentuate mistrust in the media and undermine a shared sense of social and political reality (Dasilva et al., 2021). Deepfakes, according to Rini (2020), could have an impact on our testimony practices, as recordings will no longer be trusted as a genuine source of truth in courts of law, necessitating the exploration of new testimonial standards. In a similar vein, deepfakes created to carry out cybersecurity attacks on biometric systems could damage the credibility of platforms and also result in financial losses.

4.3.4 A Sliver of Hope amidst the Chaos

The narrative around deepfakes is undoubtedly concerning with its potential to inflict harm at various levels, but the technology is not without a sliver of hope. Deepfakes hold immense potential, which can be leveraged in beneficial ways. At an individual level, deepfake videos may allow people to have experiences that may otherwise not be possible in real life owing to the dangers or challenges involved in such acts. For example, Langa (2021), in discussing how deepfakes facilitate personal expression, quotes the example of a text-to-speech technology company that helped create a voice for a radio host who had lost his voice due to a medical condition. Wiederhold (2021) discusses the possibility of combining deepfake technology with augmented reality applications, which can aid clinicians in offering personalized therapeutic sessions to patients. Deepfake videos, for example, can also be used to bring individuals from the past to life through manufactured videos, giving history lectures a new lease of life (Chesney & Citron, 2018), while Ham (2021) puts forth as an example the possibility of reviving Albert Einstein to teach students during a physics class. In a similar vein, they may be beneficial to the domain of art wherein dead performers may be resurrected on stage through a combination of live acting and technical wizardry. In this context, Mihailova (2021) discusses how deepfakes are being employed in art museums for audience engagement, advertising, and educational outreach in what may be termed 'edutainment.' However, such use for educational purposes also raises questions of moral ambiguity over medium specificity, with deepfakes exposing museum exhibits to heightened scrutiny from visitors (Mihailova, 2021).

Whittaker et al. (2021) discuss the benefits of deepfakes in the context of marketing and advertising and put forth a series of propositions that may be subject to empirical verification. They indicate that deepfake enhanced messaging, in addition to being tailored and boosting the customer's capacity to visualize the use of products and suggest that these could be more powerful than AI-based advertising messages. They also advocate the integration of deepfake technology in AI-powered services such as chatbots and self-service technologies, which could go on to enhance the customer's sense of empowerment and improve perceptions of their emotional intelligence (Whittaker et al., 2021). Furthermore, they suggest that disclosure in deepfakes could enhance business outcomes alongside protecting them from deviant use. Kietzmann et al. (2021), however, warn of a 'sleeper effect' when the consumer perceives the advertisement as fake, and this perception continues to have a lasting effect on the advertisement in the future.

4.3.5 Deepfakes for Fun

Prior literature proposes another application of deepfakes, one in which they are neither generated to reap benefits nor cause harm. In this context, there have been reports of individuals creating deepfake videos for the sake of amusement. Numerous apps and websites have appeared in recent years that enable users to produce deepfake videos with face swaps while maintaining a sense of humor (Staff, 2019). In this context, from a Kantian perspective, each human has a right to digital self-representation, and if a deepfake portrays them in a way they would not wish to be seen or heard, such fakes are morally reprehensible (de Ruiter, 2021). In a similar vein, Bode (2021) suggests that while most people think of deepfakes as digital trickery on social media, spreading them outside of their original context can lead to deception, even if the initial intent was not to deceive. Furthermore, some scholars have been critical of such deepfakes for their apolitical approach toward culture and ethnicity. Ayers (2021) analyzes videos of action stars from the 1980s and observes that such face-swaps have generally taken a neutral stance by remaining apolitical and retaining the racial differences and cultural specificities in the original video. Allison (2021) anchors the argument around this neutrality and calls for the use of deepfakes to usher in racial justice rather than assuming a racial-neutral stance.

Current legal frameworks are inadequate to address the plight of individuals whose images or videos have been deepfaked (Chesney & Citron, 2018; Harris, 2019), blurring the lines between what constitutes fun and what constitutes deception. A blanket ban on deepfakes may not be appropriate because digital content alteration is not necessarily a problem (Chesney & Citron, 2018). However, the design of a legislative framework restricting harmful applications of the technology while permitting beneficial applications may be difficult but not impossible (Chesney & Citron, 2018). In the light of challenges associated with deepfake prohibition, Chesney and Citron (2018) examine a wide range of civil and criminal liabilities, highlighting the difficulties associated with the former option if the plaintiff is unable to attribute the deepfake to its creators or if the creator or platform circulating it is located outside the country, which can make it difficult to leverage civil remedies effectively. The global nature of online platforms might pose a challenge for legal procedures that are constrained by geographical limits (Chesney & Citron, 2018). Criminal liability could also be invoked, although only to a limited extent, when deepfakes, such as those of an explicit nature, are used to defame or impersonate another person with the aim of harming that person while being aware that the video is fake (Chesney & Citron, 2018).

However, criminal liability is unlikely to be an effective alternative when deepfakes are used in the electoral process and may face constitutional obstacles similar to earlier cases involving election-related falsehood, in which the courts have required a convincing basis to limit speech (Chesney & Citron, 2018).

Harris (2019) focuses on explicit deepfakes involving non-celebrities and explores legal solutions while lamenting the absence of legal recourse for deepfaked victims. In this aspect, civil remedies are difficult because the deepfake makers may not have produced the content for commercial gain, but there is a sliver of hope if the victim can establish that the content was deliberately intended to cause serious emotional anguish (Harris, 2019). Criminal liabilities may be hampered by the fact that the deepfake creator may not have intended to harm the featured person, although laws pertaining to explicit non-consensual content may be the most effective legal option, and there is an immediate need for legislatures to examine options to protect victims of personal deepfakes (Harris, 2019). In this context, a number of nations are examining the possibility of imposing regulations on deepfakes as a means of protection against the emerging threats posed by technological advancements (Hine & Floridi, 2022), and we contend that regulations will improve as the scope of technological governance expands.

4.3.6 Deepfakes: A Concern or Hype?

Although several studies have highlighted the potential harms which may be inflicted by deepfakes at various levels, the verdict on this remains divided, with scholars also suggesting that deepfakes may not be as disastrous as they are projected to be. Chesney and Citron (2019) argue that while deepfakes are undoubtedly dangerous, they may not necessarily be disastrous as platforms will improve in their capabilities with regard to detection and timely flagging of deepfakes while democratic societies will embrace a post-truth world where nothing is accepted at face value and learn to live with lies. Furthering the argument around the unwarranted concerns over deepfakes, Ray (2021) argues that while there is little evidence to claim that deepfakes have impacted the Australian elections, they possess the potential to erode voter trust and disrupt electoral outcomes. Simonite (2020) puts forth a similar argument around the lack of impact of deepfakes on the U.S elections. In a similar vein, while some reports predicted that deepfake disinformation campaigns would increase during the COVID-19 pandemic (Avast, 2020), other studies found no evidence of deepfaked visuals in their analyses of such campaigns, despite the presence of manipulated visuals created with simple tools (Brennen et al., 2021).

However, it may not be wise to dismiss the concerns around deepfakes. A consideration of the deceptive abilities of deepfakes places it higher than other forms of digital deception we have witnessed thus far. According to Kietzmann et al. (2020), deepfakes could be alarmingly successful as a form of digital trickery due to two factors: believability and accessibility, where the former refers to the increasingly convincing deepfakes being created, making them more believable than ever before, and the latter refers to the easy access to deepfake technology, allowing for the creation of such increasingly convincing deepfakes.

Digital deception may either be message-based or identity-based, with the former referring to the communication between two or more agents while the latter refers to a false representation of one’s identity (Hancock, 2009). Prior digital deception attacks have largely taken the form of fake news items, phishing attempts, identity theft, and fake websites to deceive viewers online. In this context, it is noteworthy that these forms of digital deception have predominantly been message-based, with the exception of phishing attempts which may be both identity and message-based. The usage of rich media as a form of deception in the case of deepfakes, as opposed to lean media such as text, has a differing impact on the efficiency of communication (Short et al., 1976). Furthermore, deepfakes allow for the manipulation of an existing person's identity with the intent to deceive (Mohammed & Salam, 2021). Doing so boosts the deception's credibility and authenticity (Diakopoulos & Johnson, 2021) while also distorting people's basic understanding of reality (Conwell, 2020). This very quality of deepfakes makes it more devastating than other forms of digital deception and neglecting it as hype could prove costly in the long run.

4.4 The Ethics behind Deepfakes

Ethics is a key issue in dealing with emerging technologies (Richardson et al., 2021; Stahl, 2021).In this context, a discussion on deepfakes is incomplete without a mention of the moral and ethical implications of the technology. In an elaborate discussion on this topic, Öhman (2020) uses the levels of abstraction involved in creating the deepfake video as a measure of whether it can be considered morally permissible or otherwise. Stadler (2019) draws parallels between deepfakes and techno biological bodies (e.g., cyborgs) in that both face similar ethical concerns which they violate through a lack of consent and freedom of speech. Mihailova (2021), in discussing the use of deepfakes in museums, cites examples when visitors' faces are inserted in films as social experiments to draw on the creative prowess of deepfake technology but also raises concerns over the ethics involved in such an act where personal data may be compromised through unauthorized use and misappropriation. Diakopoulos and Johnson (2021) highlight the importance of anticipating the ethical implications of deepfakes and call for anticipatory governance, which would facilitate an examination of these ethical implications while they are still in the early stages of development and help in addressing them suitably. de Ruiter (2021) offers an alternative perspective in discussing the moral stance of deepfakes, unlike other articles which have mostly discussed the ethical aspects. The author argues that deepfakes could be morally suspect as the actions may violate fundamental moral norms, but they may not be morally wrong in that the technology can be leveraged for good through reinforcement of people's autonomy or empowerment. However, a Kantian perspective on deepfakes would lead us to the importance of respecting people as ends in themselves (Kant & Beck, 1959) and help clarify, albeit to a reasonable extent, the intrinsic moral wrong involved in creating a deepfake against the will of people who are represented in it (de Ruiter, 2021).

4.5 New Frontiers of Deepfake Technology

Extant literature has thus far mostly referred to recorded videos in deepfake related discussions. However, the frontiers of deepfake technology are fast moving from recorded to real-time synthetic media (Hancock & Bailenson, 2021). We are witnessing the launch of sophisticated and easily accessible AI tools such as DeepFaceLive (iperov, 2020/2021), which allows individuals to transform their faces in real-time and even use them on video conferencing platforms (Tran, 2021). The technology even allows the participant's gaze to be manipulated to point towards the camera when they are actually looking elsewhere (Hancock & Bailenson, 2021). While tools such as DeepFaceLive are still far from perfect, this contentious technology could spell chaos for people on live streaming and conferencing platforms (Thalen, 2021).

Downstream effects of these real-time deepfake technologies may extend to incorporate real-time filters, which alter not just the eye gaze but also facial expressions with the potential to impact interpersonal dynamics. For example, Oh et al. (2016) altered the facial expressions of participants in a virtual communication environment by enhancing their smiles and noted positive effects of doing so on interpersonal communication. In a similar vein, Leong (2021) discusses the possibility of individuals being able to view personalized role models of themselves who excel with increased confidence and creativity at tasks in real-time.

Lastly, the parallel developments in technologies powering augmented reality (AR) and virtual reality (VR) and their increasing adoption (French et al., 2020) open the possibility of moving toward a world of synthetic reality that combines the potential of deepfake technology with AR/VR. In a world of immersive synthetic reality, entire replicas of real-life individuals may be portrayed to feel part of a natural environment and even transported to places that never existed before (Pasquarelli, 2019). Synthetic media and its potential to blur lines between reality and the virtual world have been previously discussed in the literature (Schultze et al., 2008). However, with this new blend of synthetic reality alongside the emerging possibilities of leveraging real-time AI-mediated communication in real-time, much remains to be seen in terms of the impact of deepfakes on a social plane.

5 Review of the Empirical Literature

Our review of the empirical literature on deepfakes revealed progress on several fronts, while much remains to be accomplished to advance the body of knowledge on the phenomenon. The issues are reasonably settled on three counts: First, political leaders, celebrities, and explicit content are the prime targets for deepfakes with a discernible impact on democratic discourse and the electoral process. Second, social media has fostered the virality of deepfake content, fueling polarization and pushing fact-checking capabilities to the edge. Third, individuals' cognitive faculties help in critical evaluation and avoidance of falling prey to deepfakes. However, we contend that even these issues are half-baked and deserve more attention in empirical studies. We will first elaborate on the key findings from past literature (see also Appendix Table A1) and then dwell on areas that deserve empirical validation.

5.1 Key Findings from Past Empirical Studies

Empirical studies validate that content of an explicit nature, political personalities, and celebrities are among the ones capturing maximum attention among deepfakes (Dasilva et al., 2021; Holliday, 2021), with the political deepfakes garnering maximum attention in empirical studies (e.g., Vaccari & Chadwick, 2020; Yu et al., 2021), possibly due to its potential to manipulate the electoral process (Bazarkina & Pashentsev, 2019). Also, empirical studies expose the gender and racial prejudices in deepfake generation. Individuals are found to be more lenient in judging deepfake generation and dissemination when the victim is a celebrity or a male personality (Fido et al., 2022). Racial distinction in terms of white dominance has been highlighted in deepfake discussions (Ayers, 2021; Holliday, 2021), while scholars have also argued for the use of deepfakes to further racial justice (Allison, 2021)

The role of social media in deepfake engagement cannot be dismissed. While political interests may spur individuals to inadvertently share deepfakes online (Ahmed, 2021c) and also impact people's perceptions towards political candidates on viewing such microtargeted deepfakes (Dobber et al., 2021), individuals' size of social media network may moderate this influence of political leanings on inadvertent sharing of deepfakes (Ahmed, 2021c). However, deepfakes threaten the value that social media brings along. Exposure to deepfakes tends to reduce trust in social media news (Ahmed, 2021b; Vaccari & Chadwick, 2020). Such skepticism runs the risk of transforming to cynicism and clouding the benefits of accidental exposure to news on social media (Ahmed, 2021b) while journalists' concerns over deepfakes are also vivid in their challenge to render verified content to the public (Wahl-Jorgensen & Carlson, 2021; Yadlin-Segal & Oppenheim, 2021). In a not so surprising response, we are noticing significant collaboration and efforts by journalistic media and social media giants in fighting deepfakes (Vizoso et al., 2021).

Several psychological factors have been discussed in how people engage with deepfakes. Empirical studies reveal that individuals' cognitive abilities clearly have an edge in preventing inadvertent sharing of deepfakes (Ahmed, 2021c, 2021b, 2021a). However, people are predisposed to believe deepfakes are legitimate and overestimate their ability to recognize them (Köbis et al., 2021), and confirmation biases may also make them more vulnerable to deepfake news that reinforces their existing beliefs (Shin & Lee, 2022). Furthermore, incidental emotions such as anger, anxiety, and happiness have minimal effect in improving deepfake recognition (Yu et al., 2021), although information cues may aid in determining the authenticity of deepfakes (Ahmed, 2021a). In this context, concerns are raised over the blurring of lines between actual and digital reality caused by deepfakes (Conte, 2019), with exposure to deepfaked media potentially leading people to false beliefs (Fallis, 2020).

5.2 A Call for more Empirical Attention

We now examine aspects that warrant greater consideration in empirical studies.

While political personalities and celebrities have garnered the most attention in deepfake empirical literature, deepfakes' expansion into more constructive domains opens up other avenues for empirical investigation. For instance, positive aspects of deepfakes are gradually emerging in empirical studies that examine how they may enrich the narrative surrounding societal issues such as artistic creativity in museums (Mihailova, 2021), alleviating privacy concerns in healthcare (e.g., Thambawita et al., 2021), and so on. With synthetic data being investigated for use in training AI models without compromising privacy (Gooding, 2021), empirical studies must examine how such developments might alter individuals' privacy calculus and if models trained on synthetic data truly reduce biases in decisioning processes. In a similar vein, with synthetic models increasingly being leveraged in advertising campaigns (Campbell et al., 2021) and deepfakes predicted to disrupt fashion advertising (Nast, 2021), it is unclear whether brand loyalty and its determinants would remain unchanged, while empirical research may revalidate the pathways connecting engagement and brand loyalty in the face of the epistemic crisis.

Empirical studies exploring the role of social media in deepfake engagement must consider four factors. First, platform-level discrepancies may highlight the subtleties of how users may express differing degrees of mistrust toward deepfakes on different platforms, such as LinkedIn versus Instagram. Second, many research studies have reported results based on people who were aware of deepfakes (Ahmed, 2021c, 2022), while public knowledge of deepfakes is still in its infancy, with many people uninformed of the phenomenon (Ahmed, 2021c), which may have resulted in biased findings. It would be worthwhile to investigate whether the results thus far are still valid as deepfake awareness expands. Third, the hyperrealism and novelty of deepfakes pose an epistemological challenge (Fallis, 2020). In such a post-truth world, it would be important to reassess the function of social media news in promoting online participation. Fourth, while prior empirical research on fake news indicates that it spreads quicker than legitimate news due to its novelty (Vosoughi et al., 2018), it would be prudent to verify whether this holds true for deepfakes given their high degree of novelty. Additionally, across the factors discussed above, the facets of personality traits warrant empirical examination to further our understanding of their significance in the generation and transmission of deepfake media.

While gender and racial concerns surrounding deepfakes have been examined in the context of platform governance (e.g., Allison, 2021; Ayers, 2021), the cultural power of such deepfaked images and films is contingent on human viewers and their networked publics. With growing concerns about AI ethics (Pazzanese, 2020) and calls for implementing ethical guidelines (Deloitte, 2019) and regulating AI content (Oxford Analytica, 2022), it would make empirical sense to determine whether such regulations and associated punitive measures affect individuals' decision-making regarding the creation, sharing, and engagement with deepfakes and whether such policies also pave the way for provocative questions around the tensions between artistic expressions and ethical compliance.

We contend that deepfake deception and detection are fundamentally different from that of other forms of deceptive content, making deception extremely impactful and detection incredibly challenging. The impact of deception caused by deepfakes is significantly greater than that of other types of deception for the following reasons. To begin, the importance of visual signals in human perception is unequivocally established (Posner et al., 1976). Second, visual messages are considerably more likely to be recalled than verbal messages (Graber, 1990; Prior, 2014). Finally, consumers place a higher premium on audiovisual content than on verbal content due to its greater likeness to reality (Metzger & Flanagin, 2007). This pervasive nature of deceit casts doubt on the efficacy of intervention mechanisms and is best validated empirically. Additionally, while it is evident that deepfakes are a concern, considerable research remains on the types of incentives or awareness that can aid in their detection (Köbis et al., 2021). In this regard, the literature on fake news may serve as a starting point for mixing and matching combination treatments to study variations in detection abilities (Moravec et al., 2020), raising the question of whether such combinations perform effectively even when deepfakes are included. And, over time, do online filter bubbles and echo chambers alter the effectiveness of such interventions?

While we highlighted several avenues for empirical investigation, it would be beneficial to be cognizant of methodological shortcomings in previous empirical investigations in order to enhance the rigor in future research. For example, in some studies (e.g., Ahmed, 2021c, 2022), cognitive ability was operationalized as a dimension of verbal intelligence (Colom et al., 2005). If, however, this construct actually facilitates critical engagement with deepfakes, it merits further examination across additional dimensions such as numerical and visuospatial capabilities (Ahmed, 2021c). Additionally, the malleability of memory revealed by deepfakes must enrich metacognitive models of memory (e.g., Jacoby & Kelley, 1987; Mazzoni & Kirsch, 2002) on how people believe external sources of information when they lack clear recollection of events. In a similar vein, self-report data in various empirical studies (e.g., Ahmed, 2021c, 2021b) incorporates an aspect of participant subjectivity by design (Tifferet, 2021) which can be addressed by exploring mixed-method and experimental research designs that may offer a holistic view of the phenomenon. According to fake news studies that have documented cultural differences in fake news consumption (e.g., Borges-Tiago et al., 2020), potential cultural differences such as values and norms may also influence empirical findings that have thus far focused primarily on the USA and the UK. Hence, macro-level analyses may also be used to examine how country-level data on the evolution of and skills in information and communication technology (ICT) affect the commercialization of deepfakes.

We conclude this section by attempting to further push the envelope on possibilities for empirical research. When we have previously argued for empirical attention on deepfakes, we have mostly referred to recorded videos and photos containing deepfake content. However, technological advancements enable real-time use of AI filters (Tran, 2021), with these filters being utilized to enhance interpersonal dynamics (Oh et al., 2016) and so on. We encourage scholars to conduct empirical studies on the downstream implications of such novel modes of interpersonal communication on the emotional and cognitive responses of individuals. Lastly, a frequently overlooked feature of deepfake empirical research is the nonconsensual victim who is falsely accused of saying or doing something through the deepfake. The impact of deepfakes on self-identity and how such fakes can result in victim humiliation, intimidation, or extortion warrants further discussion in the literature.

6 Research Gaps and Potential Research Areas

This review of deepfake literature offers an integrated perspective on the interdisciplinary nature of deepfake research. However, some gaps in the existing literature may need to be addressed appropriately in order to further research on this topic. In this section, we examine those gaps and limitations and make recommendations for further research through a set of key research questions (see Appendix B).

6.1 Definitional Issues

The literature around deepfakes faces definitional issues. We highlight three key aspects in this context. First, it may not be accurate to refer to deepfakes as a new phenomenon, as video alterations have been documented throughout history dating all the way back to the second world war (Margry, 1992), although technological advancements have now made such manipulations highly sophisticated (Whittaker et al., 2021). Second, while the role of sophisticated AI techniques in the creation of deepfakes is clear from this review, we notice confusion around the conceptualization of deepfakes in literature. It is not clear if it only refers to a face-swapping technology (Meskys et al., 2020) or if it includes text as well over and above multimedia content (Pietro et al., 2020), while references to other forms of fakery such as cheap and shallow fakes (Hight, 2021) obfuscates the taxonomical clarity of the term. Third, the term 'fake' is etymologically erroneous because it refers to something that is not genuine and counterfeit in nature (Merriam-Webster, n.d.), and the majority of references to deepfakes are associated with those with a malicious intent to deceive, despite recent literature demonstrating the benefits of deepfakes.

These definitional discrepancies necessitate scholarly attention targeted at bridging gaps in the existing body of knowledge. Future research should address these definitional gaps by developing a precise conceptualization of deepfakes and defining what they are and are not. For instance, should modified text be regarded as deepfakes? Should photographs that have been altered without the aid of AI be deemed deepfakes?

6.2 Lack of Standard Measures

Our analysis of empirical research on deepfakes reveals that while various constructs have been included in studies (see Appendix A), there is a lack of standard measures to operationalize several of these constructs. Several variables such as deepfakes concern and exposure (Ahmed, 2021b), trust in social media news (Vaccari & Chadwick, 2020), and others, were examined using questionnaires developed particularly for the study and with limited potential to expand beyond a specific area. There is an urgent need to investigate such shortcomings and explore new manifestations, as well as to build valid and trustworthy measures for deepfake-related constructs. This will also contribute to the conceptual growth of deepfakes, given their multi-domain nature.

6.3 Antecedents and Outcomes

Scholars have investigated the deepfake phenomenon from multiple perspectives, including a variety of antecedents and their influences on specific outcomes (see Appendix A). However, as noted earlier, empirical research on deepfakes remains scarce, and much remains to be investigated in terms of antecedents and outcomes. For example, while echo chambers and filter bubbles have been discussed for their potential to fuel deepfake dissemination (Chesney & Citron, 2019; Sample et al., 2020), we find a lack of empirical studies investigating the role of these situational aspects in influencing deepfake engagement. Future research may address the shortcomings through an examination of specific individual characteristics of deepfake viewers, which determine their accuracy judgment, the deepfake receiver's intentions to disseminate it further, and the ramifications on the deepfaked person and receivers as well.

6.4 Lack of Cross-Geographic Coverage

Our analysis of empirical studies revealed that the samples were largely from the USA, the U.K, and the Netherlands, followed by Singapore, Canada, Britain, and Poland. However, a 2020 study on the state of deepfakes showed that the target countries for deepfakes were dominated by USA, U.K, South Korea, India, and Japan (Sensity, 2020). A similar study undertaken a year earlier showed that the victims of deepfakes belonged to USA, U.K, South Korea, Canada, and India (Sensity, 2019). While USA and U.K have garnered significant attention in terms of the country of origin of scholars engaged in deepfake research and also samples involved in empirical studies, other countries targeted by deepfakes, such as South Korea, Canada, and India, have lacked research in comparison.

Despite the limited cross-geographic coverage of deepfake research, investigations to date have detected geographic and cultural variations (e.g., Ahmed, 2021c). Deepfakes have been projected as a new form of fake news (Botha & Pieterse, 2020), and hence we draw parallels between the two domains. Prior research indicates that fake news capitalizes on the cultural values of the intended audience (Sample et al., 2020). Thus, we argue that geographic and cultural factors influence the production, dissemination, and influence of deepfakes on the audience and that future deepfake research must account for culturally and geographically diverse studies in order to provide a critical and holistic assessment of the phenomenon.

6.5 Investigation of Demographic Characteristics

An analysis of empirical studies in our corpus revealed that while the majority of them attempted to account for demographic characteristics (e.g., Ahmed, 2021a), the effects of these variables largely remain underexplored in deepfake literature. Also, we identified some studies which resorted to a purposive oversampling of a particular religious group (e.g., Dobber et al., 2021). We argue that taking demographic diversity into account when developing sampling strategies may yield novel insights. We draw on earlier research on fake news to support our case. For instance, prior research has revealed age-related variations in the dissemination of fake news (Guess et al., 2019) and gender-related differences in the likelihood of trusting fake news items (Shu et al., 2018). Thus, additional empirical research on the effects of demographic variables in the context of deepfake creation, dissemination, viewing, and their impact is needed before a consensus may be arrived at on the varied influences of deepfakes.

6.6 Theoretical Grounding

A concerning feature of deepfake literature is the lack of an overarching theoretical framework. This absence of a framework to guide empirical research on deepfakes has led researchers to use a variety of approaches to theoretically ground their hypotheses.

Table 2. Theoretical Underpinnings in the Extant Literature on Deepfakes

Theoretical lens	How is the theory used in deepfake literature?	Examples from deepfake literature
Dual process theory	Used to explain the role of political interest and cognitive abilities in the inadvertent sharing of deepfakes	Ahmed (2021c)
	Used to study how inadvertent sharing of deepfakes could eventually lead to social media skepticism	Ahmed (2021b)
Limited capacity model	Used to explain how individuals possess limited cognitive abilities and only process salient features of any information they are exposed to.	Ahmed (2021b)
Prominence-interpretation theory	Leveraged to explain how few salient features processed through limited cognitive abilities form the basis for them to evaluate the credibility of information online	Ahmed (2021b)
Signal detection theory	Used to examine how people distinguish authentic video clippings from deepfakes	Köbis et al. (2021)
Signaling theory	Used to demonstrate that perceptions about the cost of generating deepfakes play a critical role in determining their credibility	Shin & Lee (2022)
Misattribution theory	Used collectively to develop their hypotheses related to incidental happiness and study their effects on affective polarization and deepfake recognition	Yu et al. (2021)
Feelings-as-information theory		

Deepfake research scholars have largely leveraged theories from social psychology to conduct empirical studies (see Table 2), while a few empirical studies have taken a predominantly atheoretical approach and based their research model and hypotheses largely on findings from prior studies (e.g., Ahmed, 2021a). Certain marketing and advertising scholars have also attempted to guide deepfake research in their domains through conceptual frameworks that encourage researchers to investigate how consumers perceive and interpret hyper-personalized advertisements (Kietzmann et al., 2021). They suggest that the sophistication of manipulation enabled by deepfakes may increase the perceived authenticity and creativity of the advertisement, resulting in an overall positive effect (Campbell et al., 2021), although disclosure of deepfake use is recommended to enhance business outcomes and avoid deviant use (Whittaker et al., 2021). However, our review uncovered no related empirical work that has used these approaches to date, but we anticipate that will change in the coming days. In summary, much work remains to be done in the context of deepfake empirical research, and it must be grounded in a solid theoretical foundation. Future research should examine why people believe in and share deepfakes, as well as the personal, emotional, and logical appeals used by deepfake makers, the factors that influence a deepfake's accuracy judgment, the viewer's intent to spread it further, and the impact of deepfakes on both the viewer and the deepfaked person.

6.7 Deepfake Viewing-Sharing Dissociation

It might be the commonly held view that people are more likely to share deepfakes if they believe they are true. Recent studies on misinformation trends, on the other hand, suggest that people's accuracy judgments may differ significantly from their sharing judgments on SNSes (Pennycook et al., 2021, 2020). Scholars have attempted to discern this disconnect between viewing-sharing judgments in fake news research with several explanations. For example, this dissociation may be linked to confusion wherein the individual genuinely believes the false news item to be authentic. Otherwise, the sharing may be purposeful in nature and motivated towards advancing one’s personal agenda. Lastly, inattention due to distraction on SNSes could also be a reason for sharing misinformation without appropriate judgment. (Pennycook & Rand, 2021). However, in the context of deepfake research, much remains to be examined in terms of factors that exacerbate this disconnect between viewing and sharing of deepfakes, while a similar examination of factors may be required across domains dominated by deepfakes such as politics, art, and so on. Doing so would help frame precise distinctions for accuracy judgments and sharing intentions of deepfakes.

6.8 Thematic Tensions

The literature on deepfakes has largely focused on adult content and politically motivated video manipulations. Within this context, we observe divergent views with a lack of studies to examine them with greater clarity. We discuss three key issues. First, in the context of creating deepfakes that are of an explicit nature, issues have been raised about which acts of creating deepfakes are morally permissible and which are not, using yardsticks such as the level of abstraction and distinctions between morally wrong and suspect (e.g., de Ruiter, 2021; Öhman, 2020). However, we find little evidence of empirical investigations examining these moral tensions. Second, the literature on deepfakes has largely focused on adult content and politically motivated video manipulations, wherein scholars have discussed the potential of this technology to inflict reputational harm, distort the democratic discourse (Chesney & Citron, 2019, 2018; Diakopoulos & Johnson, 2021), erode trust in public and private institutions (Chesney & Citron, 2018), infuse a sense of skepticism in journalistic content (Chesney & Citron, 2018; Sample et al., 2020) and so on. However, even in the political context, there are divergent views on the impact of deepfakes, with some dismissing the concerns (Simonite, 2020) while some claim that it could disrupt the electoral process (Diakopoulos & Johnson, 2021; Gosse & Burkell, 2020; Paterson & Hanley, 2020). The diversity of perspectives in this context allows for an examination of deepfake influence in a variety of contexts, where much remains unexplored. Third, most empirical research has focused on analyzing interactions with political deepfakes (e.g., Ahmed, 2021c; Dobber et al., 2021; Vaccari & Chadwick, 2020; Yu et al., 2021), while a handful of studies have focused on non-political deepfakes (e.g., Ahmed, 2021b, 2021a; Köbis et al., 2021). This dual categorization of studies raises questions about generalizability, as findings from a specific category of deepfakes may not be applicable across domains (Ahmed, 2021a). In this context, it is critical to assess the applicability of deepfake research findings in a given domain, like politics, to deepfakes in other domains and to illustrate the differences and similarities between the two.

6.9 Divergent Perspectives

We encounter several divergent perspectives in the extant literature on deepfakes which we elaborate further in this section.

On moral grounds, the social consequences experienced by victims such as reputational damage and embarrassment are well established in sexual abuse literature (Bloom, 2014). Given the focus of deepfakes on explicit content, such ramifications may extend to deepfakes as well, although we have little understanding of the personal consequences of deepfake victimization. However, some scholars argue that deepfakes are morally suspect but not wrong (e.g., de Ruiter, 2021). This divergent perspective requires further evaluation by bringing together the digital, psychological, and corporeal factors to unpack the subtleties of morality.

It remains unclear if deepfake awareness improves individuals' detection abilities. While some studies find value in information cues helping determine the authenticity of deepfakes (e.g., Ahmed, 2021a), others find no improvements in detection accuracy on driving awareness or providing incentives (Köbis et al., 2021). Building on this in the context of social media, while an individual's network and usage patterns have been shown to prevent inadvertent sharing of deepfakes (Ahmed, 2021c), deepfake exposure may elevate skepticism (Ahmed, 2021b) and erode trust in social media news (Vaccari & Chadwick, 2020). Since awareness of deepfakes is still growing (Ahmed, 2021c), we may reach a broad consensus on this as time progresses.

We notice divergent perspectives on the use of deepfakes in several fields. In the domain of advertising, while deepfakes have been argued to bring novelty to advertising campaigns, they are also speculated to alter social norms and erode consumer trust (Kietzmann et al., 2021). In journalism practice, while some highlight the journalistic concerns (Wahl-Jorgensen & Carlson, 2021) and steps taken by media houses to combat deepfakes (Vizoso et al., 2021), others argue that deepfakes are being leveraged by journalists as a double-edged sword to restore faith in their own work by blaming the doubt and disbelief in society on technological advancements (Yadlin-Segal & Oppenheim, 2021). In the domain of politics, while literature alludes to a reasonable clarity on the ramifications of deepfakes on the democratic discourse (e.g., Ahmed, 2021b; Chesney & Citron, 2018), we encounter positive use cases of this technology to contextualize messages to regional dialects during the 2019 elections in India (Christopher, 2020; Jee, 2019) thus leaving the debate wide open.

Around issues of privacy and societal concerns, while deepfakes have been criticized for advancing racial differences (Ayers, 2021), deepfakes have also been encouraged to be viewed as a tool to promote racial justice (Allison, 2021) and generate empathy in people towards war-stricken countries (e.g., the joint project by UNICEF and MIT Boston) (Deep Empathy, 2017). In a similar vein, accusations over misappropriation of personal data by deepfakes (Mihailova, 2021) aimed at identity theft (Ring, 2021) are countered with positive use cases for synthetic data to protect patient privacy in healthcare applications and to train deep learning models (e.g., Crystal et al., 2020; Shin et al., 2018).

In summary, the verdict is split on moral lines, genuineness of journalistic concerns, the role of deepfakes in fields such as advertising, and interventions to improve deepfake detection. However, we see a silver lining in the evolving role of deepfakes in the electoral process, alleviating data privacy concerns and bringing about positive societal changes while we must wait and watch how positive use cases of deepfakes evolve. To gain a deeper understanding of these issues and progress towards convergence, we call for mixed methods research to present a comprehensive view wherein, to begin with, the large portion of qualitative studies, which formed over fifty five percent of the overall empirical corpus, may complement their understanding from quantitative data analysis to validate if divergences still persist or not. We also recommend more sophisticated data analysis techniques, preferably longitudinal in nature, especially to capture the evolution of trends in the context of specific scenarios highlighted earlier in this section.

7 A Theoretical Framework for Deepfake Research

We now propose a comprehensive framework (see Figure 2) based on the review and gaps identified in the published studies. The proposed theoretical framework synthesizes the process of creating, disseminating, viewing, and detecting deepfakes as well as their consequences at the individual and institutional levels. We discuss the theoretical underpinnings of this framework and then elaborate on the key aspects of the framework. Situating deepfakes within the broader context of fake news draws attention to parallels, if any, between the two phenomena. Our framework for deepfakes, however, is distinct from frameworks proposed for fake news (e.g., Domenico et al., 2021) on three counts. First, unlike fake news, which is created with a malicious intent, the motivation for creating deepfakes has a silver lining in that it may stem from a positive intent such as protecting patient privacy during doctor-patient interactions (e.g., Yang et al., 2022), creating educational videos (Griffin, 2019), and other similar motivations which we shall discuss in this section. Second, and perhaps most obviously, this positive usage of deepfakes may result in favorable consequences such as the provision of empathetic care to patients during clinical interactions (Yang et al., 2022) and so on. Third, whereas the target of fake news can range from an individual to a news item to a country, the aim of a deepfake is almost always an individual, which makes the consequences for the target a certainty and more exacerbated as in the case of fake news due to the audio-visual nature of the manipulated content.

Prior studies find disinformation messages accompanied by a deepfake video to increase its persuasiveness (Hwang et al., 2021). This inherently persuasive nature of deepfakes drives our theoretical choices, which are borrowed from two persuasion theories, namely the elaboration likelihood model (ELM) and the theory of reasoned action (TRA) (Hoewe & Sherrick, 2015; Vafeiadis et al., 2019). This dyadic theoretical perspective with ELM (Petty & Cacioppo, 1986), which is a dual-process framework (Evans & Stanovich, 2013; Kahneman, 2011) and TRA (Fishbein & Ajzen, 1975; Fishbein, 2008), helps us develop the proposed framework which can serve as a useful guide for future research.

According to dual-process theory (Evans & Stanovich, 2013; Kahneman, 2011), intuition allows for rapid responses that are usually based on heuristic cues, whereas reflecting on the content with effort can override and correct intuitive responses (Bago et al., 2020). Extending this further, the elaboration likelihood model (ELM), a dual-process theoretical framework, goes by the basic premise that individuals process information through two different paths, which depend on the person's involvement level (Petty & Cacioppo, 1986). In this context, the central path would be one of thoughtful consideration of the information in the message, while the peripheral path would lead the individual to associate with the information cues in the stimulus to make a judgment. Leveraging this, we suggest that deliberation could lead to accurate deepfake detection, while the intuitive response or lack of sufficient involvement could lead to a belief in the deepfake as authentic.

According to TRA, an individual's behavior is determined by an intent to perform it, which in turn is determined through a combination of the attitude held towards that behavior and the subjective norms (M. Fishbein & Ajzen, 1975). We leverage TRA to propose that the belief in a deepfake forms the individual's intent to share the deepfake. This intention to share the deepfake is in turn determined through a combination of attitudes, which represent an evaluation of the deepfake, and subjective norms, which could, for example, be social influences to perform the act of deepfake dissemination. This combination of attitudes and subjective norms leads to the formation of intent to share the deepfake, which results in the eventual behavior of dissemination and further engagement with the deepfake.

Based on the dyadic theoretical perspective of ELM and TRA, we develop and discuss the five aspects of this framework that require scholarly attention, namely: (1) motivations for deepfake creation (2) deepfake viewer responses (3) antecedents of deepfake dissemination (4) deepfake sharing mechanisms and (5) outcomes of deepfake dissemination. The framework is grounded in gaps identified in prior deepfake literature and possible future research avenues.

7.1 Motivations for Deepfake Creation

The motivations represent the factors behind the creation of deepfakes. We argue that the intent to create a deepfake could be tied to derive a benefit or induce harm or neither in addressing a peripheral need. A benefit is tied to a positive intent of creating deepfakes which could range from personalized clinical therapy (e.g., reskilling a physically challenged patient by offering the experience through virtual reality and deepfake technology) (Wiederhold, 2021) or simulation modeling (e.g., allotransplantation simulations to computationally model and also demonstrate the post-transplantation appearance of organ recipients) (Crystal et al., 2020) or education (e.g., historical characters could be brought back to life through videos) (Chesney & Citron, 2018) or edutainment (e.g., use of deepfakes in museums for audience engagement and educational outreach) or advertising and marketing campaigns (e.g., personalized product placements) (Campbell et al., 2021; Kietzmann et al., 2021; Whittaker et al., 2021). Deepfakes combined with matching lip movements could also be used to translate speech to different languages for wider audience outreach (Vincent, 2021). Similar techniques, such as the Apple Facetime app's gaze correction, are meant to make video calls more intimate (Barkho, 2019), but this appearance of false eye contact might also be used to fake attention on conference calls.

An intent to inflict harm is tied to a negative motivation behind the creation of the deepfake. Such motivations could include an intent to disrupt democratic discourse or damage the online civic culture by creating an environment of indeterminacy (e.g., Chesney & Citron, 2018; Vaccari & Chadwick, 2020) or an intent to cause reputational harm, which could include examples such as damage through a non-consensual video (de Ruiter, 2021) or an act of cyberbullying (Langguth et al., 2021) or harm to business competitors through sabotage (Chesney & Citron, 2018). Deepfakes may also be created to carry out cybercrime by evading facial recognition and voice biometric systems (Ring, 2021). Negative intentions behind deepfake creation could further include the instigation of civil unrest by inflaming tensions or manufacturing crises (O'Donnell, 2021) or amplification of societal polarization by stoking cultural divisions (Sample et al., 2020). Further, the creation of deepfakes could stem from a negative intent to spread economic disinformation (O'Donnell, 2021) or to attempt an extortion bid, or commit online crimes against children through the creation of fake identities (Ring, 2021). In the case of aesthetic surgeries, while a positive motivation behind deepfake creation could be linked to the simulation of the post-surgical appearance, a negative motivation could be linked to creating false patient testimonials and outcomes (Crystal et al., 2020)

Lastly, the intent behind the creation of a deepfake may be neither tied to a benefit or harm but could just be for entertainment where deepfakes are created as a mechanism for online humor (Ayers, 2021) or satire (Westerlund, 2019) or to showcase their creative deepfake creation abilities in communities engaged in deepfake creation (de Seta, 2021). In this context, deepfakes hold immense potential as a tool for creative expression (Ham, 2021).

7.2 Viewing of Deepfakes

In the context of viewing deepfakes, we posit that the deepfake viewer may either choose to engage with it by believing the content to be true and possibly disseminating it further or may detect it to be false. In the context of deepfake detection, the growing sophistication of deepfake creation technology may impede the individual's ability to detect a deepfake (Chesney & Citron, 2018; Whittaker et al., 2021). Köbis et al. (2021) tested if increasing awareness of deepfakes and encouraging through financial means enhanced people's ability to detect them. Despite the lack of evidence to support improved detection abilities through these means, the experiments included videos with no political or ideological content, so we believe these elements may still have a role in accelerating deepfake identification, which may be confirmed in future studies. A slightly different approach would be to use information priming techniques to improve accuracy judgment and deepfake detection abilities. Such priming techniques have been found to improve fake news detection (Bryanov & Vziatysheva, 2021) and may be explored in the context of deepfakes. Similarly, enhancing media literacy and education may help individuals prepare better for deepfake attacks and reduce their belief in such synthetic media (Yankoski et al., 2021) while improving their accuracy judgment. Borrowing further from the literature on fake news, combination interventions involving the flagging of information as fake and raising awareness through training messages may yield positive results by improving the detection abilities of individuals (Moravec et al., 2020), and such combinations may serve as a good starting point for designing the appropriate interventions although the novelty of deepfakes driven by technological advancements makes it harder than usual for individuals to detect (Kietzmann et al., 2020).

Several factors could play a crucial role in establishing an individual's belief in deepfakes. The individual's biases may elicit emotional responses in the person, leading to the belief that the deepfake is true (Lange et al., 2011). Differences in cognitive ability may be able to anticipate how a deepfake will be countered, with previous research showing that persons with low cognitive abilities maintain their beliefs about fake information even when provided with the correct facts (De keersmaecker & Roets, 2017). In this context, the individual's cognitive biases could be at play wherein human cognition consistently produces representations that are systematically distorted in comparison to some feature of objective reality (Haselton et al., 2015).

We elaborate on three specific cognitive biases which could prompt individuals to establish their belief in a deepfake. First, anchoring bias could result from the individual believing the received deepfake to be a true source of information by drawing upon their own experience where most information they encounter is true (Brashier & Marsh, 2020). Such a bias could be another reason for people to believe a deepfake they are exposed to (Sample et al., 2020). Second, our proclivity to pay attention to news about our own or others' health and well-being may cause us to not only believe a deepfake to be true but also disseminate deepfakes on these issues, a phenomenon known as survival information bias. (Sample et al., 2020; Stubbersfield et al., 2015). Third, conspiracy theories about celebrities or political figures circulated by deepfakes may be a sort of social norm violation that causes us to pay attention to and believe it to be true and disseminate such material due to what is known as social information bias (Mesoudi et al., 2006; Stubbersfield et al., 2015).

As predicted by the feelings-as-information theory, emotions are another predictor of whether people believe or reject fake news (Schwarz, 2012). Emotionally biased reasoning could also lead the deepfake viewer to justify their belief instead of an accurate evaluation of the deepfake's authenticity (Sample et al., 2020; Yankoski et al., 2021). Similar to emotionally evocative narratives, a logical appeal could be established through the presentation of statistical manipulations with the purpose of data-driven persuasion, leading the individual to believe that the deepfake is legitimate (Sample et al., 2020). On social media, echo chambers (Boutyline & Willer, 2017) and filter bubbles (Holone, 2016) reinforce our cognitive processes, which, via the heuristic of confirmation bias, influence our belief in the deepfake as true (Shin & Lee, 2022). Although such exacerbation of ideological polarizations via the internet was previously considered difficult, especially given the high cost of tailoring stories to a particular point of view (Gentzkow & Shapiro, 2011), the ease with which deepfakes can be created with little expertise (Gosse & Burkell, 2020) threatens to upset the equilibrium and amplify ideological polarizations. Finally, information cascades could cause people to pay insufficient attention to the information they receive, assuming instead that the sender has determined the information to be genuine and end up passing it on (Chesney & Citron, 2018).

7.3 Antecedents of Deepfake Dissemination

Antecedents represent the factors that prompt individuals to share deepfakes online. In this context, an individual may genuinely believe the deepfake to be true and share it further. However, this is not always the case. Prior literature on misinformation reveals that there is a significant gap between what individuals think and what they would post on social media and that this is mostly due to inattention rather than deliberate sharing of misinformation (Pennycook & Rand, 2021). In a similar vein, social influences could be another antecedent for the dissemination of deepfakes. For example, the size of an individual's social network has been found to play a role in influencing the relationship between political interest and the inadvertent sharing of deepfakes (Ahmed, 2021c).

An individual's core characteristics and related differences between individuals are also key determinants of sharing deepfakes online. For example, several empirical studies have recognized the role of cognitive ability behind the sharing intention of deepfakes (Ahmed, 2021b, 2021a, 2022; Sample et al., 2020). The individual's cognitive biases could also fuel efforts aimed at deepfake dissemination, as discussed in the section on viewing of deepfakes (Chesney & Citron, 2018). Similarly, an individual's interest in a particular topic could lead to an inadvertent sharing of deepfakes. For example, Ahmed (2021a) discusses the role of political interest in the unintentional sharing of political deepfakes. The individual's awareness of deepfakes could also potentially lead individuals to detect a deepfaked video. Although one prior study by Köbis et al. (2021) did not find support for this ability in the detection of deepfakes, the authors suggest the possibility that awareness of deepfake artifacts could improve detection performance which could prevent them from sharing these deepfakes online.

7.4 Deepfake Sharing Mechanisms

False news is generally more novel than true news and has a higher likelihood of being shared, with virality being particularly strong in the context of political news (Vosoughi et al., 2018). Our review revealed both the novelty in deepfakes powered by technological advancements in artificial intelligence (Kietzmann et al., 2020) and the prominence of politics and political personalities, which have received considerable attention as a theme even among deepfakes, implying the importance of synthetic media as a medium for fake news. In this context, the sharing of deepfakes may be categorized into two groups. First, intentional sharing of deepfakes online could be executed through bot spreaders (Sample et al., 2020), deepfake-for-hire services (O'Donnell, 2021), or shared with a malicious intent linked to any of the motivations discussed in the previous section. Second, unintentional sharing of deepfakes can be linked to several factors ranging from innate or acquired capabilities of the individual to biases held by the person, which were discussed as antecedents in the previous section. Unintentional sharing may also be non-malicious as it may be linked to the distribution of deepfakes created for enjoyment or to demonstrate creativity, such as deepfake commercials (Campbell et al., 2021) and museum art (Mihailova, 2021).

7.5 Outcomes of Deepfake Dissemination

Outcomes refer to the consequences of deepfakes at various levels. In the context of our current framework, we leverage the levels as per business terminology to express the outcomes at the micro (individual), meso (organizational), and macro (institutional/market/societal) levels (Jeurissen, 1997). At the micro-level, the impact may be a positive one in granting autonomy to the individual. For example, a deepfake may allow an individual to have experiences that might not be possible in real life due to the challenges and dangers involved with such an act (Chesney & Citron, 2018) or may grant the individual personal expression through AI-enabled voice regeneration techniques (Langa, 2021) or allow the individual to speak in the language of choice and not bother any eye contact while relying on deepfake technology to perform the voice dubbing and gaze correction (Barkho, 2019; Vincent, 2021). In a similar vein, deepfakes aimed at edutainment may foster deeper emotional involvement in the viewer (e.g., deepfakes in art) (Mihailova, 2021). It may also enable healthcare professionals to have sensitive conversations during, for example, counseling sessions wherein the face could be anonymized without erasure of non-verbal cues (Ham, 2021). However, the negative outcomes outweigh the positives. Viewing a deepfake may bring about attitudinal changes towards the deepfaked person. For example, political deepfakes have been found to affect the attitude of the supporters towards the candidate (Dobber et al., 2021) and generate false perceptions (Vaccari & Chadwick, 2020), while viewing a deepfake of oneself could potentially have a similar effect on one's self-perception.

Deepfakes, in a related vein, may make the viewer doubt the validity of the video being viewed. For example, Vaccari and Chadwick (2020) found that while deepfakes did not mislead the viewer, it left them uncertain about the content being viewed. This skepticism about authenticity may not be an end in itself. It may lead to a situation of epistemological pluralism, in which an individual's acquisition of knowledge leads to a questioning of reality (Porter, 2020). Finally, the non-consensual nature of the video is a considerable threat to the deepfaked individual’s privacy (Chesney & Citron, 2018), and the deepfake target may suffer reputational damage (de Ruiter, 2021) or face financial losses when deepfake videos are used to evade biometric systems (Ring, 2021). The deepfaked individual may perhaps also face legal implications until the video's validity is verified if the victim is depicted in fake unlawful behavior. The prevalence of anonymity on the internet (Caldera, 2019) amplifies the legal wrath for the victim, who bears the burden of suing the deepfaked video's creator (O’Donnell, 2021), who may be difficult to track down while the lack of confinement of cyberspace to a particular state’s jurisdiction further complicates the victim’s legal options (Delfino, 2019)

Deepfakes may also have consequences at the meso and macro levels. At the meso level, they may aid organizations in enhancing audience engagement when used for educational purposes (e.g., recreating historical figures) or as works of art in museums or in advertising campaigns (Chesney & Citron, 2018; Kietzmann et al., 2021; Mihailova, 2021). Beyond engagement, deepfakes of self can be used to enhance trust in AI and related technological interventions, which are being increasingly explored in current societies towards enhancing people’s lives. For example, prior studies have been successful in changing people’s routine behaviors in a positive way when they were confronted by virtual versions of themselves showing the benefits of exercise and healthy eating habits (Bailenson & Segovia, 2009). In a similar vein, synthetic media used for simulation modeling helps preserve personal health data and also improves the decisioning abilities in the medical practice (Shin et al., 2018). However, deepfakes may also have negative consequences by impacting credibility at a professional level. For example, they may lead people to question journalistic credibility and create a sense of mistrust in the profession (Yadlin-Segal & Oppenheim, 2021).

In the context of deepfake related consequences at the macro level, they threaten to erode our trust in institutions and society. They may stoke cultural divisions in society and amplify societal polarization (Sample et al., 2020). Public institutions might find the democratic discourse being distorted with citizens’ views and voting preferences misaligned due to deepfakes (Sample et al., 2020). Deepfake videos could be used to spread fake economic information leading to disruption of economic activities (O’Donnell, 2021). For example, releasing a deepfake video of a company’s key personnel or investors might be used to manipulate the stock market and cause its stock prices to crash. Further, they may impact our testimonial practices, where recordings will cease to be treated as evidence in courts of law (Rini, 2020). Figure 2 provides a graphical representation of the theoretical framework with all the aspects discussed thus far in this section.

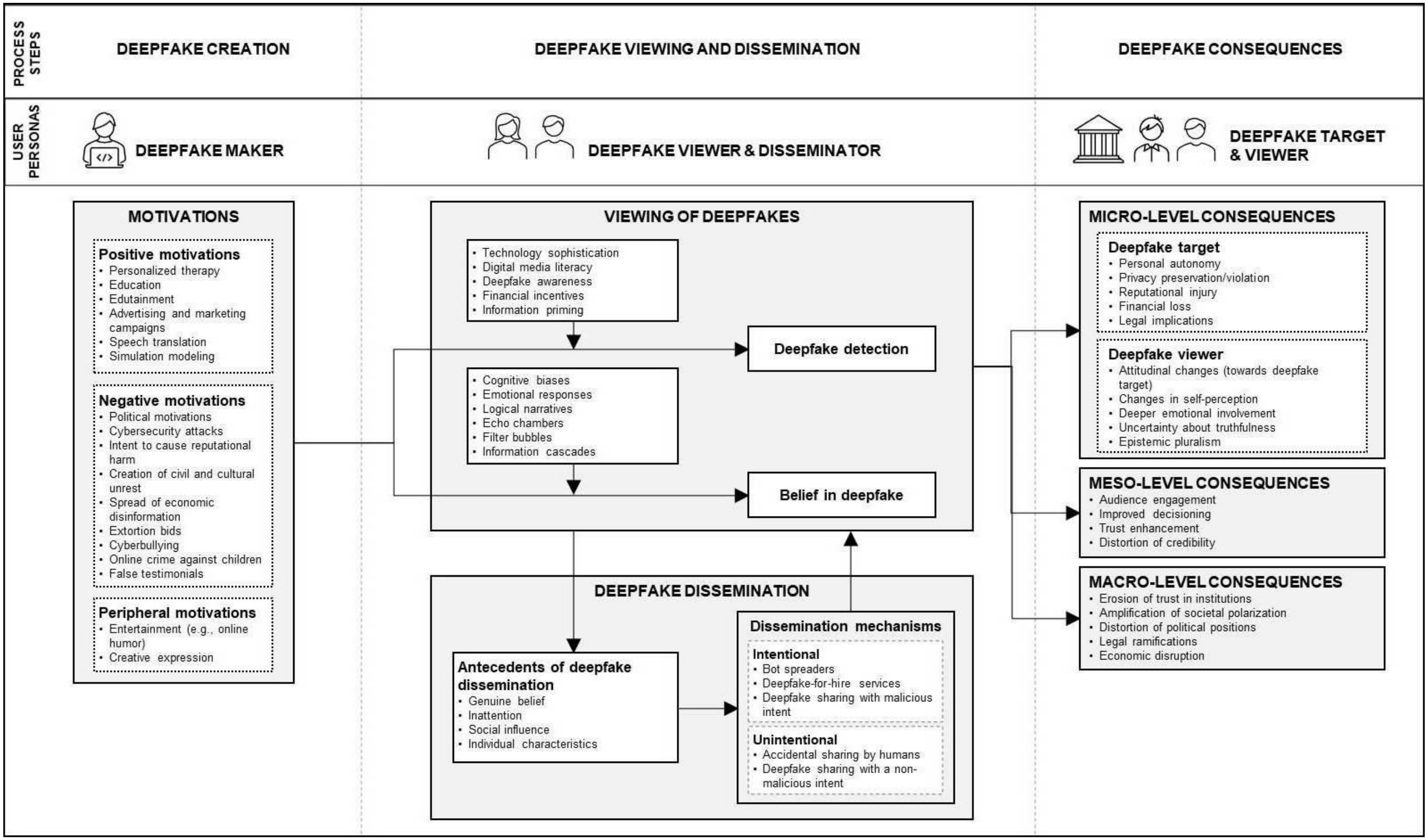

Figure 2. Theoretical Framework

8 Implications of the Study

Our research demonstrates that deepfakes are an emerging stream of research with the potential to have a wide range of applications. With this intent, this integrative review contributes significantly to our understanding of theoretical development opportunities in the field of deepfakes. In this section, we discuss the implications of our study to theory, the researchers in the domain of information systems, and policymakers.

8.1 Contributions to Research

The review makes three key contributions to theory. First, our study presents a systematically organized and contemporary structure of existing studies on deepfakes which aids in defining the current intellectual contours in this domain. Our contribution differentiates itself from prior reviews by offering a holistic discussion of existing literature within the domain of deepfakes. Second, our identification of gaps and limitations in the extant literature and theme-specific research questions offers a robust foundation for scholars interested in investigating the deepfake phenomenon. Third, our proposed framework structures the under-researched variables in the existing literature and presents avenues for future exploration. Future research based on this framework could provide novel insights into the phenomenon and expand our theoretical knowledge of this stream.

8.2 Implications for Information Systems Research

The review offers several new directions for researchers in the domain of information systems (IS). We elaborate on three key aspects below. First, deepfakes are generated using synthetic data, which is also used to power other emerging use cases such as the metaverse and virtual reality environments (Datagen, 2021; Mystakidis, 2022), enabling an immersive experience in which humans interact with deepfaked avatars. These open social environments have a disadvantage in that they may encourage individuals to engage in toxic antisocial conduct such as cyberbullying, trolling, and harassment (Chesney et al., 2009), as well as evoke traumatic experiences. This post-reality universe dominated by persuasive technologies creates an epistemic dilemma and opens up numerous pathways for IS researchers to examine how these technologies affect users' emotions, cognitive abilities, and behaviors. Second, deepfakes raise serious ethical concerns in the context of data ethics and privacy, while simultaneously opening up positive use cases for synthetic data in sectors such as healthcare. IS Researchers can make a substantial contribution towards informing the guiding principles for the design of systems that should embrace privacy and ethics from the start. Additionally, such designs enable IS researchers to ascertain any changes in individuals' privacy calculus. Third, the literature still lacks a clear understanding of the role of technology and its enabling characteristics in facilitating the spread of fake news online (Spiekermann et al., 2022), and deepfakes threaten to exacerbate these concerns by fueling additional disinformation and contributing to online radicalization. Platform providers and other third-party organizations are battling the negative consequences of this disinformation (e.g., Sharma et al., 2019). Implementing any actions to mitigate these negative consequences needs region-specific awareness that any ill-conceived countermeasures may jeopardize basic liberties such as freedom of expression (Monar, 2007; Nouri, 2019). In this precarious situation, IS researchers are well positioned to explore the connections between social and technological components in detail in order to inform platform players and regulatory agencies about the dangers of deepfakes and their associated dire effects.

Emerging interest in the metaverse merits special consideration in this context. This virtual shared space generated by the confluence of virtually enhanced physical and digital realities offers a profoundly engaging experience and possesses great value (Rimol, 2022). It offers new business models and economic opportunities, as well as the incentive for businesses to offer superior human and machine customer experiences (Furlonger et al., 2022). According to recent reports, however, what exists today are merely precursors to the metaverse, and the true prospects and adoption of the metaverse lie in accelerating enterprise innovation (Proulx et al., 2022). In this regard, the current simplistic landscapes and avatars on the metaverse have been accused of being unduly cartoonish and of undermining the gravity of its use in commercial context, so casting doubt on the technology's adoption among mainstream users (Truog, 2022). It is only when compelling experiences in the metaverse are firmly grounded in the physical world that individuals can function effectively and find fulfillment in the metaverse (Scott, 2022), allowing for the expansion of metaverse-enabled use cases. For instance, hybrid office models could evolve into virtual shared work spaces with the metaverse, caregivers could remotely interact with patients through the metaverse and tele-operated equipment, thereby enhancing the quality of healthcare service, hybrid instructors could teach in person or through a virtual presence in the metaverse (Linder et al., 2022), manufacturing plants could run virtual simulations of production cycles in the metaverse, and ecommerce players could enable customers to interact virtually with their products in the metaverse (Adair, 2022). In view of these compelling use cases, firms are aiming to evolve from simplistic to hyper-realistic metaverse representations (Scott, 2022), and deepfakes have the potential to power these transitions.

In what is referred to as the fashion metaverse, for instance, we are already experiencing an increase in deepfake apparel, in which influencers promote clothes without actually wearing it in real life, but rather by wearing digital versions of it (Darko, 2021; McDowell, 2021). In the realm of the Metaverse, where everything is synthetic, as we spend more time in this synthetic space and start to give a lot more attention to how we look and sound, the current low-polygon cartoon-like virtual avatars will evolve and deepfakes will play a role in this evolution (Gamma Telecom, 2022). Furthermore, as businesses seek to provide individuals with the technology to generate hyper-realistic synthetic avatars, they are also exploring options for users to securely store their avatars, such as non-fungible tokens, so that users retain ownership of their photos and the biometric data used to create them (Palmer, 2022). We contend that these metaverse-enabled uses of deepfake technology for legitimate and ethical purposes may eventually displace the malicious use of the technology, bring in more responsible synthetic content, and inspire creative expressions, while these technological transitions offer a wide range of research opportunities for IS researchers.

However, if the malicious usage of deepfakes persists, interactive deepfakes could become ubiquitous in the metaverse, deceiving metaverse users into believing they are dealing with the real person rather than a mere simulation (Mostert & Cruz, 2022). If our reliance on the metaverse advances to the point where it is used to authenticate individuals in the real world, our existence as disembodied participants with a database entry and a security token in cyberspace could be gravely at risk if the security of the data associated with us is compromised. Further, when we traverse to a world where our images and voices are synthetic and beyond images, even our voices can be deepfaked, and we may be required to prove the ownership of our own voices using options similar to non-fungible tokens if we wish to use our voices to authenticate banking transactions, attend meetings and so on in the metaverse (Gamma Telecom, 2022). Advanced encryption solutions that are practically inaccessible to hackers today may be an option for authentication, although the viability of such alternatives for day-to-day interactions in the metaverse may require additional investigation (Woodie, 2022). These issues posed by deepfakes present numerous avenues for IS researchers to explore in the context of cybersecurity and data protection.

8.3 Implications for Policy Makers

This review has significant implications for policymakers. First, as the technology associated with deepfakes matures, policymakers must recognize that these tools will become widespread. The current regulations surrounding deepfakes and the challenges in imposing those regulations, which are discussed in this review, provide a crucial perspective for policymakers to design rules regarding the fabrication and distribution of deepfakes, although the review suggests that there may be no perfect solution to stop the threats posed by deepfakes, with constant calibration of policies required to mitigate their negative effects. The inadequacies of present civil and criminal liabilities must compel officials to seek legislation intended at bolstering deepfake related redressal mechanisms. Second, the positive use cases of deepfakes outlined in this review offer direction to policymakers who must ensure that the implementation of regulations does not impede the useful applications of deepfakes. It is crucial for policymakers to guarantee that regulation proposals include some exemptions for the advancement of the technology's beneficial use cases. Third, the review emphasizes that the deepfake technological environment is continually evolving and that detection algorithms must keep pace with deepfake generation techniques. In this context, the review encourages technology policymakers to ensure that adequate encouragement and incentives are in place to drive investments in the development and adoption of technical solutions aimed at deepfake detection, while the review also equips technology platform players with the insights to develop guidelines and standards for platform governance.

9 Limitations

Despite the contributions presented in this work, it is not without limitations. Firstly, our research is based on secondary data gathered from academic research on deepfakes. While we tried to include grey literature in our review, the number of sources is limited, so future research can continue this effort and gain a more comprehensive understanding of the subject. In a similar vein, we encountered a few studies which were behind a paywall, and the full text could not be obtained. However, we have attempted to convey the key essence of such articles as well in our review. Second, in order to assist our integrative review, we used a set of relevant and contextual keywords and popular databases to find relevant publications. On this basis, we position our work as fully original, extensive, and critical in character, based on our keyword strategy, but the evolving nature of the phenomenon and resolution of definitional issues surrounding it may expand its horizons and incorporate more works in future reviews. In that context, our work may serve as a platform for further research into the emerging topic of deepfakes.

10 Conclusion

In this paper, we conducted an integrative review of literature on deepfakes. Integrative reviews help examine a topic in depth and synthesize the extant literature to generate new perspectives and frameworks (Torraco, 2005; Webster & Watson, 2002). This paper represents one of the first attempts to provide an extensive and critical review on deepfakes, unlike past reviews, which have been narrow in focus by limiting to news articles (Westerlund, 2019) or creation and combat methods (Albahar & Almalki, 2019; Botha & Pieterse, 2020; Verdoliva, 2020) or limited to particular domains (Godulla et al., 2021). Our study has been able to evaluate the existing literature and highlight areas in deepfake literature which warrant attention, while the framework we put forth should help structure future research on deepfakes in a methodological manner. We hope that this review can stimulate researchers across domains to collaborate with each other and engage in understanding the multi-domain nature of deepfakes as a phenomenon and counter the effects of this increasingly important technological advancement while generating value from it.

Acknowledgments

Satish Krishnan thanks the Indian Institute of Management Kozhikode’s Chair Associate Professorship for supporting this research.

References

Adair, M. (2022). Five Industries That Will Be Transformed by The Metaverse. Retrieved July 23, 2022, from Forbes website: https://www.forbes.com/sites/forbestechcouncil/2022/03/22/five-industries-that-will-be-transformed-by-the-metaverse/

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7.

Ahmed, S. (2021a). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences*, *182*, 111074.

Ahmed, S. (2021b). Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media & Society*, 14614448211019198.

Ahmed, S. (2021c). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, *57*, 101508.

Ahmed, S. (2022). Disinformation Sharing Thrives with Fear of Missing Out among Low Cognitive News Users: A Cross-national Examination of Intentional Sharing of Deep Fakes. *Journal of Broadcasting & Electronic Media*, *0*, 1–21.

Albahar, M., & Almalki, J. (2019). DEEPFAKES: THREATS AND COUNTERMEASURES SYSTEMATIC REVIEW. . . *Vol.*, 9.

Allison, T. (2021). Race and the digital face: Facial (mis)recognition in Gemini Man. *Convergence*, *27*, 999–1017.

Avast. (2020, December 8). Avast Cybersecurity Experts Predict Covid-19 Vaccination Scams and Deepfake Disinformation Campaigns for 2021. Retrieved December 31, 2021, from Avast Cybersecurity Experts Predict Covid-19 Vaccination Scams and Deepfake Disinformation Campaigns for 2021 website: https://press.avast.com/avast-cybersecurity-experts-predict-covid-19-vaccination-scams-and-deepfake-disinformation-campaigns-for-2021

Ayers, D. (2021). The limits of transactional identity: Whiteness and embodiment in digital facial replacement. *Convergence*, *27*, 1018–1037.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology. General*, *149*, 1608–1613.

Bailenson, J., & Segovia, K. (2009). *Virtual Doppelgangers: Psychological Effects of Avatars Who Ignore Their Owners*.

Barari, S., Lucas, C., & Munger, K. (2021). *Political Deepfakes Are As Credible As Other Fake Media And (Sometimes) Real Media* [Preprint]. Open Science Framework.

Barkho, G. (2019, July 3). Finally, You Can Make Eye Contact During FaceTime Calls. Retrieved December 24, 2021, from Observer website: https://observer.com/2019/07/facetime-eye-contact-ios-13-update/

Bazarkina, D. Yu., & Pashentsev, Y. N. (2019). Artificial Intelligence and New Threats to International Psychological Security. Retrieved November 24, 2021, from Russia in Global Affairs website: https://eng.globalaffairs.ru/articles/artificial-intelligence-and-new-threats-to-international-psychological-security/

Bloom, S. (2014). No vengeance for ‘revenge porn’ victims: Unraveling why this latest female-centric, intimate-partner offense is still legal, and why we should criminalize it. *Fordham Urban Law Journal*, *42*, 233–289. Scopus.

Bode, L. (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. *Convergence*, *27*, 919–934.

Borges-Tiago, T., Tiago, F., Silva, O., Guaita Martínez, J. M., & Botella-Carrubi, D. (2020). Online users’ attitudes toward fake news: Implications for brand management. *Psychology & Marketing*, *37*, 1171–1184.

Botha, J., & Pieterse, H. (2020). *Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security*.

Boutyline, A., & Willer, R. (2017). The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks. *Political Psychology*, *38*, 551–569.

Brashier, N. M., & Marsh, E. J. (2020). Judging Truth. *Annual Review of Psychology*, *71*, 499–515.

Breen, D. C. (2021). Silent no more: How deepfakes will force courts to reconsider video admission standards. *Journal of High Technology Law*, *21*, 122–161.

Brennen, J. S., Simon, F. M., & Nielsen, R. K. (2021). Beyond (Mis)Representation: Visuals in COVID-19 Misinformation. *The International Journal of Press/Politics*, *26*, 277–299.

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. *PLOS ONE*, *16*, e0253717.

Caldera, E. (2019). “Reject the Evidence of Your Eyes and Ears”: Deepfakes and the Law of Virtual Replicants. *Seton Hall Law Review*, *50*. Retrieved from https://scholarship.shu.edu/shlr/vol50/iss1/5

Campbell, C., Plangger, K., Sands, S., & Kietzmann, J. (2021). Preparing for an Era of Deepfakes and AI-Generated Ads: A Framework for Understanding Responses to Manipulated Advertising. *Journal of Advertising*, *0*, 1–17.

Carvajal, L., & Iliadis, A. (2020). DEEPFAKES: A PRELIMINARY SYSTEMATIC REVIEW OF THE LITERATURE. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2020i0.11190

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*, 493–497.

Chesney, R., & Citron, D. (2019). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. *Foreign Affairs*, *98*, 147.

Chesney, R., & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security* (SSRN Scholarly Paper No. ID 3213954). Rochester, NY: Social Science Research Network.

Chesney, T., Coyne, I., Logan, B., & Madden, N. (2009). Griefing in virtual worlds: Causes, casualties and coping strategies. *Information Systems Journal*, *19*, 525–548.

Christopher, N. (2020, February 18). Deepfakes by BJP in Indian Delhi Election Campaign. Retrieved December 11, 2021, from https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

Cole, S. (2017, December 12). AI-Assisted Fake Porn Is Here and We're All Fucked. Retrieved November 5, 2021, from Vice website: https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Colom, R., Abad, F. J., Rebollo, I., & Chun Shih, P. (2005). Memory span and general intelligence: A latent-variable approach. *Intelligence*, *33*, 623–642.

Conte, P. (2019). MOCKUMENTALITY: FROM HYPERFACES TO DEEPFAKES. *World Literature Studies*, *11*. Retrieved from http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.cejsh-d4691768-ac82-4539-a63e-dce56ff09c27

Conwell, J. (2020, January 29). 2020: A year of deepfakes and deep deception. Retrieved December 19, 2021, from Help Net Security website: https://www.helpnetsecurity.com/2020/01/29/deepfake-technology/

Crystal, D. T., Cuccolo, N. G., Ibrahim, A. M. S., Furnas, H., & Lin, S. J. (2020). Photographic and Video Deepfakes Have Arrived: How Machine Learning May Influence Plastic Surgery. *Plastic and Reconstructive Surgery*, *145*, 1079–1086.

Cybenko, G., Giani, A., & Thompson, P. (2002). Cognitive hacking: A battle for the mind. *Computer*, *35*, 50–56.

Darko. (2021, November 12). The rise of deepfake clothing. Retrieved July 23, 2022, from Indie Hackers website: https://www.indiehackers.com/post/the-rise-of-deepfake-clothing-4519753717?utm_source=indie-hackers-emails&utm_campaign=ih-newsletter&utm_medium=email

Dasilva, J. P., Ayerdi, K. M., & Galdospin, T. M. (2021). Deepfakes on Twitter: Which Actors Control Their Spread? *Media and Communication*, *9*, 301–312.

Datagen. (2021, November 30). The Metaverse and AI Edge Cases Will Drive Synthetic Data Boom: Top Predictions for 2022 by Synthetic Data Innovator Datagen. Retrieved April 16, 2022, from GlobeNewswire News Room website: https://www.globenewswire.com/news-release/2021/11/30/2343152/0/en/The-Metaverse-and-AI-Edge-Cases-Will-Drive-Synthetic-Data-Boom-Top-Predictions-for-2022-by-Synthetic-Data-Innovator-Datagen.html

David, R. J., & Han, S.-K. (2004). A Systematic Assessment of the Empirical Support for Transaction Cost Economics. *Strategic Management Journal*, *25*, 39–58.

De keersmaecker, J., & Roets, A. (2017). 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, *65*, 107–110.

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

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

Deep Empathy. (2017). Deep Empathy by MIT Media Lab. Retrieved December 11, 2021, from http://deepempathy.mit.edu

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

Deloitte. (2019, February 20). Artificial Intelligence Ethics | Deloitte Jordan | Press release. Retrieved April 16, 2022, from Deloitte website: https://www2.deloitte.com/jo/en/pages/about-deloitte/articles/artificial-intelligence-ethics.html

Denham, H. (2020). Another fake video of Pelosi goes viral on Facebook. *The Washington Post*. Retrieved from https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/

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

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

Domenico, G. D., Sit, J., Ishizaka, A., & Nunan, D. (2021). Fake news, social media and marketing: A systematic review. *Journal of Business Research*, *124*, 329–341.

Dowdeswell, T. L., & Goltz, N. (2020). The clash of empires: Regulating technological threats to civil society. *Information & Communications Technology Law*, *29*, 194–217.

European Commission. (2018). JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Report on the implementation of the Action Plan Against Disinformation, JOIN/2019/12 final [Website].

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*, 223–241.

Facebook. (2019). Deepfake Detection Challenge. Retrieved November 17, 2021, from https://deepfakedetectionchallenge.ai/

Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Green supply chain management: A review and bibliometric analysis. *International Journal of Production Economics*, *162*, 101–114.

Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00419-2

Fenn, E., Ramsay, N., Kantner, J., Pezdek, K., & Abed, E. (2019). Nonprobative Photos Increase Truth, Like, and Share Judgments in a Simulated Social Media Environment. *Journal of Applied Research in Memory and Cognition*, *8*, 131–138.

Fido, D., Rao, J., & Harper, C. A. (2022). Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography. *Computers in Human Behavior*, *129*, 107141.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behaviour: An introduction to theory and research* (Vol. 27).

Fishbein, Martin. (2008). Reasoned Action, Theory of. In *The International Encyclopedia of Communication*. John Wiley & Sons, Ltd.

Franks, M. A., & Waldman, A. E. (2018). Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions. *Maryland Law Review*, *78*, 892.

French, A., Risius, M., & Shim, J. P. (2020). The Interaction of Virtual Reality, Blockchain, and 5G New Radio: Disrupting Business and Society. *Communications of the Association for Information Systems*, *46*. https://doi.org/10.17705/1CAIS.04625

French, A., Shim, J. P., Risius, M., Larsen, K., & Jain, H. (2021). The 4th Industrial Revolution Powered by the Integration of AI, Blockchain, and 5G. *Communications of the Association for Information Systems*, *49*. https://doi.org/10.17705/1CAIS.04910

Furlonger, D., Uzureau, C., & Kandaswamy, R. (2022, June 24). Wondering how to capture opportunity in the metaverse? Retrieved July 22, 2022, from Gartner website: https://www.gartner.com/en/articles/how-to-capture-opportunity-in-the-metaverse

Gamma Telecom. (2022). Future technologies: The Metaverse, NFTs and Deepfakes. Retrieved July 23, 2022, from Gamma Telecom website: https://www.gamma.co.uk/resources/unify/future-technologies-we-can-not-put-off-until-tomorrow/

Gentzkow, M., & Shapiro, J. M. (2011). Ideological Segregation Online and Offline *. *The Quarterly Journal of Economics*, *126*, 1799–1839.

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

Gooding, M. (2021, September 24). Synthetic data may not be AI’s privacy silver bullet. Retrieved April 16, 2022, from Tech Monitor website: https://techmonitor.ai/technology/ai-and-automation/synthetic-data-may-not-be-ais-privacy-silver-bullet

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

Graber, D. A. (1990). Seeing is remembering: How visuals contribute to learning from television news. *Journal of Communication*, *40*, 134–155.

Gregory, S. (2021). Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*, 14648849211060644.

Griffin, M. (2019, August 2). Edtech company Udacity uses deepfake tech to create educational videos automatically. Retrieved April 15, 2022, from By Futurist and Virtual Keynote Speaker Matthew Griffin website: https://www.fanaticalfuturist.com/2019/08/edtech-company-udacity-uses-deepfake-tech-to-create-educational-videos-automatically/

Güera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*. https://doi.org/10.1126/sciadv.aau4586

Ham, B. (2021, December 16). Characters for good, created by artificial intelligence. Retrieved December 22, 2021, from MIT Media Lab website: https://www.media.mit.edu/articles/characters-for-good-created-by-artificial-intelligence/

Hancock, J. T. (2009, February 12). Digital deception. https://doi.org/10.1093/oxfordhb/9780199561803.013.0019

Hancock, J. T., & Bailenson, J. N. (2021). The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*, 149–152.

Harari, Y. N. (2018, September 7). Are we living in a post-truth era? Yes, but that's because we're a post-truth species. Retrieved December 23, 2021, from Ideas.ted.com website: https://ideas.ted.com/are-we-living-in-a-post-truth-era-yes-but-thats-because-were-a-post-truth-species/

Harris, D. (2019). Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke Law & Technology Review*, *17*, 99–127.

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

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The Evolution of Cognitive Bias. In *The Handbook of Evolutionary Psychology* (pp. 1–20). John Wiley & Sons, Ltd.

Hight, C. (2021). Deepfakes and documentary practice in an age of misinformation. *Continuum*, *0*, 1–18.

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

Hoewe, J., & Sherrick, B. (2015). Using the Theory of Reasoned Action and Structural Equation Modeling to Study the Influence of News Media in an Experimental Context. *Atlantic Journal of Communication*, *23*, 237–253.

Holliday, C. (2021). Rewriting the stars: Surface tensions and gender troubles in the online media production of digital deepfakes. *Convergence*, *27*, 899–918.

Holone, H. (2016). The filter bubble and its effect on online personal health information. *Croatian Medical Journal*, *57*, 298–301.

Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (2021). Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *Cyberpsychology, Behavior, and Social Networking*, *24*, 188–193.

iperov. (2021). *Iperov/DeepFaceLive* [Python]. Retrieved from https://github.com/iperov/DeepFaceLive (Original work published 2020)

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious Influences of Memory for a Prior Event. *Personality and Social Psychology Bulletin*, *13*, 314–336.

Jankowicz, N. (2020). *How to lose the information war: Russia, fake news, and the future of conflict*. I.B. Tauris.

Jee, C. (2019, February). An Indian politician is using deepfake technology to win new voters. Retrieved December 11, 2021, from MIT Technology Review website: https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/

Jeurissen, R. (1997). Integrating Micro, Meso and Macro Levels in Business Ethics. *Ethical Perspectives*, *4*, 246–254.

Johnson, D., & Diakopoulos, N. (2021, March). What To Do About Deepfakes. Retrieved November 25, 2021, from Communications of the ACM website: https://cacm.acm.org/magazines/2021/3/250701-what-to-do-about-deepfakes/fulltext

Kahneman, D. (2011). *Thinking, fast and slow*.

Kalpokas, I., & Kalpokiene, J. (2022a). Fake News: Exploring the Backdrop. In I. Kalpokas & J. Kalpokiene (Eds.), *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation* (pp. 7–17). Cham: Springer International Publishing.

Kalpokas, I., & Kalpokiene, J. (2022b). From GANs to Deepfakes: Getting the Characteristics Right. In I. Kalpokas & J. Kalpokiene (Eds.), *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation* (pp. 29–39). Cham: Springer International Publishing.

Kalpokas, I., & Kalpokiene, J. (2022c). On Alarmism: Between Infodemic and Epistemic Anarchy. In I. Kalpokas & J. Kalpokiene (Eds.), *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation* (pp. 41–53). Cham: Springer International Publishing.

Kant, I., & Beck, L. W. (1959). *Foundations of the metaphysics of morals, and What is enlightenment?* New York: Liberal Arts Press.

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics*, *8*, 81–108.

Khan, A., Krishnan, S., & Dhir, A. (2021). Electronic government and corruption: Systematic literature review, framework, and agenda for future research. *Technological Forecasting and Social Change*, *167*, 120737.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*, 135–146.

Kietzmann, J., Mills, A. J., & Plangger, K. (2021). Deepfakes: Perspectives on the future “reality” of advertising and branding. *International Journal of Advertising*, *40*, 473–485.

Kikerpill, K. (2020). Choose Your Stars and Studs: The Rise of Deepfake Designer Porn. *Porn Studies*, *7*. https://doi.org/10.1080/23268743.2020.1765851

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *IScience*, *24*, 103364.

Langa, J. (2021). DEEPFAKES, REAL CONSEQUENCES: CRAFTING LEGISLATION TO COMBAT THREATS POSED BY DEEPFAKES. *BOSTON UNIVERSITY LAW REVIEW*, *101*, 41.

Lange, P. A. M. V., Kruglanski, A. W., & Higgins, E. T. (2011). *Handbook of Theories of Social Psychology: Collection: Volumes 1 & 2*. SAGE.

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don’t Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication*, *6*, 26.

Leong, J. S. L. (Joanne S. L. (2021). *Investigating the use of synthetic media and real-time virtual camera filters for supporting communication and creativity* (Thesis, Massachusetts Institute of Technology). Massachusetts Institute of Technology.

Linder, P., Murphy, T., Nezami, Y., & Dohler, M. (2022, June 30). 12 metaverse use cases in coming years. Retrieved July 23, 2022, from Ericsson website: https://www.ericsson.com/en/blog/2022/7/10-metaverse-use-cases

Maddalena, G., & Gili, G. (2020). *The history and theory of post-truth communication*. Palgrave Macmillan.

Maddocks, S. (2020). ‘A Deepfake Porn Plot Intended to Silence Me’: Exploring continuities between pornographic and ‘political’ deep fakes. *Porn Studies*, *7*, 415–423.

Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*, *23*, 255–262.

Margry, K. (1992). ‘Theresienstadt’ (1944–1945): The Nazi propaganda film depicting the concentration camp as paradise. *Historical Journal of Film, Radio and Television*, *12*, 145–162.

Mazzoni, G., & Kirsch, I. (2002). Autobiographical memories and beliefs: A preliminary metacognitive model. In *Applied metacognition* (pp. 121–145). New York, NY, US: Cambridge University Press.

McDowell, M. (2021, September 21). Influencers are wearing digital versions of physical clothes now. Retrieved July 23, 2022, from Vogue Business website:

https://www.voguebusiness.com/technology/influencers-are-wearing-digital-versions-of-physical-clothes-now

McIntyre, L. (2018). *Post-Truth*. Cambridge, MA, USA: MIT Press.

Merriam-Webster. (n.d.). Fake Definition & Meaning—Merriam-Webster. Retrieved December 12, 2021, from https://www.merriam-webster.com/dictionary/fake

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

Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology*, *97*, 405–423.

Metzger, M. J., & Flanagin, A. J. (Eds.). (2007). *Digital Media, Youth, and Credibility*. The MIT Press.

Mihailova, M. (2021). To Dally with Dalí: Deepfake (Inter)faces in the Art Museum. *Convergence*, *27*, 882–898.

Minola, T., Criaco, G., & Cassia, L. (2014). Are youth really different? New beliefs for old practices in entrepreneurship. *International Journal of Entrepreneurship and Innovation Management*, *18*, 233.

Mohammed, F., & Salam, A. F. (2021). Me and the Other Not Me - Deepfake as Digitally Constructed Alternate Deceptive Identity: Loss of Control Over One’s Identity and Consequences. *ICIS 2021 Proceedings*. Retrieved from https://aisel.aisnet.org/icis2021/soc_impact/soc_impact/4

Monar, J. (2007). Common Threat and Common Response? The European Union’s Counter-Terrorism Strategy and its Problems. *Government and Opposition*, *42*, 292–313.

Moravec, P. L., Kim, A., & Dennis, A. R. (2020). Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on social media. *Information Systems Research*. (world). https://doi.org/10.1287/isre.2020.0927

Mostert, F., & Cruz, S. (2022, February 4). *Image Rights in the Digital Universe* [SSRN Scholarly Paper]. Rochester, NY. https://doi.org/10.2139/ssrn.4026437

Mullen, M. (2022). A New Reality: Deepfake Technology and the World Around Us. *Mitchell Hamline Law Review*, *48*. Retrieved from https://open.mitchellhamline.edu/mhlr/vol48/iss1/5

Mystakidis, S. (2022). Metaverse. *Encyclopedia*, *2*, 486–497.

Nast, C. (2021, January 11). How deepfakes could change fashion advertising. Retrieved April 16, 2022, from Vogue Business website: https://www.voguebusiness.com/companies/how-deepfakes-could-change-fashion-advertising-influencer-marketing

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, C. M., Nguyen, D., Nguyen, D. T., & Nahavandi, S. (2021). Deep Learning for Deepfakes Creation and Detection: A Survey. *ArXiv:1909.11573 [Cs, Eess]*. Retrieved from http://arxiv.org/abs/1909.11573

Nouri, L. (2019). Following the Whack-a-Mole: Britain First’s Visual Strategy from Facebook to Gab.

O’Donnell, N. (2021). Have We No Decency? Retrieved November 27, 2021, from Illinois Law Review website: https://www.illinoislawreview.org/print/vol-2021-no-2/have-we-no-decency/

Oh, S. Y., Bailenson, J., Krämer, N., & Li, B. (2016). Let the Avatar Brighten Your Smile: Effects of Enhancing Facial Expressions in Virtual Environments. *PLOS ONE*, *11*, e0161794.

Öhman, C. (2020). Introducing the pervert’s dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, *22*, 133–140.

Oxford Analytica. (2022). Beijing acts to control AI-generated content. *Emerald Expert Briefings*, *oxan-db*. https://doi.org/10.1108/OXAN-DB267101

Palmer, M. (2022, April 18). The deepfake dangers lurking in the metaverse. Retrieved July 22, 2022, from Sifted website: https://sifted.eu/articles/deepfake-dangers-metaverse/

Paris, B. (2021). Configuring Fakes: Digitized Bodies, the Politics of Evidence, and Agency. *Social Media + Society*, *7*, 20563051211062920.

Pasquarelli, W. (2019, August 6). Towards Synthetic Reality: When DeepFakes meet AR/VR. Retrieved December 22, 2021, from Oxford Insights website: https://www.oxfordinsights.com/insights/2019/8/6/towards-synthetic-reality-when-deepfakes-meet-arvr

Paterson, T., & Hanley, L. (2020). Political warfare in the digital age: Cyber subversion, information operations and 'deep fakes.' *Australian Journal of International Affairs*, *74*, 439–454.

Pavis, M. (2021). Rebalancing our regulatory response to Deepfakes with performers' rights. *Convergence*, *27*, 974–998.

Pazzanese, C. (2020, October 26). Ethical concerns mount as AI takes bigger decision-making role. Retrieved April 16, 2022, from Harvard Gazette website: https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*, 388–402.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on social media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*, 770–780.

Petkauskas, V. (2021, May 3). Report: Number of deepfakes double every six months. Retrieved December 1, 2021, from CyberNews website: https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months/

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. In R. E. Petty & J. T. Cacioppo (Eds.), *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (pp. 1–24). New York, NY: Springer.

Pietro, R. D., Raponi, S., Caprolu, M., & Cresci, S. (2020). *New Dimensions of Information Warfare*. Springer Nature.

Popova, M. (2019). Reading out of context: Pornographic deepfakes, celebrity and intimacy. *Porn Studies*, *7*. https://doi.org/10.1080/23268743.2019.1675090

Porter, A. (2020). Bioethics in the Ruins. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *45*, 259–276.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, *83*, 157–171.

Prior, M. (2014). Visual Political Knowledge: A Different Road to Competence? *The Journal of Politics*, *76*, 41–57.

Proulx, M., Ask, J., Bennett, M., Gownder, J. P., & Truog, D. (2022, March 29). There Is No Metaverse Today, but Be Prepared. Retrieved July 23, 2022, from Forrester website: https://www.forrester.com/blogs/there-is-no-metaverse-today-but-be-prepared/

Ray, A. (2021). Disinformation, Deepfakes and Democracies: The Need for Legislative Reform. *UNSW Law Journal*. Retrieved from https://www.unswlawjournal.unsw.edu.au/article/disinformation-deepfakes-and-democracies-the-need-for-legislative-reform/

Richardson, S., Petter, S., & Carter, M. (2021). Five Ethical Issues in the Big Data Analytics Age. *Communications of the Association for Information Systems*, *49*. https://doi.org/10.17705/1CAIS.04918

Rimol, M. (2022, February 7). Gartner Predicts 25% of People Will Spend At Least One Hour Per Day in the Metaverse by 2026. Retrieved July 22, 2022, from Gartner website: https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026

Ring, T. (2021). Europol: The AI hacker threat to biometrics. *Biometric Technology Today*, *2021*, 9–11.

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers' Imprint*, *20*, 1–16.

Rupapara, V., Rustam, F., Amaar, A., Washington, P. B., Lee, E., & Ashraf, I. (2021). Deepfake tweets classification using stacked Bi-LSTM and words embedding. *PeerJ Computer Science*, *7*, e745.

Sample, C., Jensen, M. J., Scott, K., McAlaney, J., Fitchpatrick, S., Brockinton, A., … Ormrod, A. (2020). Interdisciplinary Lessons Learned While Researching Fake News. *Frontiers in Psychology*, *11*, 2947.

Schick, N. (2020). *Deep Fakes and the Infocalypse: What You Urgently Need To Know*. Monoray.

Schilpzand, P., De Pater, I. E., & Erez, A. (2016). Workplace incivility: A review of the literature and agenda for future research. *Journal of Organizational Behavior*, *37*, S57–S88.

Schultze, U., Hiltz, S., Nardi, B., Rennecker, J., & Stucky, S. (2008). Using Synthetic Worlds for Work and Learning. *Communications of the Association for Information Systems*, *22*. https://doi.org/10.17705/1CAIS.02219

Schwarz, N. (2012). Feelings-as-information theory. In *Handbook of theories of social psychology, Vol. 1* (pp. 289–308). Thousand Oaks, CA: Sage Publications Ltd.

Scott, K. (2022, June 13). Why the Metaverse Will Be Hyperreal. Retrieved July 22, 2022, from Https://metaphysic.ai/ website: https://metaphysic.ai/why-metaverse-hyperreal/, https://metaphysic.ai/why-metaverse-hyperreal/

Scully-Russ, E., & Torraco, R. (2020). The Changing Nature and Organization of Work: An Integrative Review of the Literature. *Human Resource Development Review*, *19*, 66–93.

Sensity. (2019). *The State of Deepfakes 2019 Landscape, Threats, and Impact*. Retrieved from https://share.hsforms.com/1cg_h2aPnRrufZeN8HDjWPw3hq83

Sensity. (2020). *The State of Deepfakes 2020: Updates on Statistics and Trends*. Sensity.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). *Combating Fake News: A Survey on Identification and Mitigation Techniques*. https://doi.org/10.48550/arXiv.1901.06437

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., … Michalski, M. (2018). Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In A. Gooya, O. Goksel, I. Oguz, & N. Burgos (Eds.), *Simulation and Synthesis in Medical Imaging* (pp. 1–11). Cham: Springer International Publishing.

Shin, S. Y., & Lee, J. (2022). The Effect of Deepfake Video on News Credibility and Corrective Influence of Cost-Based Knowledge about Deepfakes. *Digital Journalism*, *10*, 412–432.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London; New York: Wiley.

Shu, K., Wang, S., & Liu, H. (2018). Understanding User Profiles on social media for Fake News Detection. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 430–435.

Sigerson, L., & Cheng, C. (2018). Scales for measuring user engagement with social network sites: A systematic review of psychometric properties. *Computers in Human Behavior*, *83*, 87–105.

Simonite, T. (2020, November 16). What Happened to the Deepfake Threat to the Election? *Wired*. Retrieved from https://www.wired.com/story/what-happened-deepfake-threat-election/

Somers, M. (2020, July 21). Deepfakes, explained. Retrieved December 8, 2021, from MIT Sloan website: https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Spiekermann, S., Krasnova, H., Hinz, O., Baumann, A., Benlian, A., Gimpel, H., … Trenz, M. (2022). Values and Ethics in Information Systems. *Business & Information Systems Engineering*, *64*, 247–264.

Stadler, J. (2019). Synthetic Beings and Synthespian Ethics. *Projections*, *13*, 123–141.

Staff, B. (2019, September 6). 10 Best Deepfake Apps and Websites You Can Try for Fun. Retrieved December 25, 2021, from Beebom website: https://beebom.com/best-deepfake-apps-websites/

Stahl, B. (2021). From PAPA to PAPAS and Beyond: Dealing with Ethics in Big Data, AI and other Emerging Technologies. *Communications of the Association for Information Systems*, *49*. https://doi.org/10.17705/1CAIS.04920

Stubbersfield, J. M., Tehrani, J. J., & Flynn, E. G. (2015). Serial killers, spiders and cybersex: Social and survival information bias in the transmission of urban legends. *British Journal of Psychology*, *106*, 288–307.

Tammekänd, J., Thomas, J., & Peterson, K. (2020). *Deepfakes 2020: The tipping point*.

Thalen, M. (2021, September 10). Real-time deepfakes could bring chaos to your next Zoom. Retrieved December 22, 2021, from The Daily Dot website: https://www.dailydot.com/debug/deepfacelive-deepfake-live-streaming/

Thambawita, V., Isaksen, J. L., Hicks, S. A., Ghouse, J., Ahlberg, G., Linneberg, A., … Kanters, J. K. (2021). DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Scientific Reports*, *11*, 21896.

Tifferet, S. (2021). Verifying online information: Development and validation of a self-report scale. *Technology in Society*, *67*, 101788.

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*, 356–367.

Torraco, R. J. (2016). Writing Integrative Reviews of the Literature: Methods and Purposes. *International Journal of Adult Vocational Education and Technology (IJAVET)*, *7*, 62–70.

Tran, T. (2021, September 12). New Deepfake Tool Turns Livestreamers into Someone Else in Real Time. Retrieved December 22, 2021, from Futurism website: https://futurism.com/the-byte/deepfake-livestreamers-real-time

Truog, D. (2022, April 28). Beyond The Cartoonish Metaverse—XR Designers and Users Deserve Better. Retrieved July 22, 2022, from Forrester website: https://www.forrester.com/blogs/beyond-the-cartoonish-metaverse/

UCL. (2020, August 4). ‘Deepfakes’ ranked as most serious AI crime threat. Retrieved November 17, 2021, from UCL News website: https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat

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

Vafeiadis, M., Bortree, D. S., Buckley, C., Diddi, P., & Xiao, A. (2019). Refuting fake news on social media: Nonprofits, crisis response strategies and issue involvement. *Journal of Product & Brand Management*, *29*, 209–222.

Verdoliva, L. (2020). Media Forensics and DeepFakes: An overview. *ArXiv:2001.06564 [Cs]*. Retrieved from http://arxiv.org/abs/2001.06564

Vincent, J. (2021, May 18). Deepfake dubs could help translate film and TV without losing an actor’s original performance. Retrieved December 24, 2021, from The Verge website: https://www.theverge.com/2021/5/18/22430340/deepfake-dubs-dubbing-film-tv-flawless-startup

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting Deepfakes: Media and Internet Giants’ Converging and Diverging Strategies Against Hi-Tech Misinformation. *Media and Communication*, *9*, 291–300.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151.

Wagner, T. L., & Blewer, A. (2019). “The Word Real Is No Longer Real”: Deepfakes, Gender, and the Challenges of AI-Altered Video. *Open Information Science*, *3*, 32–46.

Wahl-Jorgensen, K., & Carlson, M. (2021). Conjecturing Fearful Futures: Journalistic Discourses on Deepfakes. *Journalism Practice*, *15*, 803–820.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*, xiii–xxiii.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*, 40–53.

Whittaker, L., Letheren, K., & Mulcahy, R. (2021). The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing. *Australasian Marketing Journal*, *29*, 204–214.

Wiederhold, B. K. (2021). Can Deepfakes Improve Therapy? *Cyberpsychology, Behavior, and Social Networking*, *24*, 147–148.

Woodie, A. (2022, April 21). Deepfakes, Digital Twins, and the Authentication Challenge. Retrieved July 23, 2022, from Datanami website: https://www.datanami.com/2022/04/21/deepfakes-digital-twins-and-the-authentication-challenge/

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

Yang, H.-C., Rahmanti, A. R., Huang, C.-W., & Li, Y.-C. J. (2022). How Can Research on Artificial Empathy Be Enhanced by Applying Deepfakes? *Journal of Medical Internet Research*, *24*, e29506.

Yang, X., Li, Y., & Lyu, S. (2018). Exposing Deep Fakes Using Inconsistent Head Poses. *ArXiv:1811.00661 [Cs]*. Retrieved from http://arxiv.org/abs/1811.00661

Yankoski, M., Scheirer, W., & Weninger, T. (2021). Meme warfare: AI countermeasures to disinformation should focus on popular, not perfect, fakes. *Bulletin of the Atomic Scientists*, *77*, 119–123.

Yu, X., Wojcieszak, M., Lee, S., Casas, A., Azrout, R., & Gackowski, T. (2021). The (Null) Effects of Happiness on Affective Polarization, Conspiracy Endorsement, and Deep Fake Recognition: Evidence from Five Survey Experiments in Three Countries. *Political Behavior*, *43*, 1265–1287.

Zhao, B., Zhang, S., Xu, C., Sun, Y., & Deng, C. (2021). Deep fake geography? When geospatial data encounter Artificial Intelligence. *Cartography and Geographic Information Science*, *48*, 338–352.

Ziegler, Z. (2021). Michael Polányi's fiduciary program against fake news and deepfake in the digital age. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01217-w

Appendix A: Summary of Studies included in the Literature Review

Table A1. Empirical Research on Deepfakes

Title	Methodology	Sample and area	Antecedents	Outcomes	Key findings
Ahmed (2021a)	Quantitative survey	Respondents from USA (n=764) and Singapore (n=662)	Political interest, cognitive ability, social network size, political trust	Inadvertent deepfake sharing	Political interest is positively connected to inadvertent sharing of deepfakes, while cognitive ability is negatively correlated, with social network size moderating the association between political interest and inadvertent sharing of deepfakes.
Ahmed (2021b)	Quantitative survey	1244 US residents	Deepfakes concern, inadvertent deepfake sharing, cognitive ability	Social media news skepticism	Deepfakes exposure and concerns are positively related to social media news skepticism, frequent social media users for news are less skeptical, and individuals with higher cognitive abilities are more skeptical. Moderation effects of deepfakes concern and inadvertent sharing are more amplified in low cognitive conditions.
Ahmed (2021c)	Quantitative survey	440 US residents	Deepfake exposure, cognitive ability, perceived accuracy of claims	Deepfake sharing intention	In the absence of information cues, people believe deepfakes are authentic, and they are more likely to share them. When informational cues are present, cognitive ability has a moderating function. Surprisingly, when the informational cues are absent, these people are more likely to believe the claim is real and disseminate it.
Ahmed (2022)	Quantitative survey	Respondents from USA (n=764) and Singapore (n=534)	Social media news use	Intentional deepfake sharing	Demonstrates that social media news use and fear of missing out (FOMO) are positively associated with intentional deep fakes sharing. Additionally, moderated mediation reveals that the indirect effects of social media news use on advertent sharing via FOMO are more pronounced for low cognitive persons than for high cognitive individuals.
Allison (2021)	Case study	American film	Racial appearance	Media circulation of racial appearances	Uses the example of an American film and its use of deepfakes to argue that, rather than repeating past "colorblind" misrecognitions, deepfake tools can be used for racial justice and equity.
Ayers (2021)	Case study	Action stars in the USA	Facial substitutions	Hegemony, ideological reclamation of masculine power	Examines deepfaked videos of action stars from the 1980s in order to project hegemonic white masculinity and the preservation of racial and cultural distinctions in the videos.
Barari et al. (2021)	Quantitative survey	5750 US residents	Audio/video stimulus (authentic and deepfake), digital literacy,	Credibility, emotional appeal, and skepticism towards	Demonstrates that deepfakes can convince the viewers of scandals that never occurred but no more than other forms of misinformation while confirming the role of

Table A1. Empirical Research on Deepfakes

Title	Methodology	Sample and area	Antecedents	Outcomes	Key findings
			political knowledge, partisanship	deepfakes	motivated reasoning in facilitating persuasion. Also highlights the low effectiveness of informational treatments in alleviating deepfake effects.
Bazarkina & Pashentsev (2019)	Case study	Official publications, monographs, and research articles from Russia and other countries	Malicious use of artificial intelligence	International psychological security	Highlights the potential of deepfakes to manipulate the electoral process and impact global politics and international relations
Bode (2021)	Case study	Data from Corridor's Digital's YouTube channels	Framing contexts	Viewer's evaluation, categorization and sense-making of images	Describes how viewers can tell the difference between real and fictitious information. Deepfakes are made, indexed, and received as a trick of the eye rather than with the purpose to deceive, according to the author.
Brennen et al. (2021)	Thematic analysis	96 samples of COVID-19 misinformation visuals from fact checker sites	Visuals of misinformation related to COVID-19	Misleading understandings about the virus	Analyzes manipulated visuals related to COVID-19 misinformation and indicates the use of simple tools to create them with no examples of deepfaked visuals
Conte (2019)	Case study	Picture of a French artist and art creations from Italy, USA	Hyperrealism	Mistrust in images	Raises concerns over the increasing blurring of lines between actual reality and digital reality due to deepfakes
Dasilva et al. (2021)	Social network analysis	18000 tweets data from Twitter	Network density	Manipulation targets	Analysis reveals that while adult content dominates deepfakes, political deepfakes capture the most attention
de Seta (2021)	Case study	Huanlian (deepfakes in China) development through ZAO app, deepfake commercialization, and communities of practice	Launch of deepfake apps	Societal backlash, regulatory response	Discusses deepfakes in the Chinese context alongside regulatory responses and emerging communities of practice around synthetic media
Diakopoulos & Johnson (2021)	Scenario analysis	2020 elections in the USA	Viewing of deepfakes	Harms to voters, electoral candidates, threat to electoral integrity	Examines ethical issues related to deepfakes and its harm to voters, political campaigns, candidates, and intervention mechanisms
Dobber et al. (2021)	Experimental study	278 Netherlands residents	Deepfake stimulus, religious orientation, degree of religiosity	Attitude towards the politician, attitude towards the political party	Deepfakes are a more potent style of misinformation, with deepfakes intended to discredit a political candidate having a negative impact on people's perceptions against the depicted politician but no significant changes in attitudes toward the political party. The impacts of the deepfake are significantly stronger for the microtargeted group than for the untargeted group, according to

Table A1. Empirical Research on Deepfakes

Title	Methodology	Sample and area	Antecedents	Outcomes	Key findings
					the study.
Fido et al. (2022)	Quantitative survey	290 and 364 U.K based participants for two separate studies	Celebrity status	Victim blame, perceived criminality, victim harm, proclivity creating and sharing	Individuals' judgments around the creation and sharing of celebrity male deepfakes created for self-sexual gratification were lenient. Higher levels of psychopathy predicted individuals' leniency in judgment and their proclivity to act
Gosse & Burkell (2020)	Discourse analysis	Corpus of articles from major publications in the USA, Canada, and U.K	Deepfake production and distribution	Promotion of false beliefs, undermining of the political process, generation of non-consensual explicit content	Investigates how the news media has portrayed the issues raised by deepfakes, such as their ease of manufacturing and distribution, their use to spread false information, undermine political processes, and produce non-consensual adult content.
Holliday (2021)	Case study	Viral deepfake videos of Hollywood film stars	Hollywood's deepfake videos	Cultural politics of identity, Hegemonic discourses	Describes the complications such as the sowing of mistrust, gender troubles due to excessive fabrication of Hollywood's digitally mediated performances
Köbis et al. (2021)	Experimental study	210 U.K citizens	Deepfake stimulus, awareness, financial incentive	Deepfake detection accuracy, confidence	People overestimate their ability to detect deepfakes. Awareness or financial incentives do not improve detection accuracy.
Mihailova (2021)	Case study	2 art projects in the USA and 1 deepfake ad	Employment of deepfakes	Advertising, audience engagement, educational outreach	Uses three case studies to examine how cultural institutions' acquisition of creative deepfake works acts as a legitimizing factor that can change the narrative about the technology's artistic value and societal uses.
Paris (2021)	Case study	200 video and image samples with audio-visual manipulation from 2016 to 2021	False audio-visual content	Harms as a result of dissemination of the content	Critically analyzes audio-visual impersonations and suggests that false impersonations through audio-visual samples are shaped and strengthened by structural powers
Popova (2019)	Digital ethnography	Data from two deepfake websites	Creation of deepfakes, circulation of deepfakes	Concern for intimacy, concern for authenticity	In comparison to other communities producing adult engagement content, Deepfake communities are less concerned about issues of intimacy and authenticity of the private person behind the deepfaked image, according to the study. Furthermore, deepfake communities are attempting to keep content contained within their groups.
Rupapara et al. (2021)	Sentiment analysis	5424 tweets data from Twitter	Machine learning classifiers	Polarity of deepfake related tweets	The number of tweets on deepfake technology is quite low on Twitter, which could be due to the novelty and lack of expertise among the general public.
Shin & Lee (2022)	Experimental study	230 American adults	Video news conditions,	Viral behavioral	Emphasizes the variable susceptibility to deepfake news and

Table A1. Empirical Research on Deepfakes

Title	Methodology	Sample and area	Antecedents	Outcomes	Key findings
			exposure to low-cost knowledge	intention (pre and post-knowledge)	the impact of pre-existing attitudes, as well as the role of media literacy in overcoming reasoning biases
Vaccari & Chadwick (2020)	Quantitative survey	2005 British residents	Deceptive deepfake, baseline trust in news, uncertainty	Trust in news	The findings indicate that while political deepfakes do not always deceive individuals, they do sow uncertainty and contribute to a decline in trust in social media news.
Vizoso et al. (2021)	Case study	3 Media outlets in USA and 3 Silicon Valley based Internet giants	Collaboration among platforms and media outlets	Detection, labeling, debunking of deepfakes	Discusses how media giants and high-tech companies are dealing with deepfakes as a new form of fake news
Wahl-Jorgensen & Carlson (2021)	Thematic analysis	Corpus of stories from Nexis U.K database	Journalistic responses to deepfakes, Concerns regarding inherent believability of audio-visual nature of deepfakes	Anxiety over future of information environment and journalism's role within it, Worry over impending weaponization of deepfakes	Suggests that journalistic responses to deepfakes reveal broader concerns about the future of journalism. Expresses concerns that the audio-visual character of deepfakes makes them fundamentally more believable than prior kinds of fake news
Yadlin-Segal & Oppenheim (2021)	Narrative inquiry	105 News media articles	Journalists' framing of deepfakes as a destabilizing platform	Undermining of a shared sense of social and political reality, enablement of abuse and harassment of women online, blurring of acceptable dichotomy between real and fake	Illustrates how journalists interpret deepfakes as a destabilizing platform that undermines a common sense of social and political reality, allows for online harassment and abuse of women, and blurs the boundary between true and fake news.
Yu et al. (2021)	Quantitative survey	USA, Poland, Netherlands (N=3611) and USA and Poland (N=2220)	Incidental emotions (happiness, anger, anxiety)	Affective polarization, conspiracy endorsement, deepfake recognition	There was no indication that incidental happiness had an influence on affective polarization, conspiracy endorsement, or belief that a deepfake is true

Table A2. Non-empirical Research on Deepfakes

Source	Industry/function	Type of paper	Key findings
Albahar & Almalki (2019)	N/A	Literature review	Examines how deepfakes are created and discusses techniques for deepfake detection
Botha & Pieterse (2020)	N/A	Literature review	Conducts a review of creation and detection techniques in the area of fake news including deepfakes
Campbell et al. (2021)	Advertising	Theoretical paper	Proposes a framework to understand consumer responses to ad manipulation through synthetic media such as deepfakes
Carvajal & Iliadis (2020)	N/A	Literature review	Conducts a preliminary literature review of academic works in the area of deepfakes
Chesney & Citron (2018)	Cross-domain	Discussion paper	Discusses deepfake characteristics, uses, harms, and ways to deal with it
Chesney & Citron (2019)	Politics	Discussion paper	Highlights the emerging role of deepfakes in the disinformation war, discusses legal options and how democracies will have to deal with lies
Crystal et al. (2020)	Medicine	Discussion paper	Discusses the relevance of deepfakes for plastic surgery and the benefits and harms of the technology.
de Ruiter (2021)	Ethics	Discussion paper	Discusses factors that determine if a deepfake is morally problematic and argues that deepfakes may be morally suspect but not inherently morally wrong
Delfino (2019)	Law and justice	Discussion paper	Discusses the legal options to deal with pornographic deepfakes and proposes legislative solutions
Dowdeswell & Goltz (2020)	High-Tech industry	Discussion paper	Discusses the impact of technology platform companies on civic life and proposes guidelines for platform governance
Fallis (2020)	N/A	Discussion paper	Analyzes the threat of deepfakes to the process of acquiring knowledge and its impact on causing false beliefs
Godulla et al. (2021)	Communication studies	Literature review	Focuses on a literature review of deepfake articles in the domain of communication studies
Gregory (2021)	Journalism	Discussion paper	Discusses the activities of WITNESS, a human rights and civic journalism network and highlights the importance of investigation of deepfakes and authenticity infrastructure to address credibility challenges
Harris (2021)	N/A	Discussion paper	Suggests that the epistemic threat posed by deepfakes can be mitigated and highlights the psychological impact of deepfakes
Hight (2021)	Film industry	Discussion paper	Highlights the increasing complexity of documentary forms and how synthetic media could potentially be used to develop more openly reflexive content.
D. Johnson & Diakopoulos (2021)	N/A	Discussion paper	Discusses ways of maximizing the benefits of deepfakes alongside reducing their negative effects
Kerner & Risse (2021)	Politics and Adult content	Discussion paper	Examines the potential of deepfakes to unleash human creativity while preserving epistemic rights and justice
Kietzmann et al. (2020)	N/A	Theoretical paper	Provides a classification of different types of deepfakes and offers a framework to deal with deepfake related risks
Kietzmann et al. (2021)	Advertising	Theoretical paper	Proposes a model to explore deepfake influence on culture and consumption patterns of consumers in the advertising space
Kikerpill (2020)	Politics and Adult content	Discussion paper	Raises concerns over the escalating ramifications of deepfakes through explicit content and disruption of electoral process
Langguth et al. (2021)	N/A	Discussion paper	Discusses the recent advancements in deepfake technology and compares it with past techniques alongside technical countermeasures to deal with it
Maras & Alexandrou (2019)	Law and justice	Discussion paper	Discusses treatment of deepfakes as evidence in court and impact of authentication mechanisms

Mullen (2022)	Law and justice	Discussion paper	Discusses deepfakes in modern society, how they are analyzed as evidence in the courtroom, and avenues for redressal of victims of deepfakes
O'Donnell (2021)	Law and justice	Discussion paper	Discusses deepfake-related threats, insufficiencies of existing legal options, and proposes amendments to deal with threats
Öhman (2020)	Adult content	Discussion paper	Uses levels of abstraction as a possible solution to resolve the ethical dilemma with regard to deepfake creation
Paterson & Hanley (2020)	Politics	Discussion paper	Discusses the impact of deepfakes on democratic elections and political discourse
Pavis (2021)	Law and justice	Discussion paper	Discusses an alternative to existing legal regulations for deepfakes through performer rights as a regulatory response
Porter (2020)	Bioethics	Discussion paper	Highlights the looming danger of epistemological pluralism inflicted upon the postmodern world through deepfakes
Ray (2021)	Politics and Law	Discussion paper	Discusses how Australian law deals with political deepfakes and proposes regulations that can reduce their threat to the electoral process
Ring (2021)	Cybersecurity	Discussion paper	Discusses how deepfakes pose a threat to cybersecurity when used in subversive ways by cybercriminals
Rini (2020)	N/A	Discussion paper	Highlights the erosion of knowledge in democratic societies due to deepfakes that impact testimonial practices, debates, and public discourses.
Sample et al. (2020)	Journalism	Theoretical paper	Discusses forms of fake news, including deepfakes, and introduces an evaluation model to analyze the dissemination patterns
Stadler (2019)	Entertainment business	Discussion paper	Investigates the ethical implications of cybernetics and digital embodiment technologies.
Verdoliva (2020)	N/A	Literature review	Provides a review of methods used for the detection of manipulated images and videos with a specific focus on deepfakes
Westerlund (2019)	N/A	Literature review	Performs a literature review of deepfakes with specific focus on online news articles
Whittaker et al. (2021)	Marketing	Theoretical paper	Proposes a framework for deepfake research in the marketing area
Yankoski et al. (2021)	N/A	Discussion paper	Argues for the need to prioritize and develop artificial intelligence techniques for the detection of shallow fakes over deepfakes.
Zhao et al. (2021)	Cartography	Technical paper	Introduces deepfakes in geography where maps are manipulated. Also discusses detection approaches and coping mechanisms.
Ziegler (2021)	Journalism	Discussion paper	Proposes the creation of a standard for online news that people can trust and rely on for information

Appendix B: Summary of Key Research Agendas and Implications

Table B1: Thematic Gaps, Potential Research Questions and Contributions

Research gaps	Potential research questions (RQs) for future studies	Contributions
Theme 1: Definitional issues	RQ1: What media manipulations constitute deepfakes and what do not? RQ2: What are the existing typologies for deepfake categorization? RQ3: Can the typologies be consolidated towards a holistic definition of deepfake with an overarching schema?	• A shared understanding of deepfakes through a comprehensive, all-encompassing definition. • A unified taxonomical framework for deepfakes.
Theme 2: Lack of standard measures	RQ4: How can we reliably and consistently measure the level of trust and skepticism in a deepfake held by an individual? RQ5: How can we accurately estimate the level of deepfake awareness and exposure through a reliable index? RQ6: What is a valid and reliable measure of the level of concern held by individuals around deepfakes? RQ7: What are the temporal and contextual aspects which determine our accuracy judgment and dissemination of deepfakes? RQ8: What are the causal factors which determine our accuracy judgment and sharing of deepfakes?	• Establishment of an operational measurement for key constructs such as deepfake concern, exposure, and belief in it. • Deeper insights and potential to develop practically grounded interventions to deal with deepfake-related consequences
Theme 3: Antecedents and outcomes	RQ9: What are the characteristics which determine the belief in the deepfake or its accurate detection? RQ10: What are the deepfake receiver's intentions to respond to the received deepfake with further perpetration? RQ11: What are the impacts of deepfake on the deepfaked person and receiver?	• Development of a nuanced understanding of the mechanisms which lead to belief in a deepfake and further dissemination • Comprehensive understanding of deepfake consequences at individual and institutional levels
Theme 4: Lack of cross-geographic coverage	RQ12: How do deepfake creation motivations vary across differently developed economies? RQ13: Do antecedents of deepfake sharing show significant differences across different geographic regions? RQ14: How do cultural norms around communication and social behavior influence the sharing, viewing, and impact of deepfakes? RQ15: How do deepfake dissemination patterns vary across cultures?	• Advancement of deepfake understanding through cross-geographic and culturally diverse investigations. • Context-specific interventions to curb negative influences of deepfakes. • Improvisation of deepfake detection methodologies through cultural and geographically specific dissemination patterns.
Theme 5: Investigation of demographic characteristics	RQ16: How do deepfake viewers across different age cohorts and gender process deepfakes? RQ17: How do demographic characteristics such as education and occupation affect users' interaction with deepfakes?	• Development of tailored interventions based on empirical investigation of influences exerted by age, gender, education, occupation, and other demographic variables on deepfake engagement.
Theme 6: Theoretical grounding	RQ18: Why do people believe in and share deepfakes? RQ19: What are the ideological, cultural, political, emotional, and logical factors which deepfake creators build on?	• Standardized and holistic explanations for deepfake engagement through the use of established theoretical models from psychology, management, and other relevant fields of research
Theme 7: Deepfake viewing-sharing dissociation	RQ20: What are the determinants of disconnect between viewing and sharing of deepfakes? RQ21: How does the dissociation between accuracy judgment and sharing intentions vary by deepfake domains?	• Distinctions between and explanations for deepfake accuracy judgment and sharing intentions
Theme 8: Thematic tensions	RQ22: When does a deepfake transition from being morally suspect to morally wrong? RQ23: How do individuals and institutions engage with	• Resolution of moral tensions within deepfake engagement • Holistic understanding of deepfake

Table B1: Thematic Gaps, Potential Research Questions and Contributions

Research gaps	Potential research questions (RQs) for future studies	Contributions
	deepfakes across different domains? RQ24: How does individual and institutional level engagement with deepfakes across non-political domains vary in comparison to engagement with political deepfakes?	engagement across various domains • Comparison of engagement with political deepfakes and non-political deepfakes
Theme 9: Divergent perspectives	RQ25: Does deepfake engagement have a dual nature encompassing both positive and negative influences? RQ26: What is the role of the deepfake creator, disseminator, and viewer in determining the influence of a deepfake? RQ27: What is the threshold at which deepfake engagement becomes detrimental for the subjects involved in the process?	• Conceptual advancement of deepfakes and elucidation of its nature • Determination of the threshold point for detrimental effects of deepfakes

About the Authors

Pramukh Nanjundaswamy Vasist is a doctoral student in the Information Systems Area at the Indian Institute of Management (IIM) Kozhikode. His research interests include fake news, social media, mobile addiction, and related behavioral and managerial issues in the context of information systems. He has published in leading journals, such as the International Journal of Hospitality Management and has also published in preeminent conferences, including the Australasian Conference of Information Systems (ACIS). He has served as a reviewer in journals and conferences such as IIM Kozhikode Society and Management Review, Pacific Asia Conference on Information Systems (PACIS), ACIS and IIM World Management Conference (WMC).

Satish Krishnan received his PhD in Information Systems from the National University of Singapore. He is the Chair Associate Professor of Information Systems at the Indian Institute of Management (IIM) Kozhikode. His research includes IT resistance, fake news and disinformation, gender gap, e-government, e-business, virtual social networks, technostress, cyberloafing, and cyberbullying. He has published in leading journals, such as the Journal of Applied Psychology, Organizational Behavior and Human Decision Processes, Information and Management, International Journal of Information Management, Journal of Association for Information Science and Technology, International Journal of Hospitality Management, Communications of the Association for Information Systems, Computers in Human Behavior, Information Systems Frontiers, Scandinavian Journal of Information Systems, Technological Forecasting and Social Change, Journal of Retailing and Consumer Services, Human Resource Development Review, Journal of Global Information Technology Management, and e-Service Journal. He is on the editorial boards of various journals such as Internet Research, Technological Forecasting and Social Change, Information Systems Frontiers, International Journal of Information Management, and Computers in Human Behavior. He also serves at various conferences such as PACIS and ICIS as Track Chair or Review Coordinator or Associate Editor. He won the Outstanding Associate Editor Award for ICIS 2017 and 2019.