SCM
Studies in Communication and Media

FULL PAPER

Dealing with deepfakes – an interdisciplinary examination of the state of research and implications for communication studies

Der Umgang mit Deepfakes – Eine interdisziplinäre Untersuchung zum Forschungsstand und Implikationen für die Kommunikationswissenschaft

Alexander Godulla, Christian P. Hoffmann, & Daniel Seibert

Alexander Godulla (Prof. Dr.), Institute for Communication and Media Studies, Leipzig University, Nikolaistraße 27-29, 04109 Leipzig; Contact: alexander.godulla(at)uni-leipzig.de. ORCID: https://orcid.org/0000-0002-1011-0639

Christian P. Hoffmann (Prof. Dr.), Institute for Communication and Media Studies, Leipzig University, Nikolaistraße 27-29, 04109 Leipzig; Contact: christian.hoffmann(at)uni-leipzig.de. ORCID: https://orcid.org/0000-0002-5282-6950

Daniel Seibert (M.A.), Institute for Communication and Media Studies, Leipzig University, Nikolaistraße 27-29, 04109 Leipzig; Contact: daniel.seibert(at)uni-leipzig.de. ORCID: https://orcid.org/0000-0001-7258-8753

FULL PAPER

Dealing with deepfakes – an interdisciplinary examination of the state of research and implications for communication studies

Der Umgang mit Deepfakes – Eine interdisziplinäre Untersuchung zum Forschungsstand und Implikationen für die Kommunikationswissenschaft

Alexander Godulla, Christian P. Hoffmann, & Daniel Seibert

Abstract: Using artificial intelligence, it is becoming increasingly easy to create highly realistic but fake video content – so-called deepfakes. As a result, it is no longer possible always to distinguish real from mechanically created recordings with the naked eye. Despite the novelty of this phenomenon, regulators and industry players have started to address the risks associated with deepfakes. Yet research on deepfakes is still in its infancy. This paper presents findings from a systematic review of English-language deepfake research to identify salient discussions. We find that, to date, deepfake research is driven by computer science and law, with studies focusing on deepfake detection and regulation. While a number of studies address the potential of deepfakes for political disinformation, few have examined user perceptions of and reactions to deepfakes. Other notable research topics include challenges to journalistic practices and pornographic applications of deepfakes. We identify research gaps and derive implications for future communication studies research.

Keywords: Artificial Intelligence, AI, deep learning, deepfakes, deep fakes.

Zusammenfassung: Mit Hilfe von Künstlicher Intelligenz wird es immer einfacher, gefälschte, hochrealistische Videoinhalte – sogenannte Deepfakes – zu erstellen. Dadurch ist es kaum mehr möglich, mit bloßem Auge zu unterscheiden, ob eine Aufnahme echt ist oder von einer Maschine erzeugt worden ist. Trotz der Neuartigkeit dieses Phänomens haben Regulatoren und Akteure der Branche erst damit begonnen, sich mit den Risiken von Deepfakes auseinanderzusetzen. Die Forschung zu den Auswirkungen von Deepfakes steckt allerdings noch in den Kinderschuhen. In diesem Paper werden die Ergebnisse eines Literature Reviews zur Deepfakes-Forschung vorgestellt, um relevante Diskussionen zu diesem Thema zu identifizieren. Wie sich zeigt, wird die Deepfake-Forschung bislang vorrangig von der Informatik und den Rechtswissenschaften angetrieben, wobei sich zahlreiche Studien auf die Erkennung und Regulierung von Deepfakes konzentrieren. Während sich eine Reihe von Studien auf die potenzielle Verwendung von Deepfakes für politische Desinformation fokussieren, befassen sich nur wenige Arbeiten mit deren Auswirkungen auf die Mediennutzenden. Wichtige Forschungsschwerpunkte stellen außerdem die Herausforderungen für journalistische Praktiken und die pornografischen Verwendungen von Deepfakes dar. Anhand des Literature Reviews identifizieren wir Forschungslücken und

leiten daraus Implikationen für die zukünftige kommunikationswissenschaftliche Forschung ab.

Schlagwörter: Künstliche Intelligenz, KI, Maschinelles Lernen, Deepfakes, Deep Fakes.

1. Introduction

At the 2020 World Economic Forum Meeting in Davos, USC associate professor Hao Li kept the meeting's prominent attendees entertained and fascinated by a screen installation that allowed users to superimpose over a live recording of their own face that of a choice of Hollywood actors, such as Leonardo DiCaprio or Will Smith. The chosen actor's face would seemingly mimic each movement and utter each word performed by the recorded user. The installation served to demonstrate the power of deepfakes – artificial intelligence-based manipulated (audio-)visual material. Hao Li also contributed a talk to the meeting titled "Do Not Believe What You See" and advertised as a discussion of "a new dimension of fake news that pose[s] a danger to democracy and vulnerable groups" (Lichfield & Li, 2020).

The example illustrates that the relatively recent development of deepfakes is increasingly capturing the attention and imagination of experts and decision-makers. Deepfakes are but one element in the influence of artificial intelligence (AI) over today's society (Makridakis, 2017), and in this context, the continuous improvements that artificial intelligence achieves in imitating human capabilities are of particular interest. So-called deep learning is a driving force in this development (Jordan, 2019). In simple terms, deep learning is a method of machine learning in which large amounts of data are processed by self-learning neural networks. On this basis, the AI is enabled to generate forecasts or decisions. As a result, AI-based technology will increasingly be able to perform tasks previously considered the exclusive domain of human capabilities.

As the emergent field of human-machine interaction (also: human-machine communication; Fortunati & Edwards, 2020) illustrates, AI-based technology is expected to affect, change and challenge established social dynamics. As with any emergent innovation, the shape and scope of this influence remains as yet somewhat opaque. Communication scholars have just begun to tackle the potential influences of AI technologies on public and interpersonal communication. Recent studies in the field have largely focused on the effect AI may have on the generation and dissemination of disinformation. Ever since the 2016 US presidential election, fake news and dwindling trust in the media has received tremendous research attention (Müller & Denner, 2017). The concept of the post-factual age has emerged, in which the perceived boundaries between fact and fake are increasingly blurred (Lewandowsky, Ecker & Cook, 2017). Deepfakes are situated at the intersection of the fake news discourse, emergent AI technology and human-machine interaction. AI technologies make it possible to fake video and audio content in a way that appears deceptively true to the original. As a result, audiences will find it increasingly difficult to distinguish whether an audio or video recording is real or has been created by a machine (Mattke, 2018).

To date, only a few studies in the field of communication studies have addressed deepfakes. However, research from other disciplines has tackled the phenomenon – pursuing distinctive questions and applying various perspectives. This paper presents an overview of the current state of research on deepfakes based on a systematic literature review. The goal is to identify domains of interest and concern, highlight gaps in this literature, and point to places where communications scholars could make a useful contribution. In order to achieve this goal, a differentiated consideration of the deepfake topic from different scientific perspectives is conducted. Due to the localization of the phenomenon in the field of computer science, it is essential to consider this perspective. Based on this fundamental work, conclusions on future research in the social sciences and especially for communication studies will be drawn. The results of the literature review will allow an integration of different perspectives into a social science model for investigating deepfakes. The following research questions will be addressed:

RQ1: What are the salient perspectives and key concepts in the emergent interdisciplinary research on deepfakes?

RQ2: What are the implications for research on deepfakes from a communication studies perspective?

2. Definition, development and occurrence of deepfakes

Before reviewing and synthesizing the research of different scientific fields on deepfakes, it is necessary to provide an initial definition of the core phenomena:

Artificial Intelligence

Artificial intelligence can be described as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 17).

Deepfakes

Deepfakes can be defined as the output of an "artificial intelligence-based image synthesis technique that involves creating fake but highly realistic video content", through which it is possible to "change how a person, object or environment is presented" (CDEI, 2019). It is not only possible to change visual content, but "existing recordings of a person's [...] voice can be used to reverse engineer their speech to have them say any sentence" (Vaccari & Chadwick, 2020). The malicious use of deepfakes is strongly linked to automated propaganda and disinformation campaigns as well as the dissemination of misinformation and conspiracy theories through social media (EPRS, 2019). In addition, malicious deepfakes can be intended to deceive and mispresent individuals or companies (for economic benefits), for example, by showing a person yelling racist epithets or using drugs, or by showing a rival company's chief executive hiring underage prostitutes. In

such cases, deepfakes could have an impact on a person's reputation or career, or a company's business opportunity (Citron & Chesney, 2018). However, the dissemination of deepfakes may also be driven by satirical purposes or, more generally, the goal of entertaining an audience (Tandoc, Lim & Ling, 2018). According to the Centre for Data Ethics and Innovation (CDEI, 2019), four types of deepfake can be distinguished:

1) Face replacement: also known as face swapping – the recording of a face (source) and its transfer onto the face of someone else (target)
2) Face re-enactment: the changing of the facial features of a person
3) Face generation: the creation of new faces that do not reflect a real person
4) Speech synthesis: the creation and transmission of mimicking a real voice in tonality and frequency

2.1 Occurrence of deepfakes

While faking visual imagery (photo or video) has been feasible for some time now, similarly faking audio content, particularly human voices, used to require significant resources (Vincent, 2019). Today, software is becoming more widely available that can largely overcome this limitation. While we may enjoy the realism of AI-enabled personal assistants communicating verbally with their users, potential dangers could emerge. For example, spam calls with fake voices can be maliciously initiated to harass people or spread disinformation as supposedly real audio recordings of politicians (Vincent, 2019). Software companies like Adobe are working to establish services for intelligent processing of voice recordings on a widespread basis (ibid.). Moreover, AI technologies reduce the cost, time and expertise required to create deepfakes (Metz, 2019).

An analysis carried out by Deeptrace (now called Sensity) in summer 2019 identified 15,000 videos openly available on the web as deepfakes, including those found on platforms such as YouTube, Vimeo, LiveLeak or Dailymotion as well as deepfakes on most common pornography websites. The prevalence of deepfakes had thus doubled compared to their investigation conducted a mere seven months earlier. About 96 percent of the examined deepfakes featured pornographic content, with all individuals artificially inserted in the sequences being women. Although this analysis is certainly not exhaustive, it nevertheless indicates that deepfake technology is currently used primarily to create pornographic content (Simonite, 2019).

With continuous improvements to deepfake technology and its current proliferation, it is to be expected that deepfake content will increasingly emerge in new social contexts, including politics. Because of the immediacy of audiovisual communication, manipulated video or audio material seems especially suited to spread disinformation. Conversely, with the rising difficulty of assessing the veracity of audiovisual content, public actors may increasingly dismiss video documentation of their statements or actions as deepfakes in order to dodge public criticism (Turton & Martin, 2020).

A recent widely discussed fake video showed former US president Barack Obama apparently making negative comments about his successor Donald Trump, calling him "a complete dipshit". Following a statement of denial from Obama,

the video was confirmed as a deepfake by its creator. However, the example highlights the ease with which audiovisual statements by public figures can be faked – as well as the quality of the deceit. While distorting or even fabricating statements may be a familiar element of public political discourse, the fabrication of audiovisual documentation of such content adds a new dynamic to the challenge of separating real from fake – due to the potency of audiovisual content, users' habitual reliance on the veracity of such content, and the critical importance of trustworthy news sources (Fehrensen & Täubner, 2019).

Of particular relevance for the potential effectiveness of deepfakes is their (viral) distribution through social networking sites. On Facebook recently, an artificially slowed-down video of the US politician Nancy Pelosi giving the false impression that she was drunk during a TV show, was not removed initially. Facebook has since responded and confirmed that in the future, videos will be deleted if (1) they are synthetically manipulated, the manipulations are not easily apparent and there is a chance that someone thinks a person has said words that they have not actually said, and (2) these videos have been created by AI or machine learning and have the intention of appearing authentic. Exceptions to this rule are parodies and satire (Turton & Martin, 2020).

However, the issue is that for any of these processes to be put into effect, a platform has to identify that a video is a deepfake in the first place. Adobe Research and UC Berkeley are therefore working together to develop a method for reconstructing image edits (Hillen, 2019).

Furthermore, a digital forensics competition was organized to generate fake photographs, video, and audio recordings and to develop methods for their automated identification (Mattke, 2018). This points to one particular challenge that is difficult to resolve. While the generation of deepfakes is becoming increasingly easy and more powerful, deepfake detectors are being developed which continuously learn and optimize their capabilities. The question is (Metz, 2019): Which of these two sides will develop faster? If deepfake detection tools are trained based on the available content, will they always remain one step behind the production of deepfake content?

2.2 Deepfakes' connection to fake news

The examples discussed above illustrate how the rapid emergence of deepfakes affects various domains of personal and public communication – and poses a challenge to research in popular and political communication. While communication studies has not focused much on deepfakes yet, it appears evident that deepfakes potentially threaten trust in media and may sway public opinion. In this context, literature from the fake news and disinformation discourse is relevant to identify points of interest for future communications studies.

Following Wardle and Derakhshan (2017), who examine the "information disorder" and its related challenges, research on disinformation is essential to find solutions for its mitigation. The authors define disinformation as "information that is false and deliberately created to harm a person, social group, organization or country" (ibid., p. 20). The intention of disinformation is strongly linked to

that of malicious deepfakes, with the difference that deepfakes are created using technologies based on artificial intelligence.

To date, the fake news discourse has primarily focused on text-based items used to create a false narrative (Tandoc et al., 2018; Wardle & Derakhshan, 2017). However, it is assumed that created or manipulated visual content may be even more persuasive. Taking the further development of technological possibilities into account, Wardle and Derakhshan (2017) suggest paying more attention to (audio-)visual types of disinformation. In addition, they suggest considering the elements and phases of manipulated information. In doing so, research should consider who created and disseminated the message, what the intention was, what its characteristics are and how the recipient(s) of the message reacted or interpreted it (ibid.). In addition, it should be examined whether the creator intends to mislead recipients (in the vein of fake news) or if the material is satire and meant to entertain people (Tandoc et al., 2018).

Furthermore, the phases when the message was created, transformed into a media product and (publicly) disseminated is of interest when analyzing disinformation (Wardle & Derakhshan, 2017). Current research on fake news indicates that mainstream news media play an important role in its dissemination (e.g., Allen, Howland, Mobius, Rothschild, & Watts, 2020; Tsfati et al., 2020), while disinformation outlets rarely achieve sizable reach and exert a limited impact on the public flow of information (Cinelli, Cresci, Galeazzi, Quattrociocchi, & Tesconi, 2020).

The suggested strategy of classifying disinformation can also be applied to the deepfake phenomenon. Accordingly, it is of interest to find out the characteristics, origin, phases of formation, ways of dissemination and intention of the creator of a deepfake. At the same time, the effects of existent deepfakes need to be examined. While studies on fake news focus primarily on text-based material, the effects of manipulated (audio-)visual material on the recipients should be investigated – especially as deepfake technology evolves.

The following literature review will provide a broad interdisciplinary overview of the state of research on deepfakes published to date, highlight insights from a social science perspective, and derive further implications for future communication studies research.

3. Materials and methods

For the systematic literature review, searches in a number of databases for scientific publications were conducted (see Table 1) over the period from April 27th to May 11th, 2020. The database *IEEE Xplore Digital Library* was added to this list of databases as the phenomenon of deepfakes constitutes an emergent technology, rendering insights from technical disciplines especially relevant. For the search, the following keywords were applied:

1) deepfake / deep fake
2) fake digital content
3) fake video
4) deep learning video
5) video manipulation.

Due to the recency of the deepfake phenomenon, the search in the databases was limited to publications from the years 2018 to 2020. Furthermore, both (peer-reviewed) journal articles as well as conference papers were included in the sample. Given the breadth of literature available in *Google Scholar*, only the first 100 hits for every keyword were analyzed in their entirety. In addition, every further 50th hit (up to N=1000) was considered. Therefore, the abstracts and the table of contents of all hits were reviewed, and the articles were included in the sample if at least one of the following two criteria were fulfilled: (1) deepfakes are the research subject of the article, and/or (2) at least one chapter is dedicated to deepfakes. If none of these criteria were fulfilled, the article was excluded from the sample.

Table 1. Hits for the keyword search

Keyword	Database	Hits	Initial check	Relevant
deepfake / deep fake	IEEE *Xplore* Digital Library	30 / 9	30 / 9	16
	SAGE Journals	9 / 10	9 / 10	2
	Taylor & Francis Online	28 / 15	28 / 15	6
	Web of Science	23 / 7	23 / 7	6
	Google Scholar	1100 / 600	118 / 110	7
fake digital content	IEEE *Xplore* Digital Library	1	1	
	SAGE Journals	-	-	
	Taylor & Francis Online	-	-	-
	Web of Science	1	1	
	Google Scholar	8	8	
fake video	IEEE *Xplore* Digital Library	8	8	
	SAGE Journals	10	10	
	Taylor & Francis Online	13	13	-
	Web of Science	-	-	
	Google Scholar	503	108	
deep learning video	IEEE *Xplore* Digital Library	70	70	
	SAGE Journals	-	-	
	Taylor & Francis Online	1	1	-
	Web of Science	3	3	
	Google Scholar	296	103	
video manipulation	IEEE *Xplore* Digital Library	4	4	
	SAGE Journals	10	10	
	Taylor & Francis Online	12	12	-
	Web of Science	10	10	
	Google Scholar	749	112	
Total		3,630	833	37

The search for the separate spelling of the first keyword ("deep fake") resulted in a lot of overlaps compared to the search term "deepfake". The results for the other keywords turned out in numerous hits, but with no significance for the deepfake discourse.

In summary, a total of 3,630 hits were generated by the keyword searches. Based on the search strategy outlined above, 833 of those hits were initially checked. As it turned out, only the results of the first keyword search (i.e., "deepfake") contained relevant literature, and the majority of even the hits for the first keyword phrase (359) did not meet the pre-defined selection criteria. We found many overlaps among the hits derived from the different databases. Since Google Scholar was used last for the keyword searches, there was also a lot of overlapping here, so that despite the number of hits, only a small number of relevant (additional) articles were included in the sample. All in all, after removing the duplicates, 37 relevant studies addressing the deepfake phenomenon were identified. A closer look at this body of research reveals that more than half of the available studies are conceptual works and most of them focused on risks associated with deepfakes (see Table 2).

Table 2. Summary of analytical approach and type of deepfake approach

Items	Prevalence
Analytical approach	Conceptual: 20 Empirical: 17
Type of deepfake approach investigated	Focus on risks: 31 Focus on opportunities: 6

An analysis of the empirical studies (n = 17) showed that only two of them examined groups of people (see Table 3). These studies focused on the respondents' competences to identify deepfakes. In contrast, most studies employed an experimental design developing and testing methods to automatically identify deepfakes. In doing so, different data sets with varying numbers of deepfakes were used and examined. Further, the analysis showed that research on deepfakes is being initiated in several countries around the world.

4. Results

The categorization of the results was derived inductively from the material. In the end, individual items were integrated into the main categories of opportunities, risks and regulations associated with deepfakes.

An overview of the literature review immediately reveals a predominantly cautionary perspective on the deepfake phenomenon throughout different disciplinary and topical perspectives, and therefore numerous approaches deal with attempts to combat or regulate deepfakes. Of the 37 identified studies, 31 focus on risks, challenges or dangers associated with deepfakes. However, six studies differ notably from this predominant framing in that they focus on opportunities provi-

ded by the emergent technology (see Table 2). When differentiating disciplinary approaches to the phenomenon, we find that computer science studies focus largely on the detection of deepfakes, while studies in the field of law, unsurprisingly, focus on the legal frameworks required to mitigate potential risks associated with the technology. These two fields, computer science and law, dominate the body of research, while only seven of the selected studies contribute to the discourse from a humanities or social science perspective.

Table 3. Summary and methodological preferences of the examined empirical studies (*n* = 17)

Item	Prevalence
Examined groups of people	Experiments: 1 study: *n* = 30; 1 study: *n* =2.005 No groups of people were examined: 15
Main method applied	Content analysis: 1 Experiment: 16
Geographical place(s) of data collection	Canada: 2 China: 1 France: 1 Germany: 1 Jordan: 1 Korea: 1 the Netherlands: 1 Norway: 1 Serbia: 1 United Arabian Emirates: 1 United Kingdom: 1 United States: 5
Dataset of deepfakes examined Different type of dataset	Mean: *n* = 240; 1 study: *n* = 2; 1 study: *n* = 3; 1 study: *n* = 4; 1 study: *n* = 8; 1 study: *n* = 48; 2 studies: *n* = 49; 1 study: *n* = 100; 1 study: *n* = 175; 1 study: *n* = 600; 1 study: *n* = 620; 1 study: *n* = 640; 1 study: *n* = 820; 3 studies: not applicable 1 study: *n* = 84 (news articles)

To provide an overview of the state of research, salient perspectives and key concepts, we first distinguish studies focusing on opportunities provided by deepfakes from those focusing on risks. As there are far more of the latter, we then differentiate specific emergent discussions among the studies applying a cautionary perspective. Finally, we discuss the state of research on risk mitigation – specifically legal and technological approaches to deepfake regulation.

4.1 Opportunities of deepfakes

While most studies from technical disciplines assume potential harms caused by deepfakes and therefore focus on developing detection techniques (e.g., Agarwal et al., 2019; Afchar, Nozick, Yamagishi, & Echizen, 2018; Korshunov & Marcel, 2018; Li, Chang, & Lyu, 2018; Yang, Li, & Lyu, 2019), a few address potential benefits of the usage of deepfakes.

One domain where they could be used in this context is virtual reality applications (Bose & Aarabi, 2019). For example, deepfake technologies can be applied in the fashion industry. Under the term "deepfashion", deepfakes make it possible to try on clothes virtually or project them onto one's own body (Liu, Chen, Liu, & Lew, 2019). Beyond the fashion industry, advertising and the entertainment industry can benefit from the usage of deepfakes – for instance, due to the opportunity of employing more realistic stunt doubles (Kietzmann, Lee, McCarthy, & Kietzmann, 2020).

Another approach to the possible field of application is pursued by Floridi (2018), who discusses concepts such as originality and authenticity in the context of deepfake art. He applies the term *Ectype*, derived from Greek, and denoting a special relation between a copy and its original source. Using the example of the Rembrandt created by Microsoft, a fake Rembrandt painting which is the result of a data analysis of his complete works, he distinguishes between two types of Ectype: The painting could be considered "authentic" in style, as the algorithm has adopted the typical characteristics of a Rembrandt, but not "original" in terms of the source, since a new painting has been created that did not exist before. Conversely, an Ectype can be "original" in relation to its archetypal source, but not "authentic" in production or representation. An example of the latter would be an audio simulation of John F. Kennedy's last speech: The text is original, but the voice recording is generated by software that learned to imitate Kennedy's speech based on his recordings. Thus, art and other fields such as education can be enriched by deepfakes with new audio and visual material. For example, based on one image of Martin Luther, deepfake technology could be used to create an animation of him describing the process of the reformation of the church.

Another example of a meaningful use of deepfakes is sermons. Online audio or video recordings of sermons can be edited and modified using deepfake technology. As a result, they can be made available in other languages to overcome language barriers in the community (Anderson, 2019).

From the legal perspective, various authors address negative consequences of pornographic deepfakes and discuss their legal regulation (e.g., Delfino, 2019; Douglas, 2019; Franks & Waldman, 2019). In contrast, others address how, with the help of a consensus-based app, the legitimate creation of pornographic deepfakes could be made possible (Raffaghello, Kastalio, Kalf, & Paisley, 2019), thus increasing the autonomy of users of the application. Raffaghello et al. (2019) describe the development of an app that allows users to make their faces available to other users as downloads for pornographic deepfakes. The consent of the users is intended to create a permission-based environment for the use of deepfakes and to prevent their illegal use.

This approach is related to the idea that pornographic deepfakes can be seen as sexual fantasies (Öhmann, 2019), and therefore the application of deepfakes in a pornographic context is not necessarily unethical. However, when creating and/or using pornographic deepfakes, a distinction should be made between morally permissible and impermissible content. For example, there is a high level of abstraction if A creates a deepfake video based on B. As long as no unethical conventions, such as racism or gender inequality are breached, this could be likened to a morally permissible sexual fantasy – even if B has not consented.

In summary, the use of deepfakes results in numerous opportunities for application. First, deepfakes can support the development of virtual reality applications. Second, the advertising and entertainment industry can benefit from deepfake technology. Third, due to the creation of new, authentic audio and visual material, deepfakes can be used in the field of art and education. Finally, there is the opportunity to create legally and morally permissible applications of deepfakes in pornography – a field where, to date, the technology has found the most avid use.

4.2 Risks of deepfakes

As described above, most available studies on deepfakes apply a critical perspective to the phenomenon and explore the risks, challenges or dangers associated with deepfakes – or potential approaches to their mitigation. As this cautionary body of research is more substantive, we identify distinct emergent discussions – in particular, studies on risks in the context of deepfake pornography, and challenges to public communication, such as disinformation.

4.2.1 Overview

While many studies share a critical perspective on the emergent deepfake phenomenon (e.g., Citron & Chesney, 2018; Kietzmann et al., 2020, Yadav & Salmani 2019), we find a variety of specific challenges, risks or dangers addressed in the available body of research. Firstly, deepfakes could negatively affect the wellbeing of an individual or an organization (Kietzmann et al., 2019). For example, from a business perspective, voiceover artists may suffer from the fact that languages and lip movements can be changed more easily through deepfakes, while firms could fall victim to trickery by manipulated audio or video statements and CEOs could be presented in deepfakes in suspicious situations, damaging the reputation of their company. Further, deepfakes could be used to damage political campaigns and manipulate public opinion. In addition, through the use of deepfakes, one person can steal another's identity for financial benefit or to sabotage their reputation (Citron & Chesney, 2018).

Further harms to society caused by deepfakes could be the distortion of democratic discourse due to the circulation of false information, or the manipulation of elections through the application of audio or video deepfakes of a candidate. As a result, the spread of false information through deepfakes may reduce trust in institutions, undermine public security and diplomacy and be a threat to journalism (ibid.).

Another possible risk caused by deepfakes is related to law enforcement agencies (Yadav & Salmani, 2019). For example, evidence could be manipulated, and an innocent person could be falsely depicted in a criminal situation. On the other hand, there are currently still limitations to deepfakes, such as the large amount of data required, the time needed to achieve a realistic output, and the high costs of the technical equipment. At this point, it remains necessary to investigate to what extent these limitations change as the technology advances and improves.

4.2.2 Risks due to pornographic deepfakes

As pornography is currently the most salient context for deepfake applications, a number of studies focus on challenges in this particular domain. It should be noted that most of these papers originate from the legal perspective of and refer to the legal situation in the USA.

Non-consensual deepfake pornography can be analyzed as an extension of patriarchal power and as a form of sexual violence (Raffaghello et al., 2019), while the emergence and spread of deepfakes could reinforce gender differences in visual content (Wagner & Blewer, 2019). Accordingly, a number of conceptual publications deal with the legal regulation of deepfakes in relation to their pornographic use (e.g., Delfino, 2019; Douglas, 2019; Franks & Waldman, 2019).

As well as considering deepfakes as a form of sexual violence or negative reinforcement of gender differences, deepfake pornography can be described as a violation of freedom of opinion and speech, especially of vulnerable groups of people. By using audio, photos, or video recordings to create a deepfake and make the target say any sentence, the targets' freedom of speech is undermined (Franks & Waldman, 2019).

Further, deepfakes carry the risk of influencing people's ability to distinguish between truth and falsehood and can damage the dignity of a human being. In particular, specific groups of people, such as gender minorities or racial minorities are prone to be victims of abusive widespread deepfakes. Frank and Waldman warn that, due to the influence of "powerful actors" (2019), the needs of minorities may be overlooked in regulatory approaches to deepfakes. However, specific regulations have yet to be developed to counter this threat.

In addition, several tort laws and recent laws on non-consensual pornography in the USA do not cover deepfakes of non-prominent persons. Therefore, both Delfino (2019) and Douglas (2019) suggest the development of a federal criminal law that prohibits pornographic deepfakes of individuals who have not consented to their use.

4.2.3 Risks to public communication and journalism

The use of deepfakes causes several risks in the context of journalism and (political) communication, but to date, only a few studies focus on the impact of deepfakes on their recipients. A risk resulting from the distribution of deepfakes is that they could lead to lower levels of trust in news on social media. In addition, the public could be confused regarding the quality of online news due to the spread of deepfakes (Vaccari & Chadwick, 2020).

An important question that arises at this point is: Does the public have the skills to identify deepfakes? A study by Khodabakhsh, Ramachandra, and Busch (2019) shows that their participants' misinterpretation of real videos as fake and vice versa was less than 30%. The majority of deepfakes do not yet seem to appear real enough. For the participants of this study, who evaluated the authenticity of 24 real and 24 fake videos, the identification of a fake was based on only a few factors regarding the main protagonist of the video, such as the head position, facial expressions, movements or the audio/video synchronicity. Even if these results may alleviate some concerns as to the deceptiveness of deepfakes, the question arises as to how far the results of the study can be applied to users who are not in a constrained laboratory situation consciously focusing on determining the veracity of a video, but consume media on a day-to-day basis? And furthermore, how will this change as deepfake technology continually improves?

In addition, several authors address challenges for journalists when dealing with deepfakes (Gutsche, 2019; Walker, 2019; Westerlund, 2019). It is assumed that it will become more difficult for journalists to establish the veracity of a source and there is a risk that deepfakes could undermine the authority of the media (Gutsche, 2019). Based on this, an interdisciplinary approach and training of journalism students could be useful. For example, students could be taught "forensic techniques" (see also technological regulation) for identifying deepfakes (Walker, 2019).

In summary, numerous risks and challenges are associated with deepfakes. First of all, deepfakes can be used to defraud people, for instance, by manipulating audio or visual material of private persons or businesspeople. Second, they carry the risk of damaging the reputation of organizations or individuals and the danger of humiliating individuals through pornographic deepfakes. Third, they could influence political campaigns and public opinion and erode trust in institutions. Fourth, they could deceive law enforcement. Finally, deepfakes could influence media coverage and cause a number of challenges for journalists.

4.3 Regulation of deepfakes

Given the focus of many studies in the field on the risks associated with deepfakes, we find a corresponding discourse on approaches to risk mitigation, especially in studies from a computer science or law perspective. Depending on the disciplinary approach, such studies focus on either legal or technological regulation of deepfakes.

An interdisciplinary framework called "R.E.A.L." attempts to manage risks caused by deepfakes (Kietzmann et al., 2019). According to the authors, the following aspects could help organizations when dealing with deepfakes: Record original content to assure deniability; Expose deepfakes early (by using deepfake detection technology); Advocate for legal protection (depending on the development of further legal regulations); Leverage trust to counter credulity (strengthen brand-customer relationship).

Besides legislation and regulation, the current discourse finds it essential to further develop technology as a means of identifying and preventing the spread of

deepfakes (Maras & Alexandrou, 2019; Westerlund, 2019). In addition, corporate guidelines and voluntary action as well as the education and training of the public about deepfakes are deemed important. From the perspective of media literacy, the importance of individuals using a variety of sources to get their news and to look upon the material with a questioning mind is also highlighted (Hall, 2018; Silbey & Hertzog, 2019).

4.3.1 Legal regulations

Several papers deal with the legal regulation of deepfakes (including Caldera, 2019; Citron & Chesney, 2018, Citron & Chesney, 2019; Hall, 2018; Silbey & Hartzog, 2019).

With regard to the regulation of abusive use of deepfakes, various laws can be examined, such as the right to protect one's own image. Although this enshrined right and copyright infringement regulation protect against the unlawful use of personal media, these laws have so far not been sufficiently tailored to the characteristics of deepfakes and are therefore not fully effective. A similar problem arises with the current criminal laws against unlawful pornography. An analysis of potential public corporations that might create deepfake regulations and mitigate damages (in the USA) showed that a new agency may be needed. In this way, the problems caused by deepfakes and problems posed by the evolution of technologies such as artificial intelligence could be addressed more effectively (Caldera, 2019).

In addition, the regulation of deepfakes in some cases might involve the military (if deepfakes play a role in armed conflicts), covert investigations (if foreign governments use deepfakes as a threat) and could be punished with economic sanctions (Citron & Chesney, 2018, 2019).

Since legal articles so far relate almost exclusively to the legal situation in the USA, to what extent the legal situations in other countries worldwide are able to deal with malicious deepfakes also requires further examination. For example, Farish (2020) analyzes whether it would be reasonable for English law to adopt California's publicity right. However, in practice this right does not seem an ideal standard for combating harmfully-used deepfakes because it turns out to be rather vague. In addition to federal laws, international regulations are required. For example, it may be possible that a deepfake is created in one country, disseminated on an online platform of another country and harms a person who is under the legal protection of a third country.

Moreover, following the legal perspective, it is recommended that organizations develop policies, in which they describe how to prevent fake news and how to use algorithms to prevent the spread of misinformation and disinformation (Hall, 2018).

All in all, several publications with a legal perspective have addressed the regulation of deepfakes. Even though there are no draft laws regulating deepfakes yet, the publications in this field suggest different possibilities to deal with them, including establishing new law agencies, working on international legal regulations, and developing economic sanctions and consequences in the event of privacy violations.

4.3.2 Technological regulations

A large number of publications in the field of computer science have been dedicated to the detection and use of deepfakes. Most studies assume potential harms caused by deepfakes and therefore focus on developing detection techniques (e.g., Afchar et al., 2018; Korshunov & Marcel, 2018; Li, Chang & Lyu, 2018; Yang et al., 2019) or solutions for tracing the source of the deepfake (Hasan & Salah, 2019). With regard to the detection of deepfakes, different techniques can be classified:

- Analyzing the foreground (fake face area) and the background (original area) of image swaps to detect deepfakes (Zhang, Zhao & Li, 2020)
- Keypoints – focusing on the detection of points of interest in an image such as sections with sudden contrast changes (Đorđević, Milivojević & Gavrovska, 2019)
- Facial expressions and head movements – e.g., by evaluating the relationship between light and shadow, the angle and blurring of facial features and the complexion (Agarwal et al., 2019; Güera & Delp, 2018; Yang et al., 2019)
- Analyzing the reflection and detail in teeth and eyes (Matern, Riess, & Stamminger, 2019)
- Inconsistent movements of the mouth (Korshunov & Marcel, 2018) or lip and mouth movements (Jafar, Ababneh, Al-Zoube, & Elhassan, 2020)
- Analyzing the mesoscopic properties of images (Afchar et al., 2018)
- Eye blinking – detecting blinking anomalies based on, for instance, the repetition number or period (Jung, Kim, & Kim, 2020; Li et al., 2018)
- Tracing a deepfake back to its original source through a block chain-based solution that can track multiple copies of data via smart contracts (Hasan & Salah, 2019)

In summary, the technical proposals can be seen as a way to regulate deepfakes, as they are intended to enable their identification and thus, for instance, support social networking sites in the decision to delete potentially harmful deepfakes. Although the listed possibilities seem very promising, it should be noted that some of them still need to be further developed because they show weaknesses, including, for example, methods that have difficulty in evaluating low quality images (e.g., Zhang et al., 2020). Further, some deepfake detection methods show weaknesses when the person in the video does not look directly into the camera (e.g., Agarwal et al., 2019), while limitations for verifying the blinking rate method could be mental illness or the dopamine activity of a person, since these can affect a person's blinking (e.g., Jung et al., 2019). Finally, some of the studies cannot be generalized due to their small data set (e.g., Đorđević et al., 2019).

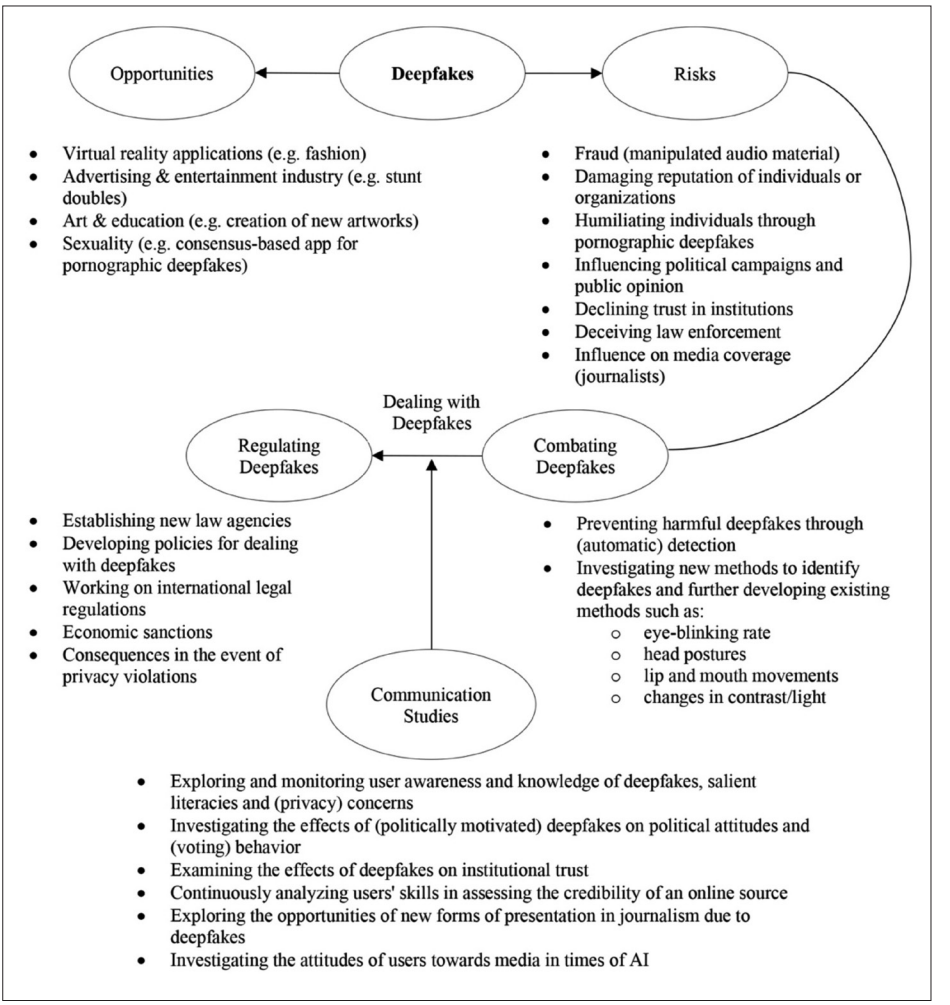
5. Conclusion

To summarize, we identify a young and therefore relatively small body of research on deepfakes, that nonetheless comprises a number of distinct discourses and key concepts. We find that deepfake research is dominated by studies from the fields of law and computer science – with both focusing strongly on risks as-

sociated with deepfakes and opportunities for mitigation through legal or technological regulation.

Our first research question (RQ1) addressed salient perspectives and key concepts in deepfake research. We find that research on deepfakes can be divided into five general categories (see Figure 1). As Figure 1 shows, there are several opportunities and risks associated with deepfakes. The risks of deepfakes raise the question of how to deal with them. This leads to the next step: Approaches to combat and regulate deepfakes as well as the potential impact of communication studies.

Figure 1. Model of dealing with deepfakes



In terms of risk mitigation, empirical research in the field of computer science focuses on AI experiments and the identification of deepfakes. To this end, methods are developed to assess the authenticity of video and audio recordings as well as

methods to trace deepfakes back to their original source. It should be noted that studies on deepfake detection focus exclusively on the determination of manipulated content but cannot shed light on intentions or causes. Even if generally accessible tools for the detection of deepfakes were to be developed, there is still insufficient basis for decision making and thus a lack of research examining when a deepfake video should be considered dangerous or harmful.

Furthermore, several publications focus on the possibilities of legal regulations when dealing with deepfakes, with the establishment of new law agencies and international laws to regulate deepfakes shown to be important. In this area, however, there is still a need for the development of precise conditions and consequences to deal with malicious deepfakes.

Our review showed that the majority of publications examined from these disciplines are conceptual articles and essays, while empirical research on deepfakes comes mostly from the field of computer science. There are only seven humanities and social science studies, of which five come from the field of communication. These publications focus on media users' skills in dealing with deepfakes, the effects of deepfakes on the users and the challenges created for journalists. Accordingly, we find very little social science research on deepfakes, leaving substantial opportunity for future research.

This leads to our second research question (RQ2), which focuses on implications for research on deepfakes from a communication studies perspective. First, it is important to focus more on the perspective of users and professional groups such as journalists. From the literature reviewed, it can be concluded that a central concern should be to analyze the user's knowledge of deepfakes, familiarity with detection options, and ability to recognize deepfakes. This angle of research touches upon media literacy, digital literacy and algorithmic literacy, as the ability to recognize deepfakes will surely differ between users. A related research gap emerges in the field of privacy studies, where deepfake applications, for instance, in the context of pornography, may affect privacy concerns and protection behavior.

Second, in regard to research on fake news indicating that mainstream news media play an important role in their dissemination, this finding can be applied to the dissemination of deepfakes, too. Accordingly, it can be assumed that deepfakes are more impactful if they reach a broad audience when shared by mainstream news media and influencers, and less dangerous if they are spread on rarely frequented disinformation outlets. Following the concern about deepfakes influencing the media coverage, professionals such as journalists also need to be supported in dealing with them. Therefore, an interdisciplinary cooperation between computer science and communication studies (i.e., computational communication studies) could be useful in order to explore how journalists can be supported in identifying deepfakes and how the impact of deepfake circulation can be regulated. At the same time, the focus should be on opportunities for new forms of presentation in journalism through deepfakes. In addition, the effects of deepfakes on users need to be continuously investigated. A continuous observation is necessary as users will increasingly become aware of the deepfake phenomenon, and thereby may become more sensitive to its detection, resulting in improvements in salient literacies. Given burgeoning research on fake news and disinformation, it seems

particularly important to consider the context of political news and election campaigns in deepfake research. As the risk of deepfakes influencing political campaigns and public opinion rises, the field of communication studies could investigate the effects of politically motivated deepfakes on recipients and their subsequent engagement with (online) media. Complementary to this line of questioning, it is also important to analyze to what extent politically motivated deepfakes have the potential to influence voting intentions and behavior.

Third, due to the continuous improvement of the quality of deepfakes, it is expected to become increasingly difficult for journalists to recognize the veracity of a source (Gutsche, 2019). As a result, if journalists are not able to identify a deepfake, they risk undermining the authority of the media by mistakenly sharing it with audiences. These dynamics in the dissemination and intermediation of deepfakes could result in decreasing institutional trust and rising media cynicism – especially among vulnerable groups such as the news-deprived, politically marginalized or lower educated individuals. Therefore, another aim of deepfake research in the field of communication studies should be to examine to what extent deepfakes undermine users' trust in institutions, including media brands and genres, political institutions and agents such as tech companies (providing detection tools) or fact-checking organizations. Similarly, the effect of deepfake awareness and perceptions on media cynicism warrants closer inspection.

Fourth, following fake news research by Wardle and Derakhshan (2017), it is important to analyze the creation, production and dissemination process of manipulated news content as well as its persuasive effect. A research gap that appears at this point is linked to the question of which characteristics make deepfakes particularly shareable and spreadable. Therefore, future research could analyze deepfakes as objects regarding their key characteristics. These could be formal aspects, such as their visuality and audibility, but also their narrative structure, such as length, editing, perspective and, in narrative terms, their relation to topics, events and persons that are of particular relevance for public communication. In addition, due to the spread of deepfakes through social networking sites, it is of interest to investigate the importance of intermediaries or influencers in the context of the dissemination of deepfakes as well as users' motives for spreading them.

Finally, the impact of AI technologies on the media sector needs to be studied broadly. Existing studies explore the automation of content generation and the replacement of human efforts by AI in journalism. AI technology provides opportunities for new forms of content generation, for illustrating stories and rendering news more accessible. However, the increasing use of AI-generated content may also negatively affect audience perceptions and journalistic brands. Media economics research is called upon to explore the structural effects and disruptive effects on business models.

6. Outlook

Since the literature review was conducted in May 2020, further studies in the social sciences have addressed the deepfake phenomenon, demonstrating the relevance of deepfakes to this field. For example, Gosse and Burkell (2020) analyzed

how news media characterize the problems presented by deepfakes. The findings of a discourse analysis of articles about deepfakes show that news media primarily discuss the phenomenon as a threat to audiences who might be misled by a deepfake and point out the hypothetical consequences of deepfakes for democratic discourse and national security. In contrast, the analysis revealed that the harm of malicious deepfakes due to the creation of non-consensual pornographic content (of women) received less attention from news media. In conclusion, both potential threats and harms should be seriously considered, and it is necessary to prepare for malicious deepfakes for both political and social processes.

A first step in this direction has been made in the study by Dobber, Metoui, Trilling, Helberger, and de Vreese et al. (2020), who created a deepfake of a politician and analyzed its effects on audiences' political attitudes in an online experiment differentiating between a microtargeted and an untargeted group. Results show that the deepfake led to a significantly more negative attitude towards the depicted politician among the untargeted group, while their attitude towards the politician's party remained similar compared to the control condition. The microtargeted group, for whom the deepfake's effects were predicted to be stronger than for the untargeted group, showed both a significantly more negative attitude towards the politician and his party (compared to the control condition). The study points to the impact of malicious political deepfakes and, in particular, microtargeting techniques that can intensify the effects of deepfakes on certain target groups.

Finally, in addition to the examination of deepfakes from the social sciences and especially communication and media studies, and in addition to the further development of deepfake detection and tracing, national and international legal regulations should be adopted, too. An interdisciplinary approach to the development of adequate regulatory responses seems to be appropriate. From the perspective of communication studies, the deepfake phenomenon is suitable for a media law, media economics and media technology approach.

Given our finding that most publications to date primarily focus on the risks of malicious deepfakes, future studies should also explore opportunities and potential advantages of this phenomenon. Initial studies hint at interesting applications not only in commercial contexts but in those of education or science communication. While it is important to be wary of dangers, emergent technologies such as deepfakes merit constructive criticism.

References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, December). Mesonet: a compact facial video forgery detection network. *IEEE International Workshop on Information Forensics and Security*, 1–7. Retrieved from: https://arxiv.org/pdf/1809.00888&hl=es&sa=X&scisig=AAGBfm3FjacexOtT3zbYNcdcxC4-TmJhmg&nossl=1&oi=scholarr

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting world leaders against deep fakes. *IEEE International Conference on Computer Vision and Pattern Recognition 2019 Workshop on Media Forensics,* 38-45. Retrieved from: http://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.html

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), 1–6. Retrieved from: https://advances.sciencemag.org/content/advances/6/14/eaay3539.full.pdf

Anderson, C. (2019, November 8). A new hermeneutics of suspicion? The challenge of deepfakes to theological epistemology. *Cursor_ Zeitschrift Für Explorative Theologie*. Retrieved from: https://cursor.pubpub.org/pub/andersondeepfakes

Bose, J. & Aarabi, P. (2019, September). Virtual fakes: DeepFakes for Virtual Reality. *IEEE 21st International Workshop on Multimedia Signal Processing*. https://doi.org/10.1109/MMSP.2019.8901744

Caldera, E. (2019). "Reject the Evidence of Your Eyes and Ears": Deepfakes and the law of virtual replicants. *Seton Hall Law Review*, *Vol. 50*(1). Retrieved from: https://scholarship.shu.edu/shlr/vol50/iss1/5/

CDEI – The Centre for Data Ethics and Innovation (2019). Deepfakes and audio-visualdisinformation. *CDEI Snapshot Series*. Retrieved from: https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai

Cinelli, M., Cresci, S., Galeazzi, A., Quattrociocchi, W., & Tesconi, M. (2020). The limited reach of fake news on Twitter during 2019 European elections. *PloS one*, *15*(6),1–13. https://doi.org/10.1371/journal.pone.0234689

Citron, D. K. & Chesney, R. (2018). Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review, Vol. 107,* 1753–1820. Retrieved from: https://scholarship.law.bu.edu/faculty_scholarship/640

Citron, D. K., & Chesney, R. (2019). 21st century-style truth decay: Deep fakes and the challenge for privacy, free expression, and national security. *Maryland Law Review, Vol. 78*(4), 882–891. Retrieved from: https://digitalcommons.law.umaryland.edu/mlr/vol78/iss4/5/

Delfino, R. (2019). Pornographic Deepfakes — revenge porn's next tragic act — the case for Federal Criminalization. *Fordham Law Review*, *Vol. 88* (3), 887–938. http://dx.doi.org/10.2139/ssrn.3341593

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2020). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *Vol. 26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Đorđević, M., Milivojević, M., & Gavrovska, A. (2019, November). Deepfake video analysis using SIFT features. *IEEE 27th Telecommunications Forum*, 1–4. https://doi.org/10.1109/TELFOR48224.2019.8971206

Douglas, H. (2019). Deepfakes: false pornography is here and the law cannot protect you. *Duke Law & Technology Review, Vol. 17*, 99–127. Retrieved from: https://scholarship.law.duke.edu/dltr/vol17/iss1/4/

European Parliamentary Research Service (EPRS) (2019). *Polarisation and the use of technology in political campaigns and communication*. Retrieved from: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634414/EPRS_STU(2019)634414_EN.pdf

Farish, K. (2020). Do deepfakes pose a golden opportunity? Considering whether English law should adopt California's publicity right in the age of the deepfake. *Journal of Intellectual Property Law & Practice*, *15*(1), 40–48. https://doi.org/10.1093/jiplp/jpz139

Fehrensen, M., & Täubner, M. (2019*)*. Warum wir nicht glauben sollten, was wir sehen. Gefälscht wurde schon immer. Sogenannte Deepfakes aber weisen in eine neue Dimension der Manipulation [Why we should not believe what we see. Counterfeiting has always

been used. But so-called deepfakes point to a new dimension of manipulation]. *Brand eins*. Retrieved from: https://www.brandeins.de/magazine/brand-eins-wirtschaftsmagazin/2019/gefuehle/social-media-warum-wir-nicht-glauben-sollten-was-wir-sehen?utm_source=zeit&utm_medium=parkett

Floridi, L. (2018). Artificial Intelligence, deepfakes and a future of ectypes. *Philosophy & Technology, Vol. 31*(3), 317–321. https://doi.org/10.1007/s13347-018-0325-3

Fortunati, L. & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in Human-Machine Communication. *Human-Machine Communication*, *Vol. 1*, 7–18. https://doi.org/10.30658/hmc.1.1

Franks, M. A. & Waldman, A. E. (2019). Sex, lies, and videotape: deep fakes and free speech delusions. *Maryland Law Review, Vol. 78*(4), 892–898. Retrieved from: https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=3835&context=mlr

Gosse, C. & Burkell, J. (2020). Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, *37*(5), 497–511. https://doi.org/10.1080/15295036.2020.1832697

Güera, D. & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. *IEEE 15th International Conference on Advanced Video and Signal Based Surveillance*. https://doi.org/10.1109/AVSS.2018.8639163

Gutsche Jr., R. E. (2019). The state and future of television news studies: theoretical perspectives, methodological problems, and practice. *Journalism Practice, Vol. 13*(9), 1034–1041. https://doi.org/10.1080/17512786.2019.1644965

Hall, H. K. (2018). Deepfake videos: when seeing isn't believing. *Catholic University Journal of Law & Technology, Vol. 27*(1), 51–76. Retrieved from: https://scholarship.law.edu/jlt/vol27/iss1/4

Hasan, H. R. & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *IEEE Access, Vol. 7*, 41596–41606. https://doi.org/10.1109/ACCESS.2019.2905689

Hillen, B. (2019, June 14). Adobe Research and UC Berkeley create AI that can find and undo portrait manipulations. *Digital Photography Review*. Retrieved from: https://www.dpreview.com/news/8837646038/adobe-research-and-uc-berkeley-create-ai-that-can-find-and-undo-portrait-manipulations

Jafar, M. T., Ababneh, M., Al-Zoube, M., & Elhassan, A. (2020, April). Forensics and analysis of deepfake videos. *IEEE 11th International Conference on Information and Communication Systems*, 53–58. https://doi.org/10.1109/ICICS49469.2020.239493

Jordan, M. I. (2019). Artificial Intelligence — the revolution hasn't happened yet. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.f06c6e61

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: deepfakes detection using human eye blinking pattern. *IEEE Access*, 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

Kaplan, A. & Haenlein, M. (2019). Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Khodabakhsh, A., Ramachandra, R., & Busch, C. (2019). Subjective evaluation of media consumer vulnerability to fake audiovisual content. *IEEE 11th International Conference on Quality of Multimedia Experience*, 1–6. **https://doi.org/**10.1109/QoMEX.2019.8743316

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

Korshunov, P. & Marcel, S. (2018). Deepfakes: a vew threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*. Retrieved form: https://arxiv.org/pdf/1812.08685.pdf

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: understanding and coping with the "Post-Truth" era. *Journal of Applied Research in Memory and Cognition, Vol. 6,* 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

Li, Y., Chang, M.-C., & Lyu, S. (2018, December). In ictu oculi: exposing AI created fake videos by detecting eye blinking. *IEEE International Workshop on Information Forensics and Security,* 1–7. https://doi.org/10.1109/WIFS.2018.8630787

Lichfield, G. & Li, H. (2020, 22 January). *Deepfakes: do not believe what you see*. [lecture at World Economic Forum Annual Meeting]. Retrieved from: https://www.weforum.org/events/world-economic-forum-annual-meeting-2020/sessions/deepfakes-seeing-is-believing

Liu, Y., Chen, W., Liu, L., & Lew, M. S. (2019). SwapGAN: a multistage generative approach for person-to-person fashion style transfer. *IEEE Transactions on Multimedia, Vol. 21*(9), 2209–2222. **https://doi.org/1**0.1109/TMM.2019.2897897

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: its impact on society and firms. *Futures*, *90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

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*, Vol. *23*(3), 255–262. https://doi.org/10.1177/1365712718807226

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. *IEEE Winter Application of Computer Vision Workshop,* 82–92. https://doi.org/10.1109/WACVW.2019.00020

Mattke, S. (2018, June 6). KI gegen KI: Wettbewerb zu Fälschung von Video-Inhalten [AI vs. AI: Competition on video content counterfeiting]. *Heise online*. Retrieved from: https://www.heise.de/newsticker/meldung/DARPA-veranstaltet-Wettbewerb-zu-Faelschung-von-Video-Inhalten-4074467.html

Metz, C. (2019, November 24). Internet companies prepare to fight the 'Deepfake' future. *The New York Times*. Retrieved from: https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html

Müller, P. & Denner, N. (2017). *Was tun gegen Fake News? Eine Analyse anhand der Entstehungsbedingungen und Wirkweisen gezielter Falschmeldungen im Internet* [What to do against fake news? An analysis based on the conditions under which fake news originates and the effects of targeted false reports on the Internet]. Berlin: Friedrich Naumann Stiftung für die Freiheit.

Öhman, C. (2019). Introducing the pervert's dilemma: a contribution to the critique of deepfake pornography. *Ethics and Information Technology 2019,* 1–8. https://doi.org/10.1007/s10676-019-09522-1

Raffaghello, I., Kastalio, L., Kalf, S., & Paisley, E. (2019). *What Does a feminist approach to deepfake pornography look like?* Retrieved from: https://mastersofmedia.hum.uva.nl/blog/2019/10/24/what-does-a-feminist-approach-to-deepfake-pornography-look-like/

Silbey, J. & Hartzog, W. (2019). The upside of deep fakes. *Maryland Law Review, Vol. 78,* 960-966. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3452633&dgcid=ejournal_htmlemail_intellectual:property:other:ejournal_abstractlink

Simonite, T. (2019, July 10). Most deepfakes are porn, and they're multiplying fast. Researchers worry that doctored videos may disrupt the 2020 election, but a new report finds that 96 percent of deepfakes are pornographic. *WIRED*. Retrieved from: https://www.wired.com/story/most-deepfakes-porn-multiplying-fast/

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital journalism, 6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. *Annals of the International Communication Association, Vol. 44*(2), 1–17. https://doi.org/10.1080/23808985.2020.1759443

Turton, W. & Martin, A. (2020, January 7). How deepfakes make disinformation more real than ever. *Bloomberg*. Retrieved from: https://www.bloomberg.com/news/articles/2020-01-06/how-deepfakes-make-disinformation-more-real-than-ever-quicktake

Vaccari, C. & Chadwick, A. (2020). Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, *6*(1), 1–13. https://doi.org/10.1177/2056305120903408

Vincent, J. (2019, May 17). This AI-generated Joe Rogan fake has to be heard to be believed. The most realistic AI voice clone we've heard. *THE VERGE*. Retrieved from: https://www.theverge.com/2019/5/17/18629024/joe-rogan-ai-fake-voice-clone-deepfake-dessa

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, Vol.3*(1), 32–46. https://doi.org/10.1515/opis-2019-0003

Walker, A. S. (2019). Preparing students for the fight against false information with visual verification and open source reporting. *Journalism & Mass Communication Educator, Vol. 74*(2), 227–239. https://doi.org/10.1177/1077695819831098

Wardle, C. & Derakhshan, H. (2017). Information disorder: toward an interdisciplinary framework for research and policy making. *Council of Europe report*, *27*. Retrieved from: https://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf

Westerlund, M. (2019). The emergence of feepfake technology: a review. *Technology Innovation Management Review*, *9*(11), 39–52. https://doi.org/10.1080/13683500.2020.1738357

Yadav, D. & Salmani, S. (2019, May). Deepfake: a survey on facial forgery technique using generative adversarial network. *IEEE International Conference on Intelligent Computing and Control Systems*, 852–857. **https://doi.org/**10.1109/ICCS45141.2019.9065881

Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. *IEEE International Conference on Acoustics, Speech and Signal Processing, 8261–8265*. **https://doi.org/**10.1109/ICASSP.2019.8683164

Zhang, W., Zhao, C., & Li, Y. (2020). A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. *Entropy*, *22*(2), 249. https://doi.org/10.3390/e22020249