The Emergence of Deepfakes and its Societal Implications: A Systematic Review

Dilrukshi Gamage
Department of Innovation Science,
Tokyo Institute of Technology,
Tokyo, Japan
dilrukshi.gamage@acm.org

Jiayu Chen
Department of Psychology and Human Developmental Sciences,
Nagoya University,
Nagoya, Japan
chen.jiayu@h.mbox.nagoya-u.ac.jp

Kazutoshi Sasahara
Department of Innovation Science,
Tokyo Institute of Technology,
Tokyo, Japan
sasahara.k.aa@m.titech.ac.jp

Abstract

The appearance of Deepfake tools and technologies in the public is proliferating. Scholarly research is very centered on technology of deepfake but sparse in understanding how the emergence of deepfakes impacts society. In this systematic review, we explored deepfake scholarly works that discuss societal implications than the technology-centered focus. We extracted studies from major publication databases - Scopus, Web of Science, IEEEXplore, ACM Digital Library, Springer Digital Library and Google Scholar. The corpus reflects patterns based on their research methodologies, area of focus, and the distribution of such research. Out of 787 works, 88 were highly relevant, with the majority of the studies being reviews of the literature. While research focus is generally drawn upon exploring security related harms, less focus is put on issues such as ethical implications and legal regularities for areas other than pornography, psychological safety, cybercrimes, terrorism, and more. The field research for Deepfake social impact research is emerging and this paper brings more insights drawn from a methodical, subject focused and distribution point of view.

1 Introduction

The rapid development of technologies such as Artificial Intelligence (AI) and Deep Learning (DL) revolutionized the way we create and consume content. As a byproduct of this revolution, we witness emerging technologies such as Deepfake which may potentially harm and distress social systems. Deepfakes are synthetic media generated using sophisticated algorithms which reflect things that did not happen for real but computer generated for manipulation purposes (Westerlund, 2019). In many cases, specific methods of Deep Learning which involve training generative neural networks — autoencoders, Generative Neural Network (GNN) in Machine Learning (ML) are utilized to generate these synthetic media.

Currently, a myriad of scholarly works concentrate on specific Deep Learning techniques — types of neural network model in which the model is trained to restore (copy) the input data known as auto encoders, GAN Models that involves a generator and discriminator in building an image closer to the original, High-definition face image generations, Conditional GANs (CGAN) that generate data while controlling attributes by giving attribute information in addition to images during training, face swapping techniques and speech synthesizing techniques (Guarnera et al., 2020). These studies are more influenced by the Deepfake generation and detection methods. However, the advancements of these scholarly works and the democratization of these technologies made it easy for any individual to generate realistic fake media content which could have been difficult previously. Apart from the incident that incepted Deepfake in 2017 where celebrity faces were used to create phonographic videos using Deepfake technologies (Burkell and Gosse, 2019), the incidences such the British energy company scammed by voice Deepfake technology (Stupp, 2019) in 2019 and recently the arrest of a Japanese student for posting pornographic videos that synthesized the face of a celebrity using Deepfake technology by training the model for about a week, using 30,000 images per video where the case is believed to be the first criminal case in Japan which Deepfake technology was abused (Times, 2020) can be highlight as emerged abuse of using Deepfakes. In addition to these, more recently(March 10th 2021), a mother in Pennsylvania used Deepfake technology to forge photos and videos to show drinking, smoking and nakedness to trap a teammate of a high school daughter who works as a cheerleader (Guardian, 2021) and the article written in

Newyorker inquires ethical implications of Deepfake voice by narrating the movie about celebrity chef Anthony Bourdain in July '15th, 2021 (Ronsner, 2021). Together, all such incidences have demonstrated the emerging threats unresting the social process.

Although Deep Learning technologies are versatile and could be useful in revolutionizing various industries, these incidents collectively raise concerns about the societal problems emerging from them. There is ample work in computer science on automatic generation (Yadav and Salmani, 2019; Caldelli et al., 2021) and detection of Deepfakes (Maksutov et al., 2020; Rana and Sung, 2020), but to date there are only a handful of social scientists who have examined the social impact of Deepfake technology. In this paper, we conducted a systematic literature review to understand the existing landscape of research that examines the possible effects Deepfakes might have on people, to understand the psychological dynamics of deepfakes and to discover how it impacts society. In particular, we hope to examine the following two research questions:

- Q1: What types of research conducted between 2017-2021 to understand the psychological and social dynamics and societal implications of Deepfake?
- Q2: What is the distribution of Deepfake research between 2017-2021 that explores any type of psychological dynamics and its societal implications?

The objective of this systematic study is to highlight the types of research carried out to understand the social dynamics of Deepfake and identify any gaps in the researches that need further discussions on social implications and concerns that arise from the technology. This exploration of research related to social processes and the implications of Deepfake will provide necessary projections, and point to scholarly work in this area where social scientists could make a useful contributions by understanding any lack of new directions. Since deepfake attributes in Deep Learning and Machine Learning, much advancement and research has occurred in the field of computer science. In addition, with the democratization of accessible technology to a wider audience, necessary attention is paramount in order to understand the societal implications of this phenomenon.

Search Database	Hits	Selected
Springer Online Database	177	17
IEEE	154	11
ACM	264	8
Web of Science	137	41
Scopus	55	2
Other (Google Scholar)	NA	9
Total	787	88

Table 1: Summary of the results retrieved by running the search query and manually filtering by reviewing according to the inclusion criteria.

2 Methods

We obtained articles for our systematic review by searching popular scientific search engines and repositories—Springer Digital Database,IEEEXplore, ACM Digital Database, Web of Science, and Scopus. Most systematic reviews incorporate Preferred Reporting Items for Systematic reviews and Meta-Analyses protocols (PRISMA) explained in details by Moher et al. (2015).We followed a similar structure to this literature review with particular interest in understanding the two previously mentioned research questions. We used the following search query in all 5 databases and in addition to this, used Google Scholar to search any other relevant preprints or non-peer reviewed articles to bring more inclusively to the research which may not have been listed in ACM, Scopus, IEEE, Web of Science or any other database.

{Deepfake OR Artificial Intelligence}
AND Misinformation

We did not restrict our search to only journal papers, but allowed any peer reviewed paper, or commentary in an article, critical review or even work-in-progress papers including the preprints. After the search terms provided the dataset, we used two experienced researchers to filter the research based on an inclusion criteria, we were particularly careful to select the results only if the manuscripts examined perceptions of Deepfake or its impact to human interaction or discussed the social implications of Deepfakes. In other words, articles that discussed a pure technology perspective (such as GAN), or studies to find new techniques for Deepfake detection's were eliminated as irrelevant to this study. Figure 1 describes the process conducted to obtain the relevant data to the analysis.

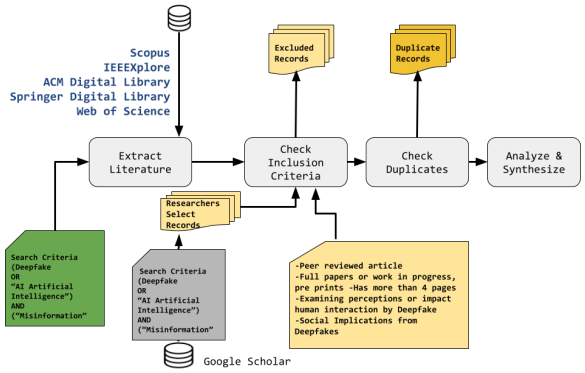

Figure 1: Flow of the systematic review

2.1 Dataset

Our initial search query extracted 787 articles from 5 databases. The extracted results were then combined to a single data file and two researchers collectively further filtered based on the inclusion criteria depicted in Figure 1 by manually reviewing the abstracts. In addition to these filtered articles, additional papers were added based on the relevant research found by Google Scholar and we labeled this source as "Other". Although a Google Scholar advance search returned 3420 hits, given the depth and spread of the articles we focused only on the first 20 pages which had 200 hits and selected 9 highly relevant papers not included in any databases.Out of these, 4 papers were from journals and, 2 universities repositories which was not listed in any of the 5 databases. Another 2 were preprints and currently under review, 1 commentary from Nature. We found 79 highly relevant papers from the 5 original databases and with the Google Scholar results had 88 papers selected for analysis.A breakdown is depicted in Table 1.

2.2 Measures

To answer RQ1, we analyzed all 88 papers using their full text, summarized the key phrases, highlighted major findings in the respective papers and identified any themes under which the article could be categorized. Based on the summary and key phrases, it was evident that the corpus can be categorized by a common methodological standpoint. For example, we realized that each article can be categorized by whether it conducted an experiment to understand social dynamics or had any sort of methodical analysis to understand social impact or if it was produced as a result of an extensive critical review by positioning any premises or even if it provided a conceptual proposal or framework beyond the review of the Deepfake social phenomenon. At the same time, we also examined whether or not the corpus focused on several domain areas addressing Deepfake social issues. We incorporated word clouds on each abstract to support subjective judgment on categories and focus areas.

To answer the RQ2, on the distribution of research in Deepfake psychological dynamics and its societal implications, we described descriptive statistics with a network analysis that understands the connections with its type of research and emphasis. At the same time, to highlight the emphasis of the paper, we highlighted the generated word clouds, specifically depict the categorical flows based on the frequencies, and used the network diagram using Gephi software to illustrate the author distributions among the selected papers.

3 Results and Discussion

Overall,the majority of the results from the query resulted scholarly work related to Deep learning, AI and ML learning technologies, and its improvements in creating or detecting Deepfake. Only 88 out of 787 were selected as those research works were found to be discussing the psychological dynamics, social implications, harms to the society, ethical standpoint, and or solutions from the a social-technological point of view.

3.1 RQ1: Types of research

Examining the abstracts and full text of the articles, we identified that each article could be categorized based on 11 types of research—Systematic review, Review based on Literature, Philosophical mode of enquiry, Examines, Experiment, Network Analysis, Content Analysis, Design, Conceptual Proposal, Commentary and Analysis by Examples. Although these categories are based on the subjective judgment of the authors, it provides a solid understanding to the conducted research based on its main objectives and methods.

A magnified view of this dataset (88) revealed that the majority (30) of the papers focused on critical reviews based on the previous literature and slightly above half of the papers (21) conducted active experiments using real users to explore the social and psychological dynamics of perceiving Deepfakes or understanding their impact. Only one study was performed a network analysis based

on Deepfake discourse and limited other research papers focused on rest of the methods as depicted in Figure 2. Apart from the methodology point of view, we also derived key categories of the papers based on its focus area. Although our key interest centered upon Deepfake and its social impact, we observed that these relevant research covered a wider range of focus areas in different subject domains. These areas ranged from security aspects, pornography, legal concerns, Deepfake media, specifically video and images, psychological perspectives, political perspective, human cognition perspectives, and more. Therefore, to specifically answer RQ1, we describe the details of these methodologies and focus areas in the following sections.

Methodology used in Deepfake social implication research

Although methodical approaches for research are not new, our analysis of the 88 highly relevant papers for the social or psychological implications of Deepfake reflected that most of the research in this domain is still developing and many researchers are critically evaluating and analyzing Deepfake phenomena from the previous literature, discussing potential future outcomes. We categorized this type of research as **Review based on Literature** and from our corpus, the earliest research on critical reviews of Deepfake social implications occurred in 2019 (although the term "Deepfake" first time in 2017 (Westerlund, 2019)). Research by Westling (2019) raise questions about to understanding whether the Deepfake phenomena is shallow or deep and how society might react to these technologies. Specifically the paper critically analysed and predominantly provided nuances to the technology that generates deep fake media and its uses, showing that society has never relied solely on the content as a source of truth.

Similarly Antinori (2019) provides an extensive narration to Deepfake and relates its consequences to terrorism. The author does not follow a systemic approach, however there is a critical discussion of the Deepfake focus on the near future of security threats by using examples of previous literature and emphasizing the need of awareness, law enforcement, and policymakers to implement effective counter terrorism's strategies. While providing this background and previous work, the author also articulates his stance on the subject emphasizing that as a globalized community, we are transitioning from e-terrorism to upcoming online terrorism, as well as the linearity to hypercomplexity by malicious use of AI and living in the post-truth era of a social system. Since his research article not only provides critical review based on past literature but also the authors theoretical and qualitative research experience with participation and working as a counter terrorism expert in related projects, we also intersected this with a new category: **Examines**. Through our full-text analysis, we observed that many other **Review based on Literature** scholarly work intersects with the **Examines** category. In these types of articles, we observed authors critically providing their experience or using their point of view as a metaphor to build constructs. All together we found 11 out of 30 papers categorized as Review based on Literature illustrated this intersection. For example, the review article by Hancock and Bailenson (2021), attempts to understand the possible effects Deepfakes might have on people, and how psychological and media theories apply. In addition, the article by Öhman (2019) brings a philosophical mode of enquiry to a pervert's dilemma, an abstraction about fantasizing sexual pornography and argues that ethical perspectives underline dilemmas by using the literature and theories. Similar placement of arguments and concepts supported by review of literature can be found in articles by Taylor (2021), Kerner and Risse (2021), Langa (2021), Ratner (2021), Harper et al. (2021), Langguth et al. (2021) and (Greenstein, 2021). However, we also derived 4 research articles that falls in the category of **Examines** without a dominating critical literature review—For example, an article was compiled while examining US and British legislation indicating legislative gaps and inefficiency in the existing legal solutions and presenting a range of proposals of legislative change to the constitutional gaps in porn (Mania, 2020). The article examines current online propaganda tools in the context of the different information environment and, provides examples of its use, while seeking to educate about Deepfake tools and the future of propaganda (Pavlíková et al., 2021). Another study examines the problem of unreliable information on the internet and its implications for the integrity of elections, and representative democracy, in the U.S. (Zachary, 2020) and another study that addresses the economic factors

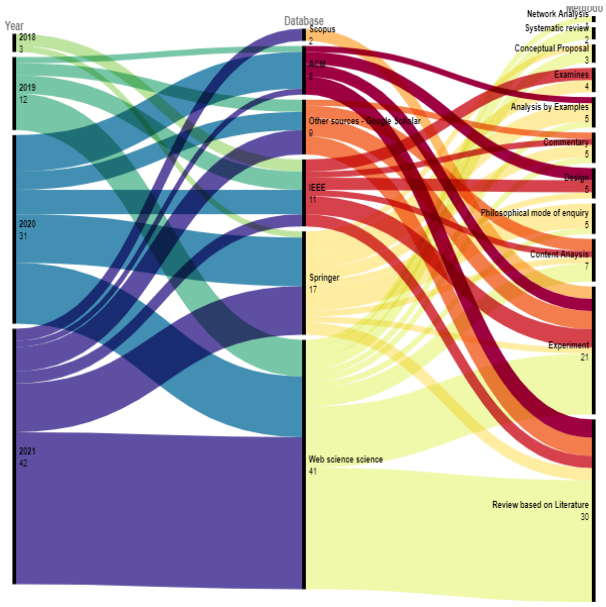

Figure 2: Scholarly work distribution based on the year it was published, the Published databases and its Methodology

that make confrontational conversation more or less likely in our era and brought viewpoints in the Deepfakes which becoming more widespread on the dark web (Greenstein, 2021) are falling in to this *Examines* category.

However, alongside review based articles and articles that conducted extensive examinination, we also derived another category. Although this category is similar to the methods we previosuly stated, it is distinguished by the way it positions its point of views. We noticed that this type of articles is extensively based on use cases, examples of incidences or more descriptions of theoretical and informational AI and Deepfake technologies. We name this category **Analysis by Example** and found 5 papers fall under its umbrella. Articles in this category includes Pantserev (2020), through their examples of Deepfakes in the modern world, and the internet-services, Amelin and Channov (2020) study the use of legal regulation in use of facial processing technologies, and Caldwell et al. (2020) study possible applications of artificial intelligence and related technologies in the perpetration of crimes, Degtereva et al. (2020) studied the general analysis of risks and hazards of the technologies and analysis examples of legal remedies available to victims. We also identified a category named **Philosophical Mode of Enquiry** which includes papers that use a philosophical point of view in premising their enquiry to the social issues found with in the Deepfake applications (Öhman, 2019; Ziegler, 2021; Floridi, 2018; Hazan, 2020; Kwok and Koh, 2021).

However, since the developments in the area of social implications of Deepfakes are yet growing, we observed only 2 **Systematic Review** types of research that explain in detail of the growing body of literature and its systematic analysis (Godulla et al., 2021; Westerlund, 2019). The first systematic review used English-language deepfake research to identify salient discussions; and the other used 84 publicly available online news articles to examine what deepfakes are and who produces them, and the benefits and threats of deepfake technology in 2021 and 2019 respectively. However, apart from these critical reviews, examiner papers, analysis by examples and systematic reviews, we found one other methods that could be classifed into the same theme but distinct in its narration of the information as it is made as a personal opinion or commentary to certain events. We named this category as **Commentary Bases** which often provides short narrative for the question of the future of technological implications (Kalpokas, 2021; LaGrandeur, 2021; Beridze and Butcher, 2019; Strickland, 2018, 2019).

As a next category of methodology, we observed that 21 out of 88 papers depicted some sort of experiment using human subjects to understand any impact and social implications of Deepfake and we named this category **Experiment**. In this category we observed researchers such as Khodabakhsh et al. (2019) used 30 users to examine human judgment on Deepfake videos, Caramancion (2021) used 161 users to explore the relationship between a person's demographic data, political ideology and the risk of him/her falling prey to Mis/Disinformation attacks. The largest study conducted by Yaqub et al. (2020) used 1,512 users to explore the impact of four types of credibility indicators on people's intent to share news headlines with their friends on social media. Similarly, Dobber et al. (2021) studied effects on political attitudes using 271 users, Köbis et al. (2021) studied the inability of people to reliably detect Deepfakes using 210 users. Their research particularly found neither by educating or introducing financial incentives improves their detection accuracy experimented and many other similar studies contained in this category. Apart from experiments, we also found research articles proposing frameworks or

solutions to Deepfake societal issues by conceptualizing theoretical frameworks (Cakir and Kasap, 2020; Kietzmann et al., 2020b,a) named as **Conceptual Proposals**. Beyond conceptual proposals, we also found that some articles consisted clear design goals with implementation plans or some some artifacts designed as solutions to the issues of Deepfake societal issues (Chi et al., 2020; Qayyum et al., 2019; Chen et al., 2018; Sohrawardi et al., 2019; Inie et al., 2020). Thus we introduced a category named **Design**.

Apart from such dominated methods to observe social implications and perceptions of Deepfakes, we also found 7 articles that followed the **Content Analysis** method. Three used Twitter data as their corpus (Maddocks, 2020; Oehmichen et al., 2019; Hinders and Kirn, 2020) and two studies analyzed the article content in news media (Brooks, 2021; Gosse and Burkell, 2020); each study conducted analyses using YouTube comment discourses about Deepfakes (Lee et al., 2021) and journalist discourse (Wahl-Jorgensen and Carlson, 2021) to understand the social implications of the Deepfakes phenomenon. Althogh, similar to these studies, we categorized one more study as **Network Analysis** and it conducted semantic content analysis using Twitter data relating to Deepfake phenomena (Dasilva et al., 2021) to understand the social discourse.

Range of focus areas examining Deepfake and its social implications

Apart from the key categorization towards research methods, we examined the significant research questions these research methods are used to solve. This aids us in categorizing the Deepfake social research based on the subject areas which it is focused. We derived 30 main focus areas these research articles primarily concentrate on, followed by 44 sub-focused areas. This flow is graphically represented in the alluvial diagram in Figure 3. At the interest of space for this paper, we highlight the top 5 focus areas of research.

As it appears, the highest interest of focus is drawn upon **Security** related issues relating to the social implications of Deepfakes. A significant number of research relating to security are foreseeing harms and threats to the society through "Review of literature" (Repez and Popescu, 2020; Taylor, 2021; Kaloudi and Li, 2020; Rickli and Ienca, 2021). More security focus research is conducted based on a "Design" of a blockchain-based

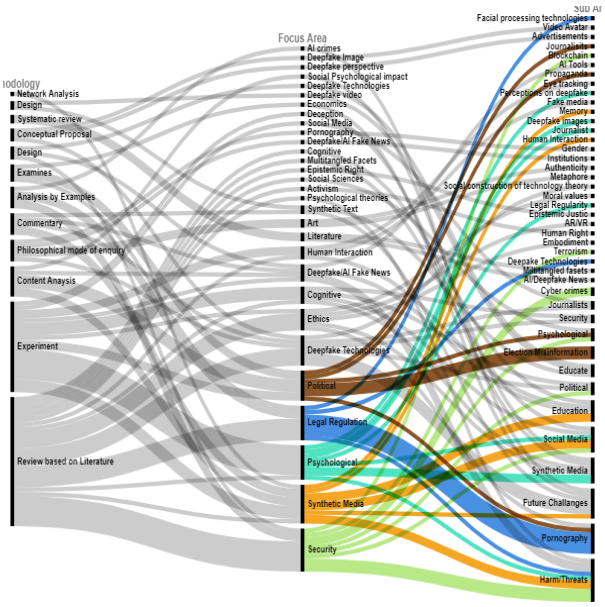

Figure 3: All 88 papers are categorised based on its main methodological and focus are of the research. Highlighted in color are the first five focus areas based in the higher frequency - Security, Synthetic Media, Psychology,Legal Regulation and Political are the top 5 focus areas.

framework for preventing fake news while introducing various design issues (Chi et al., 2020). At the same time security focus research has been visible in the research method of "Analysis by Example" where Degtereva et al. (2020) conduct a general analysis to understand the risks and hazards of the technologies used today and highlight the need for a wider application and enhancement of Deepfake technology to fight Cyber Crimes. Similarly, Pantserev (2020) analyses a wide range of examples of deepfakes in the modern world and the Internet-services that generate them with a key focus on security. Their research also depicts a clear sub focused area of **Psychological Security** as they try to understand the threats Deepfake cause to society and its impacts.

The next highest focus area of literature solves problems relating to "Synthetic Media." These are mostly considered as the Deepfake in the the mode of Videos. We observed that most researchers have used **Synthetic media** to conduct "Experiments" and "Content Analysis." For instant, Iacobucci et al. (2021) test whether a simple priming of deepfake information significantly increases users' ability to recognize Synthetic media, Hwang et al. (2021) examined the negative impact of deepfake video and the protective ef-

fect of media literacy education; and Murphy and Flynn (2021) examined how Deepfake videos may distort the memory for public events, yet found it may not always be more effective than simple misleading text. Other than these, Brooks (2021) used "Content Analysis" to analyze popular news and magazine to understand impact of Synthetic media. Interestingly, the article argues, that if fake videos are framed as a technical problem, solutions will likely involve new systems and tools or if fake videos are framed as a social, cultural, or as an ethical problem, solutions needed will be legal or behavioral ones. On the other hand, in this article, the focus of Synthetic media also expand to the sub focus to examine the societal **Harm/Threats**. Similarly, Hinders and Kirn (2020), empathize that digital photos are so easy to manipulate, yet deepfake videos are more important to understand as deepfake synthetic media (video evidence) could be deliberately misleading and not easy to recognize as fake. Apart from content analysis, focus on synthetic media narrowed the focus for a few commentary based articles: one examines Deepfake video implications on Facebook (Strickland, 2019), and two other articles focus examining Deepfake videos challenges with a sub focus on understanding **Future Challenges** (Kalpokas, 2021; LaGrandeur, 2021).

The next highest set of research articles focus mainly on the areas of **Psychological, Legal Regulation, and Politics**. Interestingly, all **Psychological** focus research conducted as experiments except for one that focuses on the Psychological impact of Deepfake through a review of literature (Hancock and Bailenson, 2021). In experiments, Yaqub et al. (2020) explore the effect of credibility signals and how they perceived any individual to share fake news Khodabakhsh et al. (2019) focus on understanding the vulnerability of Human judgement to Deepfake. Ahmed (2021b) examines the social impact of Deepfakes using an online survey sample in the United States. This investigates psychological aspects of the impact of Deepfake while examining the concerns of citizens regarding deepfakes, exposure to deepfakes, inadvertent sharing of deepfakes, the cognitive ability of individuals, and social media news skepticism. Cochran and Napshin (2021) provided psychological aspects of Deepfakes by exploring factors impacting the perceived responsibility of online platforms to regulate deepfakes and pro-

Figure 4: Word clouds from abstracts identified as focusing Pornography (top) and in all articles (bottom)

vide implications for users of social media, social media platforms, technology developers, and broader society. The research focusing on **Legal Regulation** extensively worked on Deepfake pornography, discussing its ethical perspective, consequences, and legal framework to take action (i,e '(Karasavva and Noorbhai, 2021; Delfino, 2020; Gieseke, 2020). Few others had sub-focus on discussing the threats and harms (O'Donnell, 2021), Terrorism (Antinori, 2019) and specific to facial processing technologies (Amelin and Channov, 2020). The **Political** focus researches have been extensively worked on election related consequences of Deepfakes and few focused on the journalists discourse to shape political context (Wahl-Jorgensen and Carlson, 2021), explored the relationship between political and pornographic deep fakes (Maddocks, 2020) and discussed the threat of Deepfake online propaganda tools (Pavlíková et al., 2021).

3.2 RQ2: Distribution of the research

In the previous sections, we partially stated the distributions of research methods and focus areas by utilizing Figure 2 and 3. Further, we expanded the knowledge of the landscape for Deepfake research that concentrates on its societal impacts by examining the yearly distribution of the relevant research. As depicted in Figure 2, the yearly projection reflects a trend for studies which explore the social implications by Deepfake are emerg-

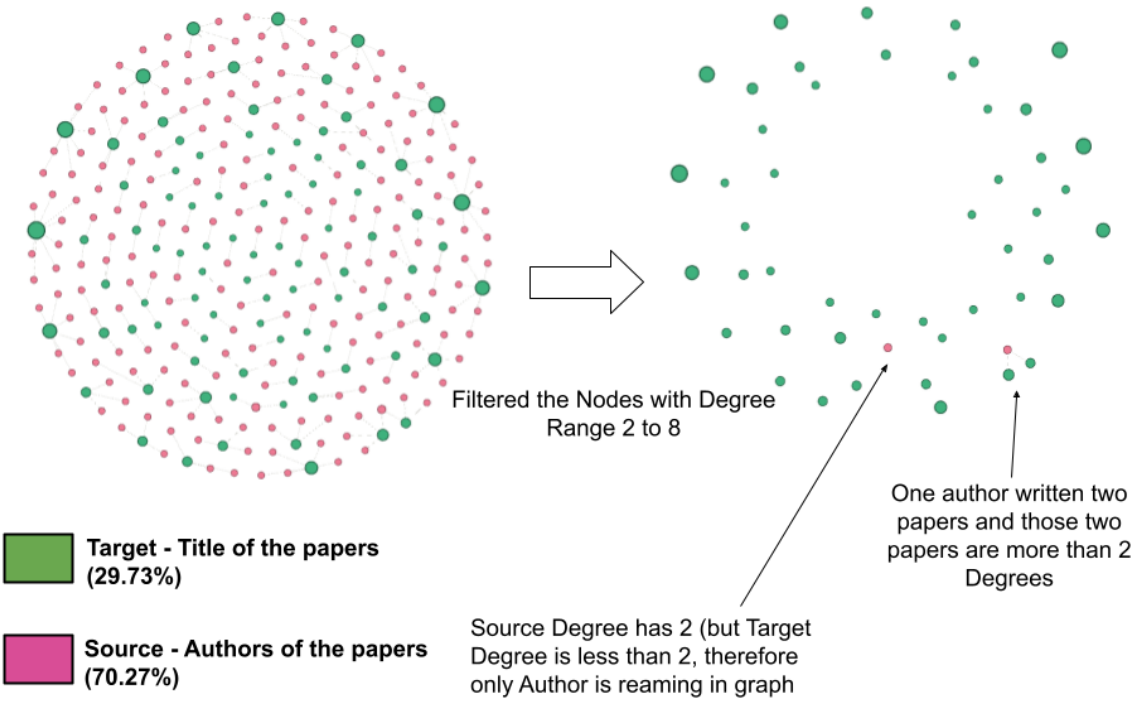

Figure 5: [Left] A bipartite graph created using source as the authors and targets as the papers. [Right] The Bipartite graph filtered based on the degree centrality larger than 2.

ing since 2019 and 2021 has the highest number of such researches(42) even before the year 2021 ends.

We generated word clouds for each abstract and one common word cloud combining all 88 abstracts to make sense of what we examined and to summarize the analysis of the full text of the articles. The top word cloud in the Figure 4 generated from a abstracts which we categorized as Pornography (Gieseke, 2020) and it hows its words are cantered on pornography; The bottom shows the word cloud from all abstracts which reflects Deepfake as the central theme and yet highlights, other focus areas we identified that greatly resonated in our categorizations. Finally, to better understand the distribution of the authors of these papers, we generated bipartite networks using the author list with the titles of the papers they have written (Figure 5). Nodes represent the authors (pink), papers (green), and the edges point from the authors to the papers. It appears that researchers who explore Deepfake social implications are almost not connected to each other as the clustering coefficient indicates 0.0 and nearly 30% of Papers written by 70% of authors and the highest number of relationship consisted one degree as a single author has written the papers. Ranked by the degree centrality (how many authors written how many papers), the graph revealed the lowest degree centrality as 1 and the highest as 8. Filtering the network to reflect if there are any 2 or more authors collaborated in writing these social research types we filtered the graph into 2 to 8 degree centrality. Interestingly, this resulted only two authors had 2 degrees relationship. in one instance, the same author wrote two different papers while collaborating with multiple other authors (Kietzmann et al., 2020b,b); in the other instance the same author has written two papers without any author collaborations (Ahmed, 2021c,a).

4 Conclusions

Our study reflects a comprehensive review of Deepfake research which discusses the social implications of Deepfake as the primary focus opposed to the reviews to the technology itself. We selected 88 highly relevant papers to our study and based on the methodical aspects, we found 11 types of studies that could be categorized. Out of all 88 papers, we also found that majority of studies focus on research relating to security and discuss the possible harms and threats to the social

echo system. Much debated issues such as ethical implications to Deepfake, the regulatory or legal solutions other than pornography, such as making awareness or educative activism to other type of harm specially, the cyber crimes and terrorism are much sparse in the landscape. Our results suggest that the social science of Deepfakes is emerging, but such research has been conducted independently thus far. Given that Deepfakes and related AI technologies are weaponizing, the social implications of Deepfakes should be more investigated with an interdisciplinary effort.

Acknowledgments

This work is generously supported by JST, CREST Grant Number JPMJCR20D3, Japan.

References

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

Saifuddin Ahmed. 2021b. Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *new media & society*, page 14614448211019198.

Saifuddin Ahmed. 2021c. Who inadvertently shares deepfakes? analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, 57:101508.

Roman Amelin and Sergey Channov. 2020. On the legal issues of face processing technologies. In *International Conference on Digital Transformation and Global Society*, pages 223–236. Springer.

Arije Antinori. 2019. Terrorism and deepfake: From hybrid warfare to post-truth warfare in a hybrid world. In *ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics*, page 23. Academic Conferences and publishing limited.

Irakli Beridze and James Butcher. 2019. When seeing is no longer believing. *Nature Machine Intelligence*, 1(8):332–334.

Catherine Francis Brooks. 2021. Popular discourse around deepfakes and the interdisciplinary challenge of fake video distribution. *Cyberpsychology, Behavior, and Social Networking*, 24(3):159–163.

Jacquelyn Burkell and Chandell Gosse. 2019. Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday*.

Duygu Cakir and Özge Yücel Kasap. 2020. Audio to video: Generating a talking fake agent. In *International Online Conference on Intelligent Decision Science*, pages 212–227. Springer.

Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. 2021. Optical flow based cnn for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, 146:31–37.

M Caldwell, JTA Andrews, T Tanay, and LD Griffin. 2020. Ai-enabled future crime. *Crime Science*, 9(1):1–13.

Kevin Matthe Caramancion. 2021. The demographic profile most at risk of being disinformed. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–7. IEEE.

Weiling Chen, Chenyan Yang, Gibson Cheng, Yan Zhang, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. 2018. Exploiting behavioral differences to detect fake ne. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 879–884. IEEE.

Hongmei Chi, Udochi Maduakor, Richard Alo, and Eleason Williams. 2020. Integrating deepfake detection into cybersecurity curriculum. In *Proceedings of the Future Technologies Conference*, pages 588–598. Springer.

Justin D Cochran and Stuart A Napshin. 2021. Deepfakes: awareness, concerns, and platform accountability. *Cyberpsychology, Behavior, and Social Networking*, 24(3):164–172.

Jesús Pérez Dasilva, Koldobika Meso Ayerdi, Terese Mendiguren Galdospin, et al. 2021. Deepfakes on twitter: Which actors control their spread? *Media and Communication*, 9(1):301–312.

Viktoria Degtereva, Svetlana Gladkova, Olga Makarova, and Eduard Melkostupov. 2020. Forming a mechanism for preventing the violations in cyberspace at the time of digitalization: Common cyber threats and ways to escape them. In *Proceedings of the International Scientific Conference-Digital Transformation on Manufacturing, Infrastructure and Service*, pages 1–6.

Rebecca A Delfino. 2020. Pornographic deepfakes: The case for federal criminalization of revenge porn's next tragic act. *Actual Probs. Econ. & L.*, page 105.

Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 26(1):69–91.

Luciano Floridi. 2018. Artificial intelligence, deepfakes and a future of ectypes. *Philosophy & Technology*, 31(3):317–321.

Anne Pechenik Gieseke. 2020. " the new weapon of choice": Law's current inability to properly address deepfake pornography. *Vand. L. Rev.*, 73:1479.

Alexander Godulla, Christian P Hoffmann, and Daniel Seibert. 2021. Dealing with deepfakes–an interdisciplinary examination of the state of research and implications for communication studies. *SCM Studies in Communication and Media*, 10(1):72–96.

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

Shane Greenstein. 2021. The economics of confrontational conversation. *IEEE Micro*, 41(2):86–88.

The Guardian. 2021. Mother charged with deepfake plot against daughter's cheerleading rivals.

Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2020. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667.

Jeffrey T Hancock and Jeremy N Bailenson. 2021. The social impact of deepfakes. *Cyberpsychology, behavior and social networking*, 24(3):149–152.

Craig A Harper, Dean Fido, and Dominic Petronzi. 2021. Delineating non-consensual sexual image offending: Towards an empirical approach. *Aggression and violent behavior*, page 101547.

Susan Hazan. 2020. Deep fake and cultural truth-custodians of cultural heritage in the age of a digital reproduction. In *International Conference on Human-Computer Interaction*, pages 65–80. Springer.

Mark K Hinders and Spencer L Kirn. 2020. Cranks and charlatans and deepfakes. In *Intelligent Feature Selection for Machine Learning Using the Dynamic Wavelet Fingerprint*, pages 297–346. Springer.

Yoori Hwang, Ji Youn Ryu, and Se-Hoon Jeong. 2021. Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, 24(3):188–193.

Serena Iacobucci, Roberta De Cicco, Francesca Michetti, Riccardo Palumbo, and Stefano Pagliaro. 2021. Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, Behavior, and Social Networking*, 24(3):194–202.

Nanna Inie, Jeanette Falk Olesen, and Leon Derczynski. 2020. The rumour mill: Making the spread of misinformation explicit and tangible. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–4.

Nektaria Kaloudi and Jingyue Li. 2020. The ai-based cyber threat landscape: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–34.

Ignas Kalpokas. 2021. Problematising reality: the promises and perils of synthetic media. *SN Social Sciences*, 1(1):1–11.

Vasileia Karasavva and Aalia Noorbhai. 2021. The real threat of deepfake pornography: a review of canadian policy. *Cyberpsychology, Behavior, and Social Networking*, 24(3):203–209.

Catherine Kerner and Mathias Risse. 2021. Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics*, 8(1):81–108.

Ali Khodabakhsh, Raghavendra Ramachandra, and Christoph Busch. 2019. Subjective evaluation of media consumer vulnerability to fake audiovisual content. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020a. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146.

Jan Kietzmann, Adam J Mills, and Kirk Plangger. 2020b. Deepfakes: perspectives on the future "reality" of advertising and branding. *International Journal of Advertising*, pages 1–13.

Nils Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice–people cannot detect deepfakes but think they can. *Available at SSRN 3832978*.

Andrei OJ Kwok and Sharon GM Koh. 2021. Deepfake: A social construction of technology perspective. *Current Issues in Tourism*, 24(13):1798–1802.

Kevin LaGrandeur. 2021. How safe is our reliance on ai, and should we regulate it? *AI and Ethics*, 1(2):93–99.

Jack Langa. 2021. Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, 101:761.

Johannes Langguth, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková, and Daniel Thilo Schroeder. 2021. Don't trust your eyes: Image manipulation in the age of deepfakes. *Frontiers in Communication*, 6:26.

YoungAh Lee, Kuo-Ting Huang, Robin Blom, Rebecca Schriner, and Carl A Ciccarelli. 2021. To believe or not to believe: framing analysis of content and audience response of top 10 deepfake videos on youtube. *Cyberpsychology, Behavior, and Social Networking*, 24(3):153–158.

Sophie Maddocks. 2020. 'a deepfake porn plot intended to silence me': exploring continuities between pornographic and 'political'deep fakes. *Porn Studies*, 7(4):415–423.

Artem A Maksutov, Viacheslav O Morozov, Aleksander A Lavrenov, and Alexander S Smirnov. 2020. Methods of deepfake detection based on machine learning. In *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 408–411. IEEE.

Karolina Mania. 2020. The legal implications and remedies concerning revenge porn and fake porn: A common law perspective. *Sexuality & Culture*, 24(6):2079–2097.

David Moher, Larissa Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. 2015. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, 4(1):1–9.

Gillian Murphy and Emma Flynn. 2021. Deepfake false memories. *Memory*, pages 1–13.

Nicholas O'Donnell. 2021. Have we no decency? section 230 and the liability of social media companies for deepfake videos. *U. Ill. L. Rev.*, page 701.

Axel Oehmichen, Kevin Hua, Julio Amador Díaz López, Miguel Molina-Solana, Juan Gomez-Romero, and Yi-ke Guo. 2019. Not all lies are equal. a study into the engineering of political misinformation in the 2016 us presidential election. *IEEE Access*, 7:126305–126314.

Carl Öhman. 2019. Introducing the pervert's dilemma: a contribution to the critique of deepfake pornography. *Ethics and Information Technology*, pages 1–8.

Konstantin A Pantserev. 2020. The malicious use of ai-based deepfake technology as the new threat to psychological security and political stability. In *Cyber defence in the age of AI, smart societies and augmented humanity*, pages 37–55. Springer, Cham.

Miroslava Pavlíková, Barbora Šenkỳ˘rová, and Jakub Drmola. 2021. Propaganda and disinformation go online. *Challenging Online Propaganda and Disinformation in the 21st Century*, pages 43–74.

Adnan Qayyum, Junaid Qadir, Muhammad Umar Janjua, and Falak Sher. 2019. Using blockchain to rein in the new post-truth world and check the spread of fake news. *IT Professional*, 21(4):16–24.

Md Shohel Rana and Andrew H Sung. 2020. Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 70–75. IEEE.

Claudia Ratner. 2021. When "sweetie" is not so sweet: Artificial intelligence and its implications for child pornography. *Family Court Review*, 59(2):386–401.

Colonel Prof Filofteia Repez and Maria-Magdalena Popescu. 2020. Social media and the threats against human security deepfake and fake news. *Romanian Military Thinking*, (4).

Jean-Marc Rickli and Marcello Ienca. 2021. The security and military implications of neurotechnology and artificial intelligence. *Clinical Neurotechnology Meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*, page 197.

Helen Ronsner. 2021. The ethics of a deepfake anthony bourdain voice.

Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2019. Poster: Towards robust open-world detection of deepfakes. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2613–2615.

Eliza Strickland. 2018. Ai-human partnerships tackle" fake news": Machine learning can get you only so far-then human judgment is required-[news]. *IEEE Spectrum*, 55(9):12–13.

Eliza Strickland. 2019. Facebook takes on deepfakes. *IEEE Spectrum*, 57(1):40–57.

Catherine Stupp. 2019. Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. *The Wall Street Journal*, 30(08).

Bryan C Taylor. 2021. Defending the state from digital deceit: the reflexive securitization of deepfake. *Critical Studies in Media Communication*, 38(1):1–17.

Japan Times. 2020. Two men arrested over deepfake pornography videos.

Karin Wahl-Jorgensen and Matt Carlson. 2021. Conjecturing fearful futures: Journalistic discourses on deepfakes. *Journalism Practice*, pages 1–18.

Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11).

Jeffrey Westling. 2019. Are deep fakes a shallow concern? a critical analysis of the likely societal reaction to deep fakes. *A Critical Analysis of the Likely Societal Reaction to Deep Fakes (July 24, 2019)*.

Digvijay Yadav and Sakina Salmani. 2019. Deepfake: A survey on facial forgery technique using generative adversarial network. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 852–857. IEEE.

Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–14.

G Pascal Zachary. 2020. Digital manipulation and the future of electoral democracy in the us. *IEEE Transactions on Technology and Society*, 1(2):104–112.

Zsolt Ziegler. 2021. Michael polányi's fiduciary program against fake news and deepfake in the digital age. *AI & SOCIETY*, pages 1–9.