Deepfakes: A Digital Transformation Leads to Misinformation

Nika Nour and **Julia Gelfand**
University of California, Irvine, USA

Abstract

Deepfakes are a product of artificial intelligence (AI) and software applications used to create convincing falsified audiovisual content. Linguistically, a portmanteau combines deep learning aspects of AI with the doctored or falsified enhancements that deem the content fake and now deepfake or misinformation results. A variety of sophisticated software programs' exacting algorithms create high-quality videos and manipulated audio of people who may not exist, twisting others who do exist, creating the potential for leading to the spread of serious misinformation often with serious consequences. The rate of detection of this digital emergence is proliferating exponentially and the sourcing is challenging to verify, causing alarms. Examples of this pervasive information warfare are is associated with deepfakes that range from identity theft, discrediting public figures and celebrities, cyberbullying, blackmail, threats to national security, personal privacy, intensifying pornography and sexual exploitation, cybersecurity, baiting hate crimes, abusing social media platforms and manipulating metadata.

Deepfakes that are difficult to cite, acquire, or track have some parallel attributes to grey literature by that definition. Often detectable, yet problematic, activities such as phishing and robocalling may be common attempts of deepfake activities that threaten and interrupt rhythms of daily life. The increasing online personas that many people create or assume contribute to this fake content and potential for escalated exploitation due to technical abilities to copy and reimagine details that are not true. AI image generators create completely false images of people that simply don't exist within seconds and are nearly impossible to track. While AI is perceived as a positive benefit for science and policy, it can have negative roles in this new AI threatened environment. Deepfakes have cross-over targets in common business applications and society at large. Examples of this blur are targeted advertising, undetected security cameras in public spaces, blockchain, tabloid press/paparazzi, entertainment, computer games, online publishing, data and privacy, courtroom testimony, public opinion, scientific evidence, political campaigns, and rhetoric.

This paper explores the impact and intersections of these behaviors and activities, products of AI, and emerging technologies with how digital grey and the optics of grey expose the dangers of deepfakes on everyday life. Applying a security and privacy lens, we offer insights of extending libel and slander into more serious criminal behavior as deepfakes become more pervasive, construing reality, endangering personal, social, and global safety nets adding to the new normal we assume today. How we became more sensitized to misinformation and fake news tells the story about deepfakes.

Introduction

Grey Literature (GL) has a history and nomenclature that includes a definition that suggests it is a "relatively recent collective noun" for "information produced on all levels of government, academia, business and industry in electronic and print formats not controlled by commercial publishing where publishing is not the primary activity of the producing body" (Farace and Frantzen 2005). Traditional scholarly literature is defined by having its submissions go through peer review. Often, grey literature is not peer-reviewed as it is not usually published in journal literature or monographs released by commercial or scholarly imprints. This suggests the parallels we see with some deepfake strategies and products since they too may not be the primary output of their sources and certainly are not peer-reviewed.

Other critical intersections include how findings from the 'Project Overview of the Grey Literature Strategies' concluded best practice guidelines for producing and managing grey literature in Australia to transform access to public interest research for various communities (Aloia and Naughton, 2017). Considering the news trajectories of the past few years and the political and social climate we are currently experiencing, this is a very timely assessment. Part of the Australian research and its outcomes demonstrated that "finding better ways to access, control, evaluate, collect and preserve grey literature is an important national and international issue" (Houghton, 2011) in our current environment, much as it was leading up to today. Publishing strategies have changed as electronic publishing and preservation options have widely expanded the creation and dissemination of information linking different formats such as text, video, data, imagery, and other content into single outputs. Five years ago, it was reported that "Australia produces $30 billion worth of 'grey literature' that we can't read" (McCallum, 2016), suggesting how government, academe, corporate entities, and other information providers issue content that is not well cited nor curated. The value of these resources is understated due to poor discovery and application. Some examples of new databases containing content from many different repositories and sources are now practicing commendable stewardship, safeguarding, indexing, and curation to protect "endangered" or lost content that is getting recognition and inclusion in contemporary research. One such example is Policy Commons, launched in 2020 by Coherent Digital to be recognized for its aggregation of premium full-text content, subject neutral but research-intensive and described by co-founder Toby Green as "Essential research done by IGOs, NGOs, think tanks and research centers simply doesn't get the attention it deserves. It's hard to find, it disappears when funding is lost, it lacks persistent identifiers, and Policy Commons fixes these issues" (Green, 2020). That is the difference in what defines something as grey.

The study of grey literature has matured over the last couple of decades as its importance has been applied to new fields, disciplines, technologies, products, and outcomes. With that has come the refinements of how specialized content that forms the nexus of grey literature has shown the importance of its applications in many fields such as sciences, technology, health, government operations, public policy, and even arts and visual studies (Pappas and Williams, 2011). The linking factor of communications, and perhaps the intersection of information, communication, and technology best described by ICTs has shown how specialized content of working papers, technical reports, illustrations, images, spatial and quantitative data, legislation, thesis/dissertations, lyrics, product descriptions, plans, blogs, tweets, and hosts of other information outputs has had a dizzying effect on the publishing landscape as each has become increasingly valuable to communicate with readers and informants about certain observations, opinions, conclusions, and other outputs challenging to find, source, cite and share in the everyday lexicon of use. This challenged how it was collected, described, transmitted, and used.

During the Coronavirus/COVID 19 pandemic, the health sciences, and government agencies, in particular, were faced with trying to convey the medical consequences of the pandemic and direct citizens to act responsibly and rationally in responding to this worldwide crisis. However, efforts at communicating how best to practice recommendations and guidelines were met with resistance. For instance, the public questioned mandates and practices to honor social distancing, personal hygiene practices, engage in tracing and testing, and be vaccinated once shots became available (Brennan, 2020). The effort to focus on scientific evidence rather than conspiracy theories and opinions that lacked medical data has been challenging, framing one of the best examples of how misinformation has flourished by different media establishments and outlets over the past two years testing credibility and legitimacy at every turn. One of the most prominent examples of COVID-19 misinformation is the long list, known as a factsheet compiled by Brennen et al. where this team analyzed a sample of 225 pieces of misinformation rated false or misleading by fact-checkers and published in English between January through March 2020, drawn from a collection of fact checks maintained by the British First Draft News (Brennen et al., 2020). Social media has contributed to both the good and negative information sharing that defines our knowledge base today.

Understanding Misinformation

We have adopted the following definitions by Wardle and Derakhshan and share their sentiments that "it's important to distinguish true and false messages, as well as messages that are and are not created, produced or distributed with intent" (Wardle and Derakhshan, 2019).

- Misinformation - false information shared by someone who believes to be true
- Disinformation - by contrast, is false information shared with knowledge of its falsity and thus intention to deceive or otherwise do harm. It is a deliberate, intentional lie.
- Malinformation - information based on a reality that is shared to harm a person, organization, or country. This term can refer to instances where private information is made public or genuine imagery is reshared in the wrong context.
- Information disorder - an umbrella term encompassing all forms of the above

Grey Literature has not been accused of serving up misinformation or fake news; however, it has been aligned with questions about why this sourcing is so fragile and why such valuable content has fallen through the cracks in what is considered a sophisticated publishing network. Several reasons begin to tell this tale. Local, federal, and international government agencies and nonprofit organizations have been forced to make critical decisions about what to archive and how to safeguard and distribute it (Langa, 2021). The financial side of the publishing enterprise is increasingly volatile as the call for open access has been made loud and clear in academic and broader circles (Fraga-Lamas and Fernandez-Carames, 2020). As "born digital" became more common, the release and citation of this material became less structured, and copyright protections were reduced. According to Suzanne Smalley, Roger Schonfeld speculates "how misinformation, politicization and other problems embedded in the open-access movement stem from a "mismatch" between the incentives in science and how 'openness and politicization are bringing science into the public discourse" (Schonfeld, 2021).

The Good, The Bad and the Future of both Grey Literature and Deepfakes

As Richard Van Hooijdonk states in his blog, "The competition between the creation and elimination of deepfakes will become increasingly fierce in the future, with deepfake technology not only becoming easier to access but its content easier to create and harder to distinguish from real" (Van Hooijdonk, R, 2021). This is not the case with grey literature as it more uniformly has reduced the gap between traditional commercial publishing and other new forms of expression and types of content, extending the open access movement. It can be perceived as Van Hooijdonk suggests that "the ability to use artificial intelligence to create realistic simulations might even be a positive thing for humanity" (2021).

Grey literature has been deeply seeded in medical and scientific areas from its inception. This transfer can also be seen in the correlated history of the open access movement when the demand for access to medical information was made by patients who had no systematic access to information about medical conditions, diagnoses, and treatments their physicians and healthcare providers were recommending as most of the information was published in subscription-based journals held by special libraries to which the unaffiliated public was not allowed.

Wardle and Derakhshan's "Misinformation and Disinformation Framework" when applied to documentary films and media and what challenges they pose to audiences and perhaps readers is yet another example of what they offer as a new lexicon to help to distinguish "between media commentary, misinterpreted material, playful content and media created to deliberately mislead" (Hight, 2021). Summarized below are several examples of different forms captured by Hight from Wardle's framework that we find insightful when describing how misinformation is created:

- Fabricated content – deliberately designed to deceive, completely false
- Manipulated content – where there has been a manipulation of genuine material;
- Imposter content – when genuine sources are impersonated;
- False content – genuine content mixed with false contextual information;
- Misleading content – misleading use of information to frame an issue or individual;
- False connection – when headlines, visuals, or captions don't support the content;
- Satire of parody – fake content intended for social commentary

In addition, Claire Wardle has extended content in her created definitional toolbox, offering a glossary. She maps the landscape and describes graphically more about Information Disorders indicating how these tools will help contextualize and categorize how we can more appropriately address issues of trust and truth in the digital age (Wardle, 2018). She contributes a bright assessment about how these categories should be used in parallel and should be commended for the work that she has contributed to *First Draft*, where she has created a collaboration to "stand up for truth in a polarized world" (Wardle, 2018).

Another lesson about misinformation relevant to grey literature is how Nossel addresses disinformation containment strategies. She says "the originators of disinformation – not just foreign governments but conspiracists, provocateurs, and paid propagandists – are too diffuse to be shut down; even trying to shut them down would unavoidably impinge on expressive rights." She continues her drift with "Stopping disinformation will also require a more refined understanding of who consumes it and why" and concludes that information consumers can be divided into the anchored, the adrift and the marooned." (Nossel, 2021). This suggests our agreement that it is the middle group, the "informationally adrift," as Nossel refers to them, who are most challenging to educate or inform because of their over exposure to content, they are "prone to perpetual doubt" and are part of growing communities of disbelief and denial. This means that they can't distinguish sourcing and determine what is rightful or credible. The challenge is to urge this diverse body to do their research in trusted sources building media literacy and grey literature when made available, archived and heavily cited contributes to that slow fix while parsing misinformation and disinformation.

With the quickly changing media landscape and generations of users relying on multiple social media platforms, findings from Aswani and colleagues are equally alarming and support that misinformation is created by misinformation propagators, and how it is being shared. Their Twitter analysis found that "misinformation was 43% of tweets were works of fiction, 27% were rumors and 22% was from vested interests of organizations and individuals for the purpose of content promotion and advertising with only 8& from government, politicians and media sources combined." (Aswani, et al, 2019).

Misinformation Leading to Deepfakes

Even though artificial intelligence and online misinformation consistently make headlines, minimal research exists on digitally-altered visual content such as AI-generated deepfake videos. This paper focuses on video experience with deepfakes rather than covering the entire spectrum of deepfake creation. Using deep learning algorithms, "deepfake" videos (or colloquially, just "deepfakes") typically substitute one person's visual and acoustic likeness for another, presenting viewers with compelling videos of individuals doing and saying things they never did or said (Vaccari et al., 2020). In addition, many politicians and media pundits believe that deepfake videos can influence elections, instigate violence, and destroy claims to the truth (Paris et al., 2010). Although verifying online content and imagery is not a new phenomenon—in fact, the entire field of image forensics exists for this purpose—academic research exploring this topic is only recently emerging. However, due to the rising prevalence of deepfake videos and their ability to promulgate violence and mistrust, it is increasingly vital for academics to deploy their technical and theoretical expertise in combating misinformation online.

What makes deepfake videos especially worrisome is the relative ease and accessibility with which adverse actors can manipulate moving images. With virtually no technical expertise, individuals can produce untraceable, deceptive videos and distribute them online from almost anywhere in the world. Altered videos have the power to propel a terrorist group's agenda or reword a politician's speech. The critical factors for assessing non-textual media distribution are the rapid pace of technology, the recent growth in digitally altered content, and users' varying levels of trust in visual imagery. While algorithmic detection for deepfake videos exists, these technologies' access and deployment remain uneven, and artificial intelligence blockers have struggled to eliminate harmful online posts and publications (Delort and Paris, 2011). Historically, this type of conduct has led to name-calling tactics, releasing personally identifiable information (doxxing), and similar intentions to create reputational harm. Deepfake videos, armed with the rhetorical persuasiveness of the moving image, represent a significant jump in the ability to inflict social damage.

Acknowledging that the algorithmic detection of deepfake videos is mired in its troubles of reliability, accessibility, and credulity, there is a robust, practical impetus to evaluate—and, where necessary—cultivate the human capacity for discerning between legitimate and deepfake manipulated videos (Biometric, 2019). In this paper, we document the process of creating deepfake videos as research instruments to test the ability of individuals to detect deepfakes in a controlled setting, outlining both the steps taken and the methodological motivations for them. We also draw comparisons to other information outputs. Finally, our documentation suggests that multiple considerations must be taken into account when creating altered content for research purposes, particularly for the sake of generalizability, concerning how individuals encounter altered content online.

The Rise and Ethics of Deepfakes

Manipulating visual media is becoming widely accessible, inexpensive, and easier to accomplish (Wade, et al, 2002). With software becoming more user-friendly, cell phone cameras increasing in quality, and the rise of technology literacy, most people have the opportunity to doctor media, an option once reserved for video editing professionals and filmmakers. Fake news and artificial intelligence consistently make headlines in today's media market; however, there is minimal research on digitally altered, visual content, specifically deepfake videos. As deepfakes rise in prevalence, there are concerns about the medium's ability to influence critical societal functions and political outcomes. Even memory researchers are studying how falsified videos can plant memories through familiarity, imagery, and credibility in how these manipulated content pieces are presented (Nash, Wade, and Brewmann, 2009). Without developing detection strategies, deepfakes can invoke personal and societal harm, threatening the foundations of trust and society. As the internet user's confidence in content fades and false information is presented as accurate, society becomes more susceptible to information warfare.

Algorithmic Detection

While some scholarly reviews of deepfake algorithms like FakeApp, Adobe VoCo, Lyrebird, and Face2Face have tended towards the ethically agnostic (Gardiner 2019), other recent research has shifted towards the creation of methods to detect deepfake manipulations, with both the explicit and implicit standpoints that deepfake videos pose social and moral dangers to the general public. Deep learning and machine learning feature prominently in the production of algorithmic attempts to detect deepfakes and—by extension of their detection—to impune on their persuasiveness of deepfake videos distributed online.

Image forensics research in the area of deepfake videos has focused on the use of neural networks (Güera and Delp, 2018; Amerini et al., 2019; Guarnera et al., 2020), classical frequency domain analysis (Durall et al., 2019), facial recognition (Korshunov and Marcel, 2019), and the algorithmic detection of visual artifacts that emerge during deepfake manipulation (Marra et al., 2018; Li et al., 2019). Problematically, as observed by Nguyen et al. (2019), the very technologies used to detect deepfake videos are the same as those used to create deepfake videos, thus perpetuating an arms race hinged on the co-advancement of shared algorithms.

Human Detection

Research on human (as opposed to machine) deepfake detection has more strongly figured around issues of literacy (Nightingale et al., 2017; Schetinger et al., 2017) and the social outcomes of deepfake circulation across social media platforms (Vaccari et al., 2020). Given that the creation of deepfake forensic algorithms consequently advances the ability of deepfake video creators to evade algorithmic detection further, there is an urgent need to cultivate and advance the human detection of deepfake manipulations. Moreover, the circumstances in which individuals encounter deepfake videos unwittingly differ dramatically from the large, high-resolution datasets that algorithmic detection models learn from. For instance, in Durall et al,'s work, as image resolution for deepfake videos fell, so too did their algorithm's predictive accuracy, suggesting that we cannot entirely rely on machines to detect manipulations in low-resolution videos correctly (2019). When coupled with Marra et al.'s observations on how the circulation of deepfake videos occurs primarily via social media channels that compress images and videos—and in doing so, obfuscate algorithmic detection "signals" with low-resolution, artifactual "noise"— it becomes increasingly evident that the need for human detection further increases (particularly when access to algorithmic detection models is uneven, impractical, and inequitable (2018).

Additional Applications

Other related technologies like drones that follow and track individuals give the public some of the same insecurity since they have not authorized or given permission to be followed or tracked. However, we can't forget how through the 2020 U.S. Presidential Election news outlets provided fact-checking after nearly every presidential debate and large campaign rally trying to force candidates to substantiate their claims with accurate history, chronology, and sourcing, rather than high-risk cases of any of the above forms of misinformation.

Data as an example of grey literature is reflected by Polonetsky, Tene, and Finch in their article, "Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification" (2016). They propose "parameters for calibrating legal rules to data depending on multiple graduations of identifiability, while also assessing other factors such as an organization's safeguards and controls, as well as the data's sensitivity, accessibility, and permanence" (Polonetsky et al., 2016). "It builds on emerging scholarship that suggests that rather than treat data as a black or white dichotomy, policymakers should view data in various shades of gray; and provides guidance on where to place important legal and technical boundaries between categories of identifiability." (Polenetsky et al., 2016, 595). They go on to create a data spectrum of key inflection points that include:

- Explicitly Personal Data
- Potentially Identifiable & Not Readily Identifiable Data
- Key-coded Data
- Pseudonymous and Protected Pseudonymous Data
- De-Identified and protected De-Identified Data
- Anonymous and Aggregated Anonymous Data

Their work continues to justify how important it is to separate between the sensitivity of a data item and its degree of identifiability and concludes with "New uses of data and technology have the potential to bring humanity a wide range of benefits, but at the same time to generate new and serious harms'' while advancing "an approach that supports benefit and deters risk by providing a practical framework for policymakers to analyze various data sets based on their degree of identifiability" (Polenetsky et al., 2016, 623).

Reflections of Technology

The Evolution of Media Manipulation

Manipulating photos and videos isn't a new concept. Instead, an early example of media manipulation includes a portrait of Abraham Lincoln in 1860. Though the image seemed real, it was actually multiple photographs of Lincoln's head and John Calhoun's body stitched together into one portrait (Caldera, 2020). As image and photo manipulation became more common, video editing and manipulated also took hold as early as the 1970s where computer animation and visual effects became a widely adopted technique in the entertainment industry (Caldera, 2020). However, the rise of hyper-realistic simulations and media have far surpassed the simple edits made to videos and photos and taken a much more nefarious turn in recent years. As media technologies advance and develop synthetic results, falsified audiovisual content is appearing and sounding more realistic (Johnson and Diakopoulos, 2021). Examples of the deepfake phenomenon include videos celebrities doctored into movie clips they never performed to Mark Zuckerberg stating that he was going to delete Facebook (Kietzmann et al., 2019). Though falsifying images, audio, and video content isn't a new concept, the phenomenon of deepfakes led to a boom of this new type of false media when an anonymous Reddit user shared the computer code to place famous personalities into pornographic clips (Kietzmann et al., 2019). The accessibility and feasibility of deepfakes have also rapidly evolved. Today's deepfakes are made with minimal coding knowledge and require no elaborate hardware. Advanced computer algorithms and apps can quickly generate, produce, and edit videos that are difficult to tell apart from the original content (Skibba, 2020). As deepfakes become more prevalent on social media platforms and mobile networks, generated adversarial networks (GANs) allow for faster dissemination and create vulnerabilities for face recognition software (Korshunov and Marcel, 2018). With facial recognition becoming less reliable in these circumstances, this technology poses a new threat to the foundation of trust online as people may be unable to detect or determine whether or not these videos are real. While some of the existing deepfake videos seem harmless, the potential consequences of this technology can impact citizenry, social welfare, business, and life in general.

The Consequences of Deepfakes in Everyday Life

As fake content becomes more prolific in the digital age, we must understand the current and potential societal impact of deepfakes on the citizenry, education, social welfare, politics, media, business, and family life. The creation of deepfake videos is becoming easier and more accessible for unskilled end-users through programs such as FakeApp, a face-swapping program, Zao, a Chinese mobile app using movie clips, and Apple's text-to-speech (TTS) editing system (Kietzmann, 2019).

Accessibility and new technologies suggest how much deepfakes are increasingly grey. Superimposing and swapping faces of individuals into movies and television clips can be an entertaining example of deepfake videos. For instance, in 2019, a clickbait video entitled "Keanu Reeves Stop A ROBBERY '' included a stuntman and voice actor with his face replaced with celebrity actor Keanu Reeves (Bode, 2021). The video was shared across multiple social media platforms, but the content was falsified, and Keanu Reeves was never involved. Though this type of content can be amusing, the ease of deepfake content creation means accountability is lacking, and entertainment becomes nefarious information dissemination. The fact that such changes are not chronicled provides no record of artifacts or tracing available. Anyone, anywhere, at any time, can create these doctored videos to convince the general public that they are real. However, the ability to track, hold accountable, or impose regulations on these perpetrators are practically nonexistent or require legal and financial resources often dismissed. For instance, with limited oversight, deepfakes can undermine trust in the news and elections, resulting in ethical implications (Diakopoulos and Johnson, 2019). In 2017, researchers successfully created deepfake stills of former President Barack Obama saying things that he never actually said but were voiced by actor, Jordan Peele (Citron and Chesney, 2019). The same technology can create false videos inciting plans to carry out political assassinations, seemingly private conversations featuring elected leaders (Citron and Chesney, 2019), or Speaker Nancy Pelosi slurring her speech (Denham, 2020). With deepfakes being the next frontier for fraud, companies like Recorded Future found multiple examples on the dark web of criminal activity using deepfakes to blackmail, create pornographic videos, and execute identity theft (Security Firm, 2021). Society and government entities need appropriate countermeasures when such actions impact personal liabilities such as non-consenting individuals, government officials, and organizations (Kietzmann et al., 2019).

Conclusions

Our exploration of this topic moves us to share the conclusion stated by Vizoso, et al, that deep fakes are highly understudied and how they will continue contributing to more misinformation will only proliferate. We speculate that there are many new research paths that can emerge as we observe how journalists, media creators, information analysts and others respond to the damage that deep fakes can cause to society and how technology and media literacy can potentially change that course. (Vizoso et al, 2021).

Grey literature may be reduced as more outputs become discoverable. As search, meta-data, and analytics become more embedded in digital materials, deepfakes are newer, multi-formatted, and potentially more prolific. With terrorism, privacy, distribution channels, relationships to blockchain, and cryptocurrency, information warfare is regularly redefined and more concerning. This leads to an unstable digital marketplace where even the definition of grey is evolving. These global concerns come with implications as stakeholders and authorities enter the wild west of policy and legal oversight. These unknown territories will only expand and cross digital borders as deepfake creators become more aggressive and confident in their abilities to create harm. The scale at which these moving targets are changing will dictate how regulatory bodies respond and develop legal frameworks in this grey area of policing the digital arena. Digital transformation is an ongoing process, especially in the space where deepfakes are proprietary but not identifiable, definable, or attributable. As society learns to govern and respond to the new grey, individuals must take it upon themselves to build media literacy and resiliency when faced with all forms of information disorder.

References

Aloia, D. and Naughton, R. (2017). The Greylit Report: Understanding the challenges of finding grey literature. *Grey Journal. 13.* 75-80. https://doi.org/10.17026/dans-2z8-x27y

Amerini, I., Galteri, L., Caldelli, R. and Del Bimbo, A. “Deepfake Video Detection through Optical Flow based CNN,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (ICCV), 2019.

Aswani, R., Kar, A.K., and Ilvarasan, P.V. (2019). Experience: Managing Misinformation in Social Media – Insights for Policymakers from Twitter Analytics. *ACM Journal of Data and Information Quality 12* (1), Article 6. https://doi.org/10.1145/3341107

Bode, L. (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. *Convergence: The International Journal of Research into New Media Technologies 27* (4) 135485652110304-934. https://doi.org/10.1177/13548565211030454

Brennen, J.S., Simon, F.M., Howard, P.N., and Nielsen, R.K. (2020). Types, Sources and Claims of COVID-19 Misinformation.

Caldera, Elizabeth (2020) " ‘Reject the Evidence of Your Eyes and Ears’: Deepfakes and the Law of Virtual Replicants," *Seton Hall Law Review*: Vol. 50 : Iss. 1 , Article 5. Available at: https://scholarship.shu.edu/shlr/vol50/iss1/5

Chesney, Robert and Danielle Citron, (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy and National Security. *California Law Review* 107.

Citron, D.K. and Chesney, R. (2019). Deepfakes and the New Disinformation War. *Foreign Affairs*. Available at: https://scholarship.law.bu.edu/shorter_works/76

Coherent Digital (2020, November 2). Press Release: *World’s largest resource for public policy launches today*. Retrieved November 5, 2021 from https://coherentdigital.net/policycommonslaunch

Deepfake videos easily fool face systems, researchers warn. (2019). *Biometric Technology Today.* (10), 3–3. https://doi.org/10.1016/s0969-4765(19)30137-7

Delort, J., Arunasalam, B., and Paris, C. (2011). Automatic Moderation of Online Discussion Sites. *International Journal of Electronic Commerce, 15* (3), 9-30. doi:10.2753/jec1086-4415150302

Denham, H. (2020, August 3). *Another fake video of Pelosi goes viral on Facebook*. The Washington Post. Retrieved December 2, 2021 from https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/.

Diakopoulos, N. and Johnson, D. (2019). Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections. *SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3474183*

Durall, R., Keuper, M., Pfreundt, F., and Keuper, J. (2019). Unmasking DeepFakes with simple Features. https://arxiv.org/abs/1911.00686

Farace, D., Frantzen, J., Schöpfel, J., Stock, C. and Boekhorst, A. (2006). Access to Grey Content: An Analysis of Grey Literature based on Citation and Survey Data: A Follow-up Study. In: Seventh International Conference on Grey Literature: Open Access to Grey Resources, Nancy, France, December 5-6, 2005. - *GL7 Conference Proceedings*, 194-203.

Fraga-Lamas, P. and Fernández-Caramés, T. (2020). Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality. *IT Professional. 22.* 53-59. https://doi.org/10.1109/MITP.2020.2977589

Gardiner, N.J. (2019). *Facial re-enactment, speech synthesis and the rise of the Deepfake*. https://ro.ecu.edu.au/theses_hons/1530

Gelfand, J. and Lin, A. (2017). Social Media Matters: Showing Up Online as Well as Ontime. Paper presented at the 19th International Conference on Grey Literature. Rome, Italy. October 23, 2017. https://av.tib.eu/media/37267 https://greyguide.isti.cnr.it/attachments/category/34/Gelfand_and_lin.pdf

Gelfand, J. and Tsang, D. (2015). Data: Is it Grey, Maligned or Malignant? Paper presented at the 16th International Conference on Grey Literature, Washington, DC, December 9, 2014. *The Grey Journal. 11*(1), 2015: 30-40.

Green, T. (2020). World's Largest Resource for Public Policy Launches Today. Press Release by Coherent Digital, November 2. https://coherentdigital.net/policycommonslaunch

Guarnera, L., Giudice, O., & Battiato, S. (2020). DeepFake Detection by Analyzing Convolutional Traces. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2841-2850.

Guera, David & Delp, Edward. (2018). Deepfake Video Detection Using Recurrent Neural Networks. 1-6. 10.1109/AVSS.2018.8639163.

Hight, C. (2021). Deepfakes and Documentary Practice in an Age of Misinformation. Continuum, DOI: 10.1080/10304312.2021.2003756

Johnson, D. G. and Diakopoulos, N. (2021). What to do about deepfakes. *Communications of the ACM. 64*. 33-35. https://doi.org/10.1145/3447255

Kietzmann, J., Lee, L., McCarthy, I., and Kietzmann, T. (2019). Deepfakes: Trick or treat? *Business Horizons. 63.* https://doi.org/10.1016/j.bushor.2019.11.006

Korshunov, P., & Marcel, S. (2018). DeepFakes: A New Threat to Face Recognition? Assessment and Detection. *ArXiv, abs/1812.08685*.

Langa, J. (2021). Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes. *Boston University Law Review 101(2)*, 761-801.

Lawrence, A., Houghton, J., Thomas, J. and Weldon, P. (2014). Where Is the Evidence? Realising the Value of Grey Literature for Public Policy & Practice, A Discussion Paper. https://apo.org.au/node/42299

Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. Detection of GAN-Generated Fake Images over Social Networks. (2018). 384-389. https://doi.org/10.1109/MIPR.2018.00084

McCallum, T. (2016). "How Australia Produces $30 Billion Worth of 'Grey Literature' That We Can't Read." *The Conversation,* April 27, 2016. Retrieved November 11, 2021 from https://theconversation.com/how-australia-produces-30-billion-worth-of-grey-literature-that-we-cant-read-56584

McGann, James. (2020). *2020 Global Go to Think Tank Index Report. https://repository.upenn.edu/think_tanks/18/*

Mercer, K., Weaver, K.D., and Waked, K. (2021). Navigating Complex Authorities: Intellectual Freedom, Information Literacy and Truth in Pandemic Stem Information. *IFLA Journal*, https://doi.org10.1177/03400352211048915

Mirsky, Y. and Lee, W. (2020). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys, 54* (1), Article 7. https://doi.org/10.1145/3425780

Nash, R. A., Wade, K. A., & Brewer, R. J. (2009). Why Do Doctored Images Distort Memory? *Consciousness and Cognition*, *18*(3), 773–780. https://doi.org/10.1016/j.concog.2009.04.011

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can People Identify Original and Manipulated Photos of Real-World Scenes? *Cognitive Research: Principles and Implications*, *2*(1), 30. https://doi.org/10.1186/s41235-017-0067-2

Nguyen, T.T., Nguyen, C.M., Nguyen, D., Nguyen, D.T., and Nahavandi, S. (2019). Deep Learning for Deepfakes Creation and Detection. https://arxiv.org/abs/1909.11573

Nossel, S. (2021). How to Save People from Drowning in a Sea of Misinformation
Slate, December 15. https://slate.com/technology/2021/12/information-consumers-misinformation-adrift-media-literacy.html?fbclid=IwAR3REIsINufZrqM29U_FEDIOorTdYXH-T8x6BkHLzlzqM4HoraB52v6G-x8

Pappas, C. and Williams, I. (2011). Grey Literature: Its Emerging Importance, *Journal of Hospital Librarianship, 11*(3), 228-234, https://doi.org/10.1080/15323269.2011.587100

Paris, Britt, and Donovan, Joan. (2019). Deepfakes and Cheap Fakes. *Data & Society.* *https://datasociety.net/library/deepfakes-and-cheap-fakes/*

Polensky, J., Tenem, O., and Finch, K. (2016). Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification." *Santa Clara Law Review, 56* (3), 593-628.

Policy Commons. (n.d.). Retrieved December 2, 2021 from https://policycommons.net/

Saunders, L. and Budd, J. (2020). Examining, Authority and Reclaiming Expertise.
Journal of Academic Librarianship 46, 102077. https://doi.org/101016/j.acalib.2019.102077

Schetinger, Victor & Oliveira, Manuel & Silva, Roberto & Carvalho, Tiago. (2015). Humans Are Easily Fooled by Digital Images. *Computers & Graphics, 68. https://doi.org/*10.1016/j.cag.2017.08.010

Schonfeld, R. (2021). Is Scientific Communication Fit for Purpose? *Scholarly Kitchen*, November 1. Retrieved from https://scholarlykitchen.sspnet.org/2021/11/01/is-scientific-communication-fit-for-purpose/

Security firm: Deepfakes are 'fraud's next frontier.' (2021). *Biometric Technology Today*, *(6)*, 2–3. https://doi.org/10.1016/s0969-4765(21)00064-3

Skibba, Ramin. (2020). Accuracy Eludes Competitors in Facebook Deepfake Detection Challenge. *Engineering, 6.* 1339-1340. https://doi.org/10.1016/j.eng.2020.10.008

Smalley, S. (2021, April 13). *As Misinformation Grows, Scholars Debate How to Improve Open Access*. Inside Higher Ed. Retrieved from https://www.insidehighered.com/news/2021/11/08/open-access-science-misinformation-era

Vaccari, C. and Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. S*ocial Media + Society. 6.* 205630512090340. https://doi.org/10.1177/2056305120903408

Van Hooijdonk, R. (2021, April 23). The Good, the Bad, and the Future of Deepfakes [web log]. Retrieved November 11, 2021 from https://richardvanhooijdonk.com/

Vizoso, A. Vaz-Alvarez, M. and Lopez-Garcia, X. (2021). Fighting Deepfakes: Media and Internet Giants' Converging and Diverging Strategies Against Hi-Tech Misinformation*. Media and Communication*, *9* (1): 291-300. https://doi.org/10.17645/mac.v9i1.3494

Wade, K., Garry, M., Read, J., and Lindsay, D. (2002). A Picture is Worth a Thousand Lies: Using False Photographs to Create False Childhood Memories. *Psychonomic Bulletin & Review. 9*. 597-603. https://doi.org/10.3758/BF03196318

Wardle, C. (2018, July 9). *Information disorder: The definitional toolbox*. First Draft. Retrieved November 2, 2021 from https://firstdraftnews.org/articles/infodisorder-definitional-toolbox/

Wardle, C. (2019). First Draft's Essential Guide to Understanding Information Disorder." *First Draft News.* Retrieved November 11, 2021 from https://firstdraftnews.org/wpcontent/uploads/2019/10/Information_Disorder

Zimdars, M. and K. McLeod, eds. (2020). *Fake News: Understanding Media and Misinformation in the Digital Age.* Cambridge, MA: MIT Press, 2020.