Research Article

CONVERGENCE

Convergence: The International
Journal of Research into
New Media Technologies
2021, Vol. 27(1) 36–51
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1354856520923963
journals.sagepub.com/home/con
SAGE

Whose dystopia is it anyway? Deepfakes and social media regulation

Aya Yadlin-Segal
Hadassah Academic College, Israel

Yael Oppenheim
The University of Haifa, Israel

Abstract

This study explores global journalistic discussions of deepfake applications (audiovisual manipulating applications based on artificial intelligence (AI)) to understand the narratives constructed through global coverage, the regulatory actions associated with these offered narratives, and the functions such narratives might serve in global sociopolitical contexts. Through a qualitative–interpretive narrative analysis, this article shows how journalists frame deepfakes as a destabilizing platform that undermines a shared sense of social and political reality, enables the abuse and harassment of women online, and blurs the acceptable dichotomy between real and fake. This phenomenon is tied to discussions of dis/misinformation, manipulation, exploitation, and polarization in the media ecosystem these days. Based on these findings, the article then provides broader practical and theoretical insights about AI content regulation and ethics, accountability, and responsibility in digital culture.

Keywords

Artificial intelligence, deep learning, deepfake, dystopia, gender, internet policy, journalism, narrative analysis, news, social media regulation, thick data

Introduction

But what happens when deepfakes stop being the fare of redditors with free time and a fixation on Game of Thrones actresses and become a weapon for state actors interested in destabilization of governments like the United States? Are we ready for that? Cause it kind of looks like fake news in

Corresponding author:
Aya Yadlin-Segal, The Department of Politics and Communication, Hadassah Academic College, 37 HaNevi'im St., Jerusalem 9101001, Israel.
Email: ayayad@hac.ac.il

2016 was the opening salvo in a continuous, deep mind fuck, and that attempts to influence midterm elections have already begun. (VentureBeat February 8, 2018)

Like other journalists who covered recent developments in the field of artificial intelligence (AI) technologies, Johanson, the author of the above quotation, grappled with the issue of social media content regulation and the implications of 'deepfakes', a recent development in the field of algorithm-based deep-learning applications. Deepfakes are a set of AI algorithms used to synthesize multiple audiovisual products into one manipulated media item (usually videos), for example through face-swap (Floridi, 2018). The result of the manipulation is a *fake* video built through *deep* learning algorithm (hence, deepfake) in which a person is seen doing/saying something they, in fact, never did. Deepfakes surfaced on Reddit in December 2017 and have since gained attention. In the context of this article, deepfakes are approached as a case study that shed an important light on larger trend of manipulations, fakery, and disinformation in the contemporary media ecosystem.

By means of a qualitative–interpretive narrative analysis, we examine in this article journalists' framing of the need to regulate social media platforms. Given that journalistic framing of technologies becomes an important arena for unpacking political, social, and cultural trends (Marciano, 2019), we analyze contemporary journalistic reporting about deepfakes published on global news platforms to address how the call for such regulation interacts with three important topics: exploitation and harm of vulnerable groups, the state of concepts such as 'truth' and 'reality' in the current digital, post-truth ecosystem, and the future of traditional gatekeeping and accountability in a fake-saturated society. These are further unpacked in the following sections of the article.

We open the article with a review of literature related to two loci: the dystopian versus utopian rhetoric dichotomy in the field of new media technologies, and the role news framing and journalistic discourses play in the context of regulation as part of internet policymaking. We then turn to discuss the unique data collection and analysis processes employed in this study. Here we present the combination of automated big-data collection with manual data sifting that produced a thick data corpus analyzed through a qualitative narrative inquiry method. Following this section, we discuss the main narratives found in the analysis process, addressing (as elaborated below) three perspectives to social media regulation: gendered, factual, and professional.

We conclude the article by discussing two aspects of the deepfakes phenomenon: One theoretical and one practical. First, given the tension between traditional media gatekeepers (e.g. editors and journalists) and Internet-media users, which was an integral part of the narratives constructed by journalists, we contextualize deepfakes in the broader discussion of media responsibility, ethics, and accountability. Here we specifically explore the implications of deepfakes for vulnerable groups. Second, based on the findings in our analysis, we ask to provide some practical approaches to journalists working in our current manipulation saturated media ecosystem, stressing the evolving role of journalists in producing discussions about digital media literacy. Thus, this article contributes empirical insights to our understanding of social media content regulation and of AI implications in society.

When society meets new media – On technology, dystopia, and utopia

The scope to which new media forms are considered opportunities or potential risks has been discussed extensively both in scholarly and popular contexts (Spilioti, 2016). To unpack journalistic framings of deepfakes in the context of dystopian and utopian narratives, we first need to understand the broader utilization of said rhetoric in sociocultural discourses. The prophetic evaluation of new media and their impact on society dates to early documentation outlets where, for example, Plato, citing Socrates, warned society against the devastating nature of the written (rather than spoken) word. Writing, he proposed, was analogous to the death and decay of societies; holding the potential to destroy human memory, weaken the mind, and lead to personal and social passivity (Ong, 1982). Similar dystopian messages found place in discussions about other media forms and outlets and their meeting places with society. Critics of the spread of the telephone in the early 20th century deemed it a pervasive medium that would shred privacy and promote indecent content (Fisher and Wright, 2001). Concerns were also voiced, for example, regarding the dangerous impact the radio would have on society, undermining social morality (Fischer, 1992).

Utopian predictions about the use of media technology represent the opposite extreme. The hype around different new media outlets attributed liberating, empowering, and democratizing virtues to new media technologies and the content disseminated through them. The television was perceived, at one point, as a medium that physically brought families together into a shared space like never before, a medium that could facilitate social unity through a simple decoding of content (Meyrowitz, 1985; Spigel, 1992). In the same vein, the cinema and films were argued to hold a revolutionary power that might lead society to critical thinking about power, hierarchies, and injustices (Benjamin, 2001).

Internet-based media platforms were also often explained through the hype versus hysteria dichotomy. From early days of Web 1.0 bulletin boards, newsgroups, and multiuser domains to recent uses of instant-message applications, computational image correction, selfies, and social networking websites, utopian and dystopian readings of Internet adoption, so it seems, appear again and again. Dystopian approaches toward the Internet tied connective media usage with aggressive online acts such as rape, deception, and forgery (Dibbell, 1993); loss of social and interpersonal communication abilities (Turkle, 2012); erosion of existing geographical communities (Verschueren, 2006); and tendencies toward narcissism and psychopathy (Fox and Rooney, 2015). Utopian approaches, in comparison, overplayed the liberating nature of Internet-based media – for example, allowing minority groups to overcome decades of social marginalization (Pearson and Trevisan, 2015).

In this manner, '[T]he meanings that are attributed to new technologies', (Sturken and Thomas, 2004) in 2004 proposed, 'are some of the most important evidence we can find of the visions, both optimistic and anxious, through which modern societies cohere' (p. 1). Understanding where and how narratives regarding the promises, whether positive or negative, of media technologies take shape can give us a glimpse into the larger political and cultural questions a society is faced with. Journalism, as a field of practice, affords one important social arena in which these pressing questions are asked.

News, journalists, and frames for understanding new media content regulation

Studying journalistic framing of new media and communication technologies becomes an important space for understanding political, social, and cultural trends, as well as potential influence on audiences (Marciano, 2019). In this context, journalistic framing practices bring forward the salient topics discussed in specific social settings; highlight the importance of specific decoding frames in political, economic, and cultural discussions; and activate ties or associations between specific sociocultural issues and existing decoding structures in society (Entman, 1993;

Iyengar and Kinder, 2010; McCombs and Shaw, 1972). Here, news sources are situated as a prime sphere in which social and political issues are framed, that is – 'shaped, defined, negotiated, and contested' (Greenberg and Hier, 2009: 463), playing a crucial role in public debates over media regulation through crafting, emphasizing, and situating information for both citizens and decision makers (Kosicki, 1993).

Through the act of framing, news media hold the potential to raise awareness of specific topics, direct attention to specific actions, and highlight solutions in contexts such as public, health, and foreign policy (Baumgartner et al., 2009; Cobb and Elder, 1981; Gamson, 2004; Ophir, 2019). Journalists covering policymaking in its broad sense are not merely selecting and treating specific issues in their news coverage but are rather impacting social perceptions and agendas by portraying an issue as pressing while attributing positive or negative frames to it (Iyengar and Kinder, 2010). Thus, news outlets might play an important role in 'disseminating the required information to successful policy diffusion' (Crow, 2012: 38). Given that social media content regulation falls under the broader contextual umbrella of internet policies, it is important to understand the current trends of these two topics. Put in simpler terms, what news sources tell readers about the need for regulatory actions or changes matters. Yet, less academic attention has been given to the ways in which news sources frame Internet policies as a whole and social media content regulation in particular.

Given recent findings about the negative place of online platforms in international public life – whether concerning 'fake news,' illicit user-information exploitation, political manipulation, or intensifying polarization of opinions and hatred (Cision, 2018; Lotan, 2014; Zelenkauskaite and Niezgoda, 2017; Zimmer, 2008) – it is imperative to contextualize the above understandings about news and regulation in the field of Internet studies. To this end, we ask in this study questions about a specific case study that can shed light on news, regulation, and Internet media: What are the journalistic narratives constructed through the coverage of deepfake technology? What are the regulatory actions associated with these offered narratives? And, what might such narratives serve?

Studying social media content regulation through deepfake applications

Since '[C]ontent-related concerns . . . have sparked debates over the appropriate scope of social responsibility that should be required of social media platforms in their gatekeeping and content-curation functionality' (Napoli, 2019: 3); studying discussions about the evolving regulation of deepfake content on social media becomes a promising venue for understanding the future of our digital media ecosystem. Through the review of related literature, we were able to identify two possible gaps that the current article seeks to fill: First, by providing insights about the place of news sources in advocating regulatory actions, we addressed the growing need to unpack the specificities of calls for social media regulation and content moderation. Second, we address the implications of dystopian versus utopian framing in new media coverage, particularly as they relate to manipulation, disinformation, and fakery online. Thus, we contribute to the growing scholarly exploration of the place disinformation and misinformation take in global socio-political discussions through a unique perspective on user-based content appliations. In the following section, we describe in detail the data collection and analysis procedures designed to enable this important exploration.

Data collection and analysis: Tools and methods

To reach a nuanced and in-depth understanding of the journalistic narration of deepfake technology, we have constructed a multilayered data sampling and analysis process combining four automated and manual stages. To this end, we have utilized *Buzzilla*, a big-data system that collects data from multiple online sources based on popularity, among them message boards, blogs, news websites, and social media platforms. Through the system, we have identified a total of 3642 public items in the English language containing the keywords ‘deepfake’ or ‘deep fake’. Focused on journalistic narratives, we have compiled only news items ($n = 988$ news items out of the total 3642 items from web sources), gathered based on *Buzzilla* crawlers of *Alexa* news rank. These items were published over a full year, from the day the application was inaugurated on Reddit in December 2017 up to December 2018.

After achieving this exclusive, yet comprehensive, data set of news items ($n = 988$), we embarked on a second data-sifting stage to achieve a manageable and carefully filtered data set. Given that ‘a headline is . . . the most powerful framing device of the [news-related] syntactical structure’ (Pan and Kosicki, 1993, p. 59), we restricted the data corpus to only include items in which the term deepfake appeared in the headline or sub-headline ($n = 233$). While it can be argued that these two aspects are not mutually exclusive (a terms not being included in the headline but included in the body of a journalistic piece), this sifting aspect allows the exclusion of items that did not focus on the media technology as the main topic or those that mentioned it in passing, which are of less relevancy for the needs of this study. These 233 items were sifted based on natural trend behavior over the examined year to focus on a unique data spikes related to deepfakes that appeared around February, June, and September of 2018.

We then embarked on a third sifting stage, in which we manually read each item published in these 3 months. We discovered that the volume pattern found in February ($n = 120$) was significantly larger than June and September. It was also more diverse in terms of news sources and content as in these June and September news outlets often feature the same news article in two different sections (e.g. technology section and news section), and the same item was circulated by news agencies and published in its entirety by multiple outlets on the same day. Thus, the February circulation spike represented a diverse pool of sources, which made this point in time significantly more fruitful for analysis. After eliminating repetitions in the set, the multilayered automated and manual data collection process yielded the final data corpus of 105 news articles from February 2018, which were independently analyzed by each of the authors through the fourth step – a holistic mode of qualitative narrative analysis.

Holistic narrative analysis

In this study, we employed the holistic mode to understand the story told by journalists through key moments in the development (or ‘life story’) of a media technology, where sections of a text were explored and understood in the context of additional parts of the narrative (Lieblich et al., 1998). To reach these different parts of the narratives constructed by journalists, we analyzed the news items according to the ‘narrative inquiry’ approach (Smith and Sparkes, 2006), by identifying six components that create a frame for holistic interpretive decoding of story patterns in narratives (Mishler, 1986):

1. Abstract: a summary of the story and its points, or, as Mishler (1986) suggests, our interpretation of the story's main point; how we perceive the meaning of the account
2. Orientation: providing a context such as place, time, and character to familiarize the reader with key issues related to the story
3. Complicating action: an event or events that cause a problem or a conflict that express and reflect some broader cultural frameworks of meanings related to the narrative
4. Evaluation: appreciative comments on events, justifications of its telling, or the meaning that the teller gives to an event. Given that the narrative is intended to be an identity-claiming story (Mishler, 1986), the evaluation is where we infer, as researchers, how the narrator perceive the story as a whole.
5. Result or resolution: the closure of the story or the conflict presented earlier
6. Coda: the point of bringing the narrator and listener (or reader) back to a shared present

Together, these six categories helped us unpack the narratives journalists constructed through coverage of the deepfakes phenomenon and answer the questions presented above.

Whose dystopia is it, anyway?

As a whole, deepfake applications were approached by journalists as an extremely negative outlet that carries the potential to harm vulnerable groups, to undermine indicators of a shared reality, and to jeopardize the ability to distinguish between real and fake. Journalists covering the deepfake phenomenon used a cautionary voice to alarm readers of the outcomes this new media form might hold. This voice, in turn, led to an extensive discussion of the regulatory steps necessary for restraining manipulation and spread of disinformation. 'Combine face-swapping tech with the ability to mimic someone's voice', suggested the writers at *The Verge* (February 8, 2018) in this context, 'and you have the potential for misinformation on a catastrophic scale. Donald Trump declaring war on North Korea. Hillary Clinton caught praising the Illuminati . . . our political future is 'hackable' . . . . the end of reality as we know it'. Here, 'deepfakes add a new layer of complexity to what could be used to harass and shame people', argued journalists at *BBC News* (February 3, 2018). Thus, posited a piece in *The Baltimore Sun* (February 19, 2018),

> [I]t's not hard to imagine a world in which social media is awash with doctored videos targeting ordinary people to exact revenge, extort or to simply troll . . . The danger there is not just believing hoaxes, but also dismissing what's real.

In this sense, the story that news sources told their readers (or, the *abstract*) was, for the most part, about social media content regulation: 'Reddit and Twitter ban AI-generated porn "deepfakes"' (*New York Magazine* February 7, 2018), 'Deepfake pornographic videos banned by Twitter' (*BBC News* February 7, 2018) and 'Reddit (finally) bans deepfake communities, but face-swapping porn isn't going anywhere' (*Slate* February 8, 2018) for example. To contextualize the ban on deepfakes, journalists provided their readers with *orientation* to the location, characters, and timeline of the story, building a dystopian image of the future, to the extent of imagining a society governed by disinformation and computer-manipulated content.

As we unpack in the following section, the method of constructing this dystopian image clustered around three main narratives, or three perspectives, related to the societal implications of deepfakes' spread (and hence the need to regulate it): first, reporting on the case of deepfakes from a gendered perspective, through concrete terms related to pornographic content and the possible harm to vulnerable communities (women and children); second, covering the issue from a factual perspective, by contemplating the effect of losing reliable indicators of reality through anxiety-inducing futuristic consequences; and third, employing a professional perspective by negotiating the place and importance of journalists as traditional trustworthy gatekeepers.

The gendered perspective

The gendered narrative presented concrete social concerns regarding the harm to vulnerable groups, namely women and children. The narrative produced under this approach focused primarily on the loss of consent and control over one's representation, where deepfake content regulation (in the context of pornographic videos) will ultimately protect women from abuse and harassment. Here journalists informed readers that '[T]he fake, but realistic "deepfakes" videos show popular actresses such as Emma Watson and Gal Gadot in explicit scenarios, with their faces mapped onto porn stars' bodies using artificial intelligence' (*Breitbart News* February 7, 2018), and '[T]he backlash against deepfake videos has been gathering pace amid concern about their potential for exploitation in revenge porn' (*The Independent* February 7, 2018). Journalists referred to the 'potential abuse for anyone who puts their face online' (*The Verge* February 9, 2018), the reinforcement of 'offensive gender or racial stereotypes' (*CNN* February 8, 2018), and overall fear for innocent female celebrities and children, as 'there has been a backlash following reports that some people had created illegal child abuse imagery by using photos of under-16s' (*BBC News* February 7, 2018). Journalists informed readers that '[W]hat's particularly unsettling here – beyond the obvious complete lack of consent – is the popularity of female celebrities who first found fame as children' (*The Daily Telegraph* February 14, 2018).

In this context, the writers of *VICE* noted that 'If someone uses the faceset that contains images of Watson as a child to make a deepfake, that means that a face of a minor was in part used to create a nonconsensual porn video' (February 27, 2018). They added, 'Sadly, the issue of virtual child porn isn't new, but deepfakes, which crowdsources and automates the process of creating fake videos, makes that issue more complicated'.

This gendered interpretation of deepfakes was prevalent in journalistic coverage across outlets and contextualized social media content regulation in a larger biological sex (females) or gender (women) discourse. 'Pornographic deepfakes are non-consensual and deeply gross', shared the staff of *Cosmopolitan* (February 13, 2018), adding, '[T]hey depict extremely intimate, explicit acts, and are created without permission from the celebrity *or* the porn actor'. 'Think about it like revenge porn or hacked celebrity nudes, which can be weaponised to humiliate a person (usually a woman) in front of millions of people', the writers at *Cosmopolitan* continued, concluding, '[B]ottom line: She doesn't have any say in what her body is made to seem like it's doing, and strangers are getting off on that' (*Cosmopolitan* February 13, 2018).

In this sense, journalists univocally supported regulating deepfake-generated pornographic content. Journalists framed the reported ban as the solution to this 'new' problem. 'For now, banning deepfakes – especially as a tool for revenge porn – is the only recourse', argued the writers at *Fox News* (February 16, 2018). They continued, '[A]nd, we'll likely start hearing about new regulations and laws. Hopefully, programmers will find positive uses for the AI and machine learning before that happens'.

Given that communication technology advancements and pornographic content have a long-standing relationship (Coopersmith, 1998), it is important to unpack the constructed rhetoric of a new, original, or unfamiliar threat and the need to regulate it. On the one hand, finding that multiple news and social media outlets supported the establishing of regulatory approach to social media content through a ban on nonconsensual pornographic depictions of women is extremely encouraging. On the other hand, most writers approached deepfakes as a new type of female objectification, and thus the ban of deepfakes as somewhat of a concluding remark on sexual and gendered exploitation via this media form. In fact, deepfake pornographic content, albeit produced through a new technological application, is not new in the broader sense of the systematic sexual oppression and objectification of women in media, and in society overall.

Why, then, did journalists covered much of the deepfake-phenomenon as a new gendered phenomenon? We propose that this has to do with the 'issue attention cycle' (Downs, 1972), where public discovery of a social issue in news coverage 'is often accompanied by the optimistic belief that, by taking some measures, the problem will be solved' (Shih et al., 2008: 146). Thus, an issue that is brought to public awareness is usually framed through the lens of *resolution*. In the coverage of deepfakes, the censorship of pornographic deepfake content on social media reflects this type of optimistic solution to the problem. Here, journalists describe a techno-societal change – new abilities to abuse women – but in fact ground this change in historical continuity – gender-based aggressive representations and abuse of women's sexuality.

'Many of us now live much of our lives online, yet the internet can be a hazardous home for far too many women', emphasized the writers of *Stylist* (February 2, 2018) in this regard. 'Against a dark backdrop of tech company inaction – not to mention the fact that deepfakes are not currently covered by revenge porn laws', the writers continued, '"websites" decisions to block deepfake videos should be seen as a small spot of light. It makes the internet a fractionally more hospitable place for women, and that can only be a good thing'. Thus, most outlets dismissed the problem of pornographic deepfake content on social media by emphasizing the ban as a means for positive public regulation through control. The narrative might argue for a positive conclusion (protecting women) but posits a gloomy prediction regarding the negative use of technology and the lack of regulatory capabilities to combat abuse online.

Given the role news media can play in policy diffusion (Crow, 2012), we must further unpack the ways journalists frame social media regulation as a whole and that of deepfakes in particular. It seems like journalists propose that merely regulating deepfakes online will help solve sexual exploitation as described in the examples provided above. This is a problematic approach given that communication technologies and their advancements cannot, and should not, be separated from the thick and sticky histories they represent, especially in the context of sociopolitical equality. The gendered framing of new communication technologies as a threat must also include a discussion of the status of legislation regarding sexual assaults, nonconsensual sex, and gender inequality. Instead, what journalists focused on was the narrative, by which Internet regulation, on its own, might be able to stop, even for a while, abuse of women. Yet, this was not the only narrative employed by journalists' reporting on deepfakes. According to journalists, while content regulation of deepfakes might help protect children and women in the context of pornographic content, what it cannot do is save humanity from the loss of a sense of truth and shared reality. This narrative is further unpacked in the following paragraphs through the factual perspective.

The factual perspective

If the gendered perspective touched on the victims of the studied phenomenon, the factual perspective constructed by journalists looked at manipulation and misinformation as an issue related to human–machine interaction and regulation. More specifically, in the journalistic coverage of deepfakes, the factual narrative depicted dystopian social occurrences related to humans' interaction with manipulated content, ones that erode markers of truth and blur the line between real and fake. Through this perspective, journalists considered online realms as unregulated domains representing extremely immoral and dangerous behaviors that risk societal grasp on a shared reality; hence, they need to be regulated.

In this context, it was interesting to find that journalists narrated deepfakes as a politically and socially counterproductive phenomenon. Deepfake content was framed by journalists as videos 'made possible by neural network technology that can learn the features of anyone's face and map it onto bodies in videos' (*The Register* February 9, 2018), or as consumer applications 'using a machine learning algorithm' (*VICE* February 7, 2018) that 'will appear on Apple Store and Google Play in the near future, offering video alteration through neural networks for general consumer amusement' (*Fox News* February 16, 2018). Thus, much of the coverage went well beyond human-created content to discuss the agency of machines.

The Hamilton Spectator, for example, informed its readers that 'AI systems are increasingly adept at generating believable audio and video on their own', a fact that, according to them, will 'make it easier for bad actors to spread misinformation online' (February 21, 2018). 'The trend is disturbing on a few levels', argued the writers at *TechCrunch* (February 7, 2018), who added, 'machine learning technology that makes these kind of manipulated videos work will only grow more sophisticated over time, making it even harder for internet-goers to determine what's real and what's fabricated' (*TechCrunch* February 7, 2018).

'Combine face-swapping tech with the ability to mimic someone's voice, and you have the potential for misinformation on a catastrophic scale', declared the writers of *The Verge* regarding the topic. 'This breeds a sort of media nihilism, a belief that no audiovisual content can ever be definitively said to be "real"' (*The Verge* February 8, 2018). 'How could we possibly know what is real?' asked journalists at *Media Post*, concluding, 'Thomas Jefferson wrote that, "A well-informed electorate is a prerequisite to democracy". Imagine if we could not trust any video or audio to be authentic' (February 20, 2018). 'In that scenario, where Twitter and Facebook are algorithmically flooded with hoaxes, no one could fully believe what they see', added the writers at the *Baltimore Sun* (February 19, 2018). 'Truth, already diminished by Russia's misinformation campaign and President Trump's proclivity to label uncomplimentary journalism "fake news," would be more subjective than ever', they continued, and concluded, 'The consequences could be devastating for the notion of evidentiary video, long considered the paradigm of proof given the sophistication required to manipulate it' (*Baltimore Sun* February 19, 2018).

Similarly, *VentureBeat*'s writers asked their readers,

> Will these deepfake videos evolve to the level whereby no human, no matter how smart, could differentiate real footage from fake footage? And if that's where we're headed, shouldn't we really be having conversations about if it's where we want to go?. (February 8, 2018)

These questions, it seems, were not rhetorical. Rather, *VentureBeat*, like most other news outlets, pointed at a possible crisis of losing a shared sense of truth and a sense of a shared reality. It is important to acknowledge here that the ability to determine what is reality and define the elusive term 'truth' goes well beyond the scope of this article. However, the rhetoric used by journalists in news items does teach us about the new factual, almost existential, predicament that occupies our contemporary public sociopolitical discourse.

The Verge expanded on this rhetoric: 'Experts taking a more pessimistic view say that this technology is going to improve to the point where humans can't tell the difference between AI fakes and real footage' (February 8, 2018). *TechCrunch* took this argument even further:

> If you're the target of malicious propaganda you'll very likely find the content compelling because the message is crafted with your specific likes and dislikes in mind. Imagine, for example, your trigger reaction to being sent a deepfake of your partner in bed with your best friend. That's what makes this incarnation of propaganda so potent and insidious vs. other forms of malicious disinformation (of course propaganda has a very long history – but never in human history have we had such powerful media distribution platforms that are simultaneously global in reach and capable of delivering individually targeted propaganda campaigns. That's the crux of the shift here). (February 18, 2018)

In these excerpts and others, the acknowledgment (and to a large extent, the fear) of machines being able to perform human tasks is amplified through depictions of content personalization via 'machine learning', 'neural networks', and 'algorithms' as the new, unrestrained structure enabling deepfakes. In these cases, Internet content is attributed an existence that is independent of 'real', 'authentic', or 'human' factors, working free from human intervention, but with a frightening resemblance to human neurological systems and learning abilities. Thus, the issue here was not merely that deepfake content 'dangerously blurs the lines between reality and mimicry' (*Cosmopolitan* February 13, 2018), but rather that it might get to a point where the reality we consume via online platforms is personalized to the point of abuse.

These journalistic accounts built a clear narrative: machine agency leads to the demise of audiovisual content as an indicator of reality, which in turn leads to the loss of a shared social sense of truth, and the undermining of democratization of not only online media spaces, but the international political arena as a whole. 'From here on in', reporters told their readers, 'the question of "Is it real?" will linger over anything particularly outrageous – whether Russian *kompromat* of an American president, or something as prosaic as a bit of showy science' (*The Register* February 12, 2018).

Deepfakes were framed by journalists as the sum of threats posed by combining machine learning, algorithmic content personalization, and political weaponization. Deepfakes both distort reality markers and allow new forms of covering up political messes and dishonesty. Similarly to what we found in the gendered perspective, regulation of deepfake content on social media was advocated by news sources in the factual context. However, as we unpack in the following section, for journalists, deepfake content becomes a double-edged sword. While this content puts society at risk, deepfakes in fact work as a handy method for restoring faith in the journalistic practice.

The professional perspective

A third narrative constructed by news sources in the context of deepfakes turned to evaluating the need for skilled professionals in a time of technological, social, cultural, and political uncertainty. To better understand this narrative, we first need to unpack the concept 'fake news', as it was used extensively by journalists throughout the coverage of the deepfake phenomenon, contextualizing AI fakery as part of a larger trend of manipulations and disinformation in the media ecosystem.

The term 'fake news' signifies the questioned status of journalistic credibility in a post-truth era. Overlapping with other negative online information-spreading phenomena such as misinformation and disinformation, the creation of fake news erode traditional journalistic standards, editorial norms, and gatekeeping positionality (Lazer et al., 2018; Tandoc et al., 2018) alongside concepts such ‘truth’, ‘reliability’, and ‘credibility’, which are traditionally associated with the journalistic field (Blitz, 2018; Gingras, 2018). The technological fabrication created by deepfake applications was tied by journalists to the fake news phenomenon, where positioning deepfakes as a threat to society became a functional tool journalists used to renegotiate their role and importance in Western societies.

‘In an age of fake news and filter bubbles, these instances [of spreading deepfake content] shred our collective sense of reality and could continue to create distance between agreed-upon facts’, determined the writers of *VentureBeat* (February 10, 2018). They continued, ‘though there is some good deepfake detection software available today, we don’t know yet if tech platforms can keep everyone safe’. ‘Just think how much worse it will be when fake news becomes fake video’ stressed the writers of *CNN* (February 8, 2018). And while ‘for individuals whose likenesses appear in these videos without their consent, enlisting the help of gatekeepers like Twitter will be the best way to stop this content from spreading’ (*The Verge* February 7, 2018), journalists agreed that ‘AI-moderators fighting AI-generated porn is the harbinger of the fake news apocalypse’ (*VICE* February 16, 2018).

As part of the hyper-dystopian narrative constructed by journalists, reporters also provided their audiences with a preview of a fake-saturated political future, where ‘we could see similar pseudo-videos that are used to spread disinformation, panic, and fear in the same way as we’ve witnessed with various recent ‘fake news’ scandals’ (*VentureBeat* February 8, 2018). The writers of *VICE* (February 21, 2018) continued this line of thought, adding that ‘one need only imagine a fake video of Trump declaring war on North Korea surfacing in Pyongyang and the fallout that would result to understand what is at stake’. Bringing together the two types of fakery – fake news and deepfakes – was not unintentional. We argue that fusing these two related, yet separate, social phenomena allowed journalists to negotiate and validate their own importance in society alongside the need to regulate social media content.

‘There are plenty of concerns, though, about the dangers posed by politically-themed faceswap videos, especially in the era of Russian interference and the proliferation and weaponization of “fake news”’, explained the writers of *Mashable* (February 20, 2018), mixing up the dystopian technological future with the political past. ‘More importantly, though’, concluded the writers at *The Verge* (February 8, 2018),

> there’s a danger that by hyping the threat of AI fakes, we increase its influence. Think about how the label ‘fake news’ was applied overzealously by the media, becoming a buzzword without a clear meaning. Pretty soon, it was turned against outlets by the same populist and partisan forces whose power it was intended to blunt. In the short term, the actual technology of AI fakery might be less of a threat than its perception. Like ‘fake news,’ it will become a shield for liars and conspiracy theorists, used to dismiss any evidence that runs counter to their own beliefs. In the age of AI, the next ‘grab them by the pussy’ video will be even more easily shrugged off as a fake under a miasma of reasonable doubt. This breeds a sort of media nihilism, a belief that no audiovisual content can ever be definitively said to be ‘real’.

With an overall mistrust in the media industry (and within it, news media; see Edelman Research, 2019), the data analyzed in this study reveal that for journalists, fakery (as a contemporary online-media phenomenon) became practical in contesting concerns regarding the practice of journalism in the age of post-truth. ‘We need to counter this sort of alarmist thinking, and the recent clampdown on AI fake porn is a salutary example in this fight’, argued journalists at *The Verge* (February 8, 2019). They went on to say, ‘It shows that gatekeepers’ authority to police content doesn't evaporate just because it was made by machine learning', and added, '[A]nd unlike other categories of questionable content, which platforms have had trouble defining and therefore limiting (e.g. hate speech and abuse), AI fakes present a more clear-cut category' (*The Verge* February 8, 2019). The concerns regarding deepfakes' societal implications (harm to vulnerable groups, loss of a shared sense of reality, harbingering apocalypse) and supportive attitudes toward an overall censorship policy became handy for advocating the need for traditional gatekeeping institutions, such as journalism. Thus, deepfakes, which were framed as a *complicating action* or a threat in the narratives constructed by journalists, also became a means for restoring the relationship between the media, the public, and politicians. Thus, in terms of *coda*, journalists left readers with a projection about mending the place of traditional journalistic gatekeepers against the many levels of fakery experienced in society nowadays.

By reading the dystopian coverage of deepfake content on social media through this perspective, it is possible to argue that painting an almost apocalyptic future was functional for restoring faith in journalists' work. In a way, pinning the blame for the many levels of doubt and disbelief in society today on technology relieved journalists from criticism about the appearance of fake news and positioned them, alongside the public and politicians, as victims. This rhetoric created ideological and social distance between fakery and journalistic practices. It is possible to argue that by scapegoating malicious Internet applications, journalists were able to reinstate their important role in society. Inseparable of this discussion was the need to regulate human–machine interactions and social media content, identified in all three narratives found in this study.

Conclusions

Through the interpretive narrative analysis of news items covering deepfake technology, we have discussed the construction of three possible narratives resulting from three perspectives, all of which imagined a techno-dystopian future in one way or another, necessitating social media content regulation. The first was a gendered perspective that positioned deepfakes as a threat to vulnerable groups such as women and children. This threat came to resolution (according to news reporting) through social media content regulation (in the form of a ban) employed by social media platforms. The second narrative looked at the issue of deepfake applications from a factual position, imagining the future of a society that loses a shared understanding of reality and truth and the ability to differentiate between humans and machines. In this sense, journalists constructed an understanding that regulating deepfake content on social media becomes a primary way to ensure a shared sense of reality among people, where facts are perceived as a core evidentiary component of society and can be reached via representations of reality (such as in videos). The third perspective on the topic, presenting journalists' arguments regarding their own place within society, used deepfake content and its possible societal implications as a means for restoring trust in journalists as sociopolitical gatekeepers in our digital era of post-truth.

Before discussing the meanings of these narratives and the article's main contributions, we wish to highlight the potential limitations of the study. First, given the analyzed time frame (one year), the insights provided here lack a longitudinal perspective that future studies can benefit from. Moreover, in this study, we have focused on mainstream news media sources. While this allowed an in-depth reading of the journalistic discourse and motivations, future research can benefit from utilizing a comparative lens to better understand how additional groups and information sources – such as technology and/or feminist bloggers or alternative news sources – narrate the phenomenon. Finally, future research, hopefully even our own research, should also focus on analyzing the voices of consumers and producers of deepfake materials, and the ways they explain and understand the phenomenon to provide a panoramic perspective on the topic.

With these limitations in mind, this study also presents a novel data collection strategy that is based on a human–computer hybrid approach. In qualitative research, data collection can be a labor-intensive task, particularly when it relies on manual collection and analysis of media items in libraries or archives offline and online. Similarly, attaining reliable, consistent, and valid research data through coding and web crawling requires extensive technological skills and expertise. In this context, utilizing social media monitoring software, such as in this study, provides an effective source for gathering large quantities of information about a focused topic, in a timely manner, without the need to use elaborated code. Our method for selecting, evaluating, filtering, and testing data items, and in turn building the data corpus, has substantial practical benefits as it enabled us to obtain a consistent data set in a relatively short time, with the methodological rigor required for qualitative research study. By utilizing big-data tools for building a thick data set (Latzko-Toth et al., 2017; Wang, 2016), we were able to reach both the breadth and scope of big-data collection as well as the holistic, meaningful, and context-grounded depth of small-scale qualitative analysis.

The finding yielded from the abovementioned analysis lead us to a discussion of two main conclusions, one about the theoretical perception of online platforms and one about the practical role of news sources in our post-truth digital era, as reflected in the relationship between the three found narratives. First, attention should be paid to the ways in which the journalistic discussions about deepfakes help us think of accountability and responsibility in digital culture. With lines between media consumers and producers getting increasingly blurred online, it is important to address the intersection of representation, minorities, and content dissemination via Internet platforms. In this context, the rise of ‘the former audience’ or ‘produsers’ (Bruns, 2005; Gillmor, 2006) not only allow users to take part in the creation of content but also put the traditional ‘gatekeeping’ function of media professionals in question. While some find online participatory features important for democracy, inclusion, and positive self-representation (Jenkins, 2006; Yadlin-Segal, 2018), the case study presented in this article reveals that more consideration should be given to ethical and moral responsibility of media content production online (or lack thereof), and specifically to the expansion of these features from traditional media producers to media users.

Thus, when considering the meeting place between the gendered perspective (looking at the issue of exploitive minority representation) and the professional perspective (looking at issues of accountability and gatekeeping), we find that journalists raised an important question – with more and more media representations produced by users, did the accountability for these representations also expand? The interaction between the professional and the gendered perspectives reveals that the answer is no. Deepfakes content that is spread over social media platforms lacks the enacted responsibility against exploitation of vulnerable groups, manipulation, and disinformation that are required from mainstream forms of media. While the narratives regarding destabilizing nature of deepfake content can be seen, at points, as exaggerated; the perspectives about accountability for generating and disseminating AI content as a whole, and in the context of minority or marginalized groups in particular, require a renewed thinking about online spaces as liberating, democratizing, and freeing.

Yet, when it comes to the fear-inducing factual narrative constructed by journalists, the dystopian tone of discussing deepfakes might overshadow an important call for shared societal thinking of media accountability and ethics. In the context of techno-dystopia, Sturken and Thomas (2004) argued that a ‘society’s capacity to project concerns and desires on technology operates as a primary form of social denial’ (p. 3). Journalists’ dystopian narration of the deepfake content, and in turn, the instrumentality of banning it for a shared greater good, calls society to cohere around a shared threat. If dystopian accounts are considered social denial, then the dystopia constructed through journalistic coverage of deepfakes and the regulatory aspects related to them reflect a professional self-denial. Instead of calling for a renewed understanding of the terms ‘real’, ‘authentic’, and ‘credible’, and responsible regulation of social media content in a new, evolving media ecology; journalists stick to their professional guns and reiterate the traditional real versus fake dichotomy associated with the meeting place between new media and society.

Considering internet regulation, as related to the public policymaking and the news, we suggest that instead of running away in fear and dystopian denial (Sturken and Thomas, 2004), journalists can restore the credibility of news through advocacy. This can be seen as the practical aspect of this concluding discussion, where suggestions can be made based on the empirical data analyzed in this research study. Thus, alongside regulation enacted by social media platforms (which is important in of itself), in the context of new media as a whole, and AI political and cultural implications in particular, journalists can assume a social role and advocate for an educated and critical use of the new technologies. This is crucial to understand, as ‘[M]edia makers apply a range of persistent frames, and as such they possibly control the number of alternatives that are available to the receivers when they are constructing social reality’ (van Gorp, 2006: 62). Journalists can shift the dystopian framing to promote consensual, fair, and respectful representations (of women or otherwise) via deepfakes. They can surface the possible positive potential of using deepfakes and call for diffusion of additional technological tools that will help us sort out and make sense of deepfake content. They can promote initiatives and public policy geared toward technological literacy in an age of deep-learning content, ones that will help the public gain tools for making sense of online information in our digital era. Such promotion should focus on understanding technology, not only predicting its best or worst outcomes. This responsibility, we argue, will not only help reenvision the journalistic practice in the age of post-truth but will also help facilitate an important understanding of the place of human agency and the reciprocal relationship between humans and machines.

ORCID iD

Aya Yadlin-Segal https://orcid.org/0000-0002-7199-6664

References

Baumgartner FR, Linn S, and Boydstun AE (2009) The decline of the death penalty: How media framing changed capital punishment in America. In: Schaffner BF and Sellers P (eds) *Winning With Words: The Origins and Impact of Framing*. New York: Routledge, pp. 159–184.

Benjamin W (2001) The work of art in the age of mechanical reproduction. In: Leitch VB (ed) *The Norton Anthology of Theory and Literary Criticism*. New York: Norton, pp. 1166–1186.

Blitz MJ (2018) Lies, line drawing, and (deep) fake news. *Oklahoma Law Review* 71(1): 59.

Bruns A (2005) Axel Bruns at iDC. Institute for distributed creativity. Available at: https://eprints.qut.edu.au/4863/1/4863_1.pdf (accessed 26 April 2020).

Cision (2018) *State of the Media Report*. Available at: https://investors.cision.com/2018-04-24-Cision-State-of-the-Media-Report-Reveals-Rising-Public-Trust-Amidst-Journalisms-Fake-News-Battle (accessed 26 April 2020).

Cobb RW and Elder CD (1981) Communication and public policy. In: Nimmo D and Sanders KR (eds) *Handbook of Political Communication*. London: SAGE, pp. 391–416.

Coopersmith J (1998) Pornography, technology and progress. *Icon* 4: 94–125.

Crow DA (2012) Policy diffusion and innovation: Media and experts in Colorado recreational water rights. *Journal of Natural Resources Policy Research* 4(1): 27–41.

Dibbell J (1993) A rape in cyberspace. *The Village Voice*. Available at: http://www.villagevoice.com/news/a-rape-in-cyberspace-6401665 (accessed 26 April 2020).

Downs A (1972) Up and down with ecology – The 'issue-attention cycle'. *The Public Interest* 28: 38–51.

Edelman Research (2019) *2019 Edelman Trust Barometer Global Report*. Available at: https://www.edelman.com/trust-barometer (accessed 26 April 2020).

Entman RM (1993) Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4): 51–58.

Fischer CS (1992) *America Calling: A Social History of the Telephone to 1940*. Berkeley: University of California Press.

Fisher DR and Wright LM (2001) On utopias and dystopias: Toward an understanding of the discourse surrounding the Internet. *Journal of Computer-Mediated Communication* 6(2). Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1083-6101.2001.tb00115.x (accessed 26 April 2020).

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

Fox J and Rooney MC (2015) The dark triad and trait self-objectification as predictors of men's use and self-presentation behaviors on social networking sites. *Personality and Individual Differences* 76: 161–165.

Gamson W (2004) Bystanders, public opinion and the media. In: Snow DA, Houle SA, and Kriesi H (eds) *The Blackwell Companion to Social Movements*. London: Wiley-Blackwell, pp. 242–261.

Gillmor D (2006) *We the Media: Grassroots Journalism by the People, for the People. Sebastopol*. California: O'Reilly Media.

Gingras R (2018) News then, news now: Journalism in a digital age. In: *Growing Between the Lines* (Osservatorio Permanente Giovani-Editori), Bagnaia, Italy, 25 May 2018.

Greenberg J and Hier S (2009) CCTV surveillance and the poverty of media discourse: A content analysis of Canadian newspaper coverage. *Canadian Journal of Communication* 34(3): 451–486.

Iyengar S and Kinder DR (2010) *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.

Jenkins H (2006) *Convergence Culture: Where Old and New Media Collide*. New York: NYU.

Kosicki GM (1993) Problems and opportunities in agenda-setting research. *Journal of Communication* 43(2): 100–127.

Latzko-Toth G, Bonneau C, and Millette M (2017) Small data, thick data: Thickening strategies for trace-based social media research. In: Sloan L and Quan-Haase A (eds) *The SAGE Handbook of Social Media Research Methods*. Los Angeles: SAGE, pp. 199–214.

Lazer DM, Baum MA, Benkler Y, et al. (2018) The science of fake news. *Science* 359(6380): 1094–1096.

Lieblich A, Tuval-Mashiach R, and Zilber T (1998) *Narrative Research: Reading, Analysis, and Interpretation*, Vol. 47. Los Angeles: SAGE, pp. 199–214.

Lotan G (2014) Israel, Gaza, war & data: Social networks and the art of personalizing propaganda. *Huffington Post* 8 July. Available at: https://www.huffpost.com/entry/israel-gaza-war-social-networks-data_b_5658557 (accessed 26 April 2020).

Marciano A (2019) The discursive construction of biometric surveillance in the Israeli press: Nationality, citizenship, and democracy. *Journalism Studies* 20(7): 972–990.

McCombs ME and Shaw DL (1972) The agenda-setting function of mass media. *Public Opinion Quarterly* 36(2): 176–187.

Meyrowitz J (1985) *No Sense of Place*. New York: Oxford.

Mishler EG (1986) The analysis of interview-narratives. In: Sarbin TR (ed) *Narrative Psychology*. New York: Praeger, pp. 233–255.

Napoli PM (2019) User data as public resource: Implications for social media regulation. *Policy & Internet* DOI: 10.1002/poi3.216.

Ong WJ (1982) *Orality and Literacy*. London: Methuen & Co. Ltd.

Ophir Y (2019) The effects of news coverage of epidemics on public support for and compliance with the CDC – An experimental study. *Journal of Health Communication* 24(5): 547–558.

Pan Z and Kosicki GM (1993) Framing analysis: An approach to news discourse. *Political Communication* 10(1): 55–75.

Pearson C and Trevisan F (2015) Disability activism in the new media ecology: Campaigning strategies in the digital era. *Disability & Society* 30(6): 924–940.

Shih TJ, Wijaya R, and Brossard D (2008) Media coverage of public health epidemics: Linking framing and issue attention cycle toward an integrated theory of print news coverage of epidemics. *Mass Communication and Society* 11: 141–160.

Smith B and Sparkes AC (2006) Narrative inquiry in psychology: Exploring the tensions within. *Qualitative Research in Psychology* 3(3): 169–192.

Spigel L (1992) *Make Room for TV: Television and the Family Ideal in Postwar America*. Chicago: University of Chicago Press.

Spilioti T (2016) Digital discourses: A critical perspective. In: Georgakopoulou A and Spilioti T (eds) *The Routledge Handbook of Language and Digital Communication*. New York: Routledge, pp. 133–145.

Sturken M and Thomas D (2004) Introduction. Technological visions and the rhetoric of the new. In: Sturken M, Thomas D, and Ball-Rokeach SJ (eds) *Technological Visions*. Philadelphia: Temple University Press, pp. 3–18.

Tandoc EC, Lim ZW, and Ling R (2018) Defining 'fake news': A typology of scholarly definitions. *Digital Journalism* 6(2): 137–153.

Turkle S (2012) *Alone Together: Why We Expect More From Technology and Less From Each Other*. New York: Basic Books.

van Gorp B (2006) The constructionist approach to framing: Bringing culture back in. *Journal of Communication* 57(1): 60–78.

Verschueren P (2006) From virtual to everyday life. In: Servaes J and Carpentier N (eds) *Towards a Sustainable Information Society: Deconstructing WSIS*. Bristol, UK: Intellect Books, pp. 169–184.

Wang T (2016) Big Data needs Thick Data. Web log post, 20 January. Available at: https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7.

Yadlin-Segal A (2018) What's in a smile? Politicizing disability through selfies and affect. *Journal of Computer-Mediated Communication* 24(1): 36–50.

Zelenkauskaite A and Niezgoda B (2017) 'Stop kremlin trolls:' Ideological trolling as calling out, rebuttal, and reactions on online news portal commenting. *First Monday* 22(5). Available at: https://uncommonculture.org/ojs/index.php/fm/article/view/7795/6225 (accessed 26 April 2020).

Zimmer M (2008) The externalities of search 2.0: The emerging privacy threats when the drive for the perfect search engine meets web 2.0. *First Monday* 13(3). Available at: https://firstmonday.org/article/view/2136/1944 (accessed 26 April 2020).

Author biographies

Aya Yadlin-Segal, PhD, is a lecturer in the Department of Politics and Communication at Hadassah Academic College. Her research explores the roles new media platforms play in processes of cultural negotiation in a globalized mediascape.

Yael Oppenheim, MA, is a PhD student in The Faculty of Management at The University of Haifa. She is a market and users researcher and consultant who specializes in digital and social media analysis. In her work, Oppenheim employs quantitative and qualitative methods based on innovative digital tools to conduct content and discourse analysis, sentiment analysis, topic analysis, and more.