Routledge
Taylor & Francis Group

Information, Communication & Society

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rics20

The tensions of deepfakes

Benjamin N. Jacobsen & Jill Simpson

To cite this article: Benjamin N. Jacobsen & Jill Simpson (13 Jul 2023): The tensions of deepfakes, Information, Communication & Society, DOI: 10.1080/1369118X.2023.2234980

To link to this article: https://doi.org/10.1080/1369118X.2023.2234980

Published online: 13 Jul 2023.

INFORMATION, COMMUNICATION & SOCIETY
https://doi.org/10.1080/1369118X.2023.2234980

The tensions of deepfakes

Benjamin N. Jacobsen[a] and Jill Simpson[b]

[a]Department of Geography, Durham University, Durham, UK; [b]University of York, York, UK

ABSTRACT

In recent years, deepfakes have become part and parcel of contemporary algorithmic culture. It is regularly claimed that they have the potential to introduce novel modes of societal disruption, violence, and harm. Yet, over-emphasising the power of deepfakes risks occluding frictions, struggles, and logics that already persist in the digital landscape. Arguing for a conceptualisation of deepfakes as an assemblage of differential tensions in society, we explore how they represent both a rupture and a continuation of the variegated politics of the image in the social world. The paper analyses the tensions of deepfakes through three distinct case studies: bodies, politics, and ideas of objectivity. Ultimately, we argue that the tensions and ethicopolitical implications of deepfakes are not reducible to a problem that can be solved through a logic of algorithmic detection and verification.

ARTICLE HISTORY
Received 30 September 2022
Accepted 14 June 2023

KEYWORDS
Algorithms; deepfakes; image; social media; detection; generative adversarial networks (GANs)

Introduction

Our media landscape has recently seen the emergence and proliferation of so-called deepfakes. Indeed, they have become part and parcel of our contemporary algorithmic culture (Hallinan & Striphas, 2016; Striphas, 2015). Broadly stated, deepfakes refer to photorealistic images, videos, or voice recordings that have been algorithmically generated or manipulated. In many cases, they have been used to algorithmically transpose the faces of women (often female celebrities) unto the bodies of unknown others, often in pornographic contexts (Compton, 2021). They have also been seen as highly problematic in political contexts. The reason being that in, say, a deepfake video political figures can, in theory, be altered to say just about anything. The tools and techniques to generate deepfakes have also become increasingly sophisticated and accessible, thereby raising a series of political and civic concerns (Yadlin-Segal & Oppenheim, 2021), with some warning of an impending 'infocalypse' (Schick, 2020). Others have called for 'an anticipatory approach' to deepfakes as a form of intervention in political contexts such as presidential elections (Diakopoulos & Johnson, 2020). As a result, discussions around deepfakes have mainly focused on issues such as gender discrimination, political legitimacy, misinformation, transparency, integrity, and trust. These discussions are often

CONTACT Benjamin N. Jacobsen benjamin.jacobsen@durham.ac.uk Department of Geography, Durham University, Durham DH1 3LE, UK

underpinned by a general anxiety about deepfakes and how they may become a disruptive model of visualisation and (mis)representation in society. Yet, as some scholars have also emphasised, current debates around the multifaceted impact of deepfakes in society are built around concepts, such as agency and the fact-fiction dichotomy, that are neither fixed nor settled (De Vries, 2020).

Diverse aspects of the power and politics of machine learning algorithms have been well documented in recent critical scholarship (Amoore, 2020; Beer, 2017; Bucher, 2018; Chun, 2021; Kitchin, 2017; Willson, 2017). The same goes for deepfakes. But while deepfakes arguably represent a rupture in society's trust of the image, they can also be said to represent a continuation in the subjectivity of all images, the indeterminacies and contingencies fundamental to all modes of image making (Crary, 1990; Mitchell, 2005). Fundamentally, therefore, they perpetuate tensions inherent in all visual representations of the world, algorithmic or otherwise (Uricchio, 2011). While photographic and filmic images are often portrayed as providing objective and truthful accounts of events, they can never provide a view from no-where (Haraway, 1988). As John Berger (1972) reminds us, 'the relation between what we see and what we know is never settled' (p. 7). Indeed, the question of whether something is a deepfake or not is not always settled nor clear.

We understand deepfakes as *an assemblage of differential tensions in society*. By 'assemblage', we emphasise, following Jane Bennett (2010), 'the ad hoc groupings of diverse elements' in the context of deepfakes and how these interact, interweave, overlap, and constitute frictions. Deepfakes are never simply one thing, but are rather the uneasy amalgamations of algorithms, data, media, socio-cultural and political contexts, financial and other incentives, ideas and applications of detection, as well as societal harms, both real and perceived. Drawing on David Beer's (2023) recent work in *Tensions of Algorithmic Thinking*, we think deepfakes through a notion of 'tensions'. Beer's book aims to 'think about the tensions that arise from the competing forces at play in the advancement and application of automation of different types' (p. 4), such as the competing push towards humanlessness and having 'a human-in-the-loop' in algorithmic systems. In other words, Beer investigates 'the push and pull around automation' (p. 7). Similarly, by thinking through the tensions of deepfakes, we want to foreground how deepfakes reflect and embody multiple social, cultural, and political competing forces that already persist in society. In other words, this article explores how deepfakes represent both a rupture and a continuation of the variegated politics of the image in the social world. Ruptures and perpetuations, discontinuities and continuities. By analysing the tensions of deepfakes we not only gain a better understanding of the specific ways they function (how they are generated and deployed) but also of some of the underlying tensions, ambivalences, and issues characterising contemporary algorithmic society.

The paper analyses the tensions of deepfakes through three distinct case studies: bodies, politics, and ideas of objectivity. The rationale for selecting these case studies is mainly a question of domains: where deepfakes are mainly used, where they are perceived to be used, and where their solutions are currently being developed and implemented. The 'Bodies' section explores deepfakes in the context of pornography, which is where deepfakes were initially observed and are still mostly used. Here we show how deepfakes have the capacity to intensify the exploitation, abuse, and objectification of women in digital spaces through the emergence of non-consensual deepfake pornography. Yet, more crucially, we argue deepfakes occupy a contested space between

rupture and continuity in this context, that is, between female objectification as always emerging and always already having emerged. 'Politics' was selected as case study because this is one of the main concentration points of anxiety in society. As Graham Meikle (2023) aptly put it, 'synthetic media [deepfakes] are so far mostly *used* to create non-consensual porn, and mostly *imagined* to be used in politics and news' (p. 4). It is therefore crucial we interrogate these deepfake imaginaries. Here we showcase how the potential impact of deepfakes has been conceived in the political domain. We claim that not only is history filled with altered images, but that anxieties around the societal impact of deepfakes can occlude how politics is always already fundamentally antagonistic and difficult. Lastly, we examine the idea of 'objectivity' because one of the main institutional and industry responses to deepfakes has been the idea of algorithmic detection as solution. In this case study, we showcase how deepfakes are often understood as a problem that can be solved through an array of algorithmic detection techniques, which have often been developed by tech companies such as Meta and Google. We argue that the attempt to 'solve' the problem of deepfakes through a framework of algorithmic detection can constitute a problematic attempt to erect clear boundaries between the real and the fake.

Through a critical examination of the tensions of deepfakes – their competing forces, continuities and discontinuities – we suggest that there is a need to 'stay with the trouble' of deepfakes (Haraway, 2016). This means not reducing the ethico-political implications of deepfakes to a problem that can be solved through a framework of algorithmic detection. Differently put, they are not reducible to a problem that is solvable through computational means. As we argue, deepfakes have the potential to disrupt which, in turn, raises the question, *what is being obfuscated by this claim to disruption*? Staying with the tensions of deepfakes is to remain attentive to their inherent plurality and multiplicity, how they constitute both ruptures as well as continuities in relation to broader societal issues such as gender, politics, and notions of objectivity.

Generating deepfakes

How are deepfakes made? How do they come to matter in society? One of the main reasons for the proliferation of deepfakes has been the emergence and widespread availability of deep learning algorithms such as generative adversarial networks (GANs) and variational autoencoders (VAEs). GANs, for instance, are a particularly popular and widespread example of such generative algorithms, especially in the context of deepfakes. The core idea is to simultaneously train two neural network algorithms (a 'discriminator' network and a 'generator' network) through an adversarial and iterative process. Over time, the generator learns to produce increasingly realistic outputs and, as a result, the discriminator finds it increasingly difficult to distinguish between the real data it was trained on and the data produced by the generator (Goodfellow et al., 2014). GANs are an example of a 'deep generative model' (Goodfellow et al., 2016) along with VAEs and more recent diffusion models such as DALL-E 2 and Midjourney. Trained on a training dataset, these generative models iteratively learn to model a set of inputs and, in turn, output a similar set of data examples (Offert, 2021). This means that they are capable of generating new data points – whether that is text, images, videos, or new features within data points. These new data are approximations, but not complete

replications, of the data on which the algorithm has been trained. Instead of learning how to discriminate between data (such as in the case of facial recognition or fraud detection), deep generative algorithms learn to create new data. In other words, algorithms such as GANs and VAEs are trained to generate synthetic data that resemble the data they are trained on – whether that is a training dataset of dogs, cats, landscapes, or human faces. The aim with generative models is often to produce synthetic or 'fake' outputs that are as proximate as possible to the desired outputs.

In the paper 'You never fake alone: Creative AI in action', Katja De Vries (2020) seeks to conceptualise the role of what she calls 'creative AI', referring to the capacity of algorithmic models such as GANs to generate realistic photo and video representations. De Vries outlines three types of ways in which creative AI is mostly used: understanding, representation, and creation. That is, the capacity of models to generate synthetic outputs is said to signal that they understand the world and are able to represent (or be representative of) it in some particular way (p. 2116). Creation, however, refers to the capacity to generate objects, images, or videos that 'stand on their own' (p. 2117). However, the power of generative models or creative AI, de Vries points out, is how all three purposes overlap and are simultaneously implicated in the creation of new data. Whereas for some, machine learning models such as GANs can be understood predominantly as algorithmic 'copy machines' (Zeilinger, 2021), for others they have the potential to disrupt common conceptions of newness, creativity, art as well as the nature and creative role of 'the artist' (Zylinska, 2020).

However, it has been observed that early GAN models, although capable of generating outputs that resembled the training data, achieved only limited results in relation to photorealism, especially with regards to human faces (Offert, 2021). One of the principal reasons for this was the kinds of images GANs were originally trained on. The issue with the traditional GAN architecture was that it was able to produce relatively realistic images when trained on smaller and lower-dimensional datasets, such as the black-and-white images of MNIST handwritten digits (Deng et al., 2009). What made possible the algorithmic generation of photorealistic deepfakes was the introduction of modified versions to the traditional GAN architecture such as StyleGAN and StyleGAN2, proposed by researchers at NVIDIA in 2019 and 2020 (Karras et al., 2019, 2020). These models were designed to be able to be trained on higher-resolution and higher-dimensional input images and, as such, it became possible to not only generate images of human faces at scale, but these generated images had become increasingly realistic and indistinguishable from real images.

We want to argue that whilst deepfakes constitute a novel danger in society, they should also be seen as embedded in a much larger assemblage of events and powers that were in place long before deepfakes emerged. In this view, they constitute both a rupture in terms of their technological capacity, scope, and scale, but also a continuation in terms of the questions and issues they raise. Firstly, we want to explore this in relation to the question of the politics of (female) bodies.

Bodies

The first place deepfakes were observed was in the context of pornography. While not all deepfakes are pornographic, Emily van der Nagel (2020) cites a study which found that

'96% of deepfake images are non-consensual porn' (p. 424). This genre of deepfake video representation is where the face of one person, often a female celebrity, is placed onto the body of another. This is done to make it appear as though the viewer is watching the celebrity engaging in a sexual act. Part of the widespread notoriety of deepfakes stems precisely from the fact that they were first generated to create non-consensual sexual videos online, often depicting hardcore pornography, which mainly targets women (Gosse & Burkell, 2020). In fact, it has been reported that 100% those targeted and harmed in deepfake pornography are women, the main reason being that the algorithms have only been trained on images of women (Deeptrace, 2019). In a 2021 article in Vogue, Sophie Compton recounts the stories of women alerted to pornographic videos that appear to show them engage in often violent sexual acts. As Compton (2021) writes about one British author, 'Whenever she left the house, she felt exposed. On runs, she experienced panic attacks. Helen still has no idea who did this to her.' Here, deepfakes appear disturbing and dangerous because they represent a loss of agency over our bodies, making it appear as though we have done things or said things that did not actually happen. Unsurprisingly, this can cause major distress and humiliation – particularly to those women who are the subject of the deepfake pornographic videos. As Compton (2021) emphasises, 'these videos may be fake, but their emotional impacts are real.' Deepfakes, then, represent a new form of digital sexual abuse. They have become another means through which women lose control over their image, another means through which women are objectified and sexualised.

Other forms of online sexual abuse, such as revenge porn, require existing intimate or sexualised images, which are then shared online without a woman's permission. This is done with the aim of causing distress and humiliation to the victim. However, the emergence of deepfake technology means there is no longer a requirement for the perpetrator to possess 'real' intimate images of their victim. The creators of deepfakes only require sufficient images of their target's face and thus, as Gosse and Burkell (2020) argue, 'any woman can be made to appear in pornography' (p. 498). This is especially pernicious for, In patriarchal societies, women are encouraged to take personal responsibility for their own safety from physical and sexual violence. Moreover, it has also been shown that deepfakes are having detrimental impact on women in non-Western contexts as well. A 2019 report showed that non-Western female subjects featured in almost a third of deepfake pornography websites, 'with South Korean k-pop singers making up a quarter of the subjects targeted' (Deeptrace, 2019). The harms and gendered nature of deepfakes are a global phenomenon.

Yet how do women protect themselves from becoming the subject of deepfake pornography? Particularly when all that is required to produce this kind of content is a bank of facial images (Gosse & Burkell, 2020) and increasingly accessible algorithmic technology. For both the woman whose face is placed on another's body, and the female performer whose face is replaced, deepfakes represent a loss of control over womens' image. It is worth adding, however, that the worries about the visual representation of women produced through deepfakes has repeatedly emphasised a particular *kind* of female body. Media coverage of deepfake pornography has focused predominantly on cis-gendered women, (Gosse & Burkell, 2020) especially high-profile female celebrities such as Emma Watson, Taylor Swift, and Gal Gadot, all of which have been victims of non-consensual deepfake porn (Compton, 2021).

Despite it mostly being obvious that these videos are fake, Brown and Fleming (2020) argue that the emergence of deepfake imagery has also resulted in a kind of 'moral panic'. This hints at the trust placed in images to represent a reliable record of an event, which is arguably what makes deepfakes so disturbing. However, while the technology may be new, deepfakes represent a continuation in how images of women are used to control them and to limit their representation in society. In her groundbreaking 1975 paper 'Visual Pleasure and Narrative Cinema', Laura Mulvey provides a psychoanalytic critique of the ways in which cinema reinforces pre-existing patterns of (gendered) fascination within viewers. At stake here, she states, is that 'the film reflects, reveals and even plays on the straight, socially established interpretation of sexual difference which controls images, erotic ways of looking and spectacle' (Mulvey, 1989, p. 14). In this view, cinema perpetuates an unconscious patriarchal order where 'man can live out his fantasies and obsessions through linguistic command by imposing them on the silent image of woman still tied to her place as bearer, not maker of meaning' (p. 15). Both cinema and the wider patriarchal order, Mulvey claims, is therefore dependent upon a particular view of women: docile objects of looking, passive carriers of male fantasies and obsessions. The dominant structures of seeing in society, as demonstrated by the cinema screen, are always already formed by patriarchal desires. A male gaze: 'Unchallenged, mainstream film coded the erotic into the language of the dominant patriarchal order' (p. 16). Similarly, it can be argued that deepfakes, rather than simply representing a rupture in the sense of an accelerated loss of agency of female bodies, should be understood to form part of a long genealogy of male-dominated ways of seeing. With deepfakes, as with Mulvey's cinema, women figure as docile and passive carriers of male fantasies and obsessions.

That being said, one of the main differences is that this mode of seeing, this male gaze, is increasingly algorithmically generated and perpetuated. As Mulvey (1989) put it, 'the determining male gaze projects its fantasy onto the female figure, which is styled accordingly' (p. 19). With deepfakes, the male gaze becomes algorithmic. Yet, it is not so much that deepfakes introduce a new logic of sexist exploitation, but rather that they constitute new modes in which the fantasy of the female body can be *styled*, new arrangements and regimes of being looked at and displayed. With deepfakes, bodies and faces become increasingly interchangeable. Any face can be transposed unto anybody. The female face and body become raw materials for the endless recomposition and stylisation of the algorithmic male gaze. In other words, deepfakes can be understood as a form of algorithmic stylisation of the female (sexualised) body. As such, deepfakes arguably constitute an intensification of what Shaun Denson (2020) has called 'discorrelated images'. That means that they generate a further disjuncture between visual representations of women and their lived, embodied socio-cultural contexts. In other words, the female body is not simply subject to further abstraction, but rather there is a radical dissonance between the embodied realities of women and the ways in which they can be visually represented and algorithmically processed.

Therefore, we argue that the production and consumption of pornographic deepfake videos is less about how well the videos pass for being 'real' and more about reinforcing a patriarchal view of women, a view perpetuated and transformed by the possibilities and logics of algorithmic culture (Brown & Fleming, 2020). In this sense, there is nothing new nor surprising about deepfakes. In fact, it is arguably the male gaze, the dominant cultural

structure of seeing, that has prompted the penetration of deepfake technologies into the domain of pornography – and not the other way around. Yet, despite the fact that deepfakes do not depict real events, they can still make what they represent a reality. Brown and Fleming (2020) argue that pornographic 'deepfake makes patriarchal society real, giving it the very 'depth' that it needs in order to be made real' (p. 362). They argue that it does not matter that most videos are obviously fake, their production and consumption work to reinforce women's subjugated position in patriarchal society. Therefore, we argue that the 'real' image in deepfake pornography is a predominantly male fantasy through which the male gaze is validated. 'Algorithms,' Louise Amoore writes (2020) in the book *Cloud Ethics*, 'are giving accounts of themselves all the time. These accounts are partial, contingent, oblique, incomplete, and ungrounded' (p. 19). Similarly, deepfakes generated through GANs are also giving account of themselves; these accounts, however, are not only partial, incomplete, and ungrounded, but also point to, in this case, an underlying patriarchal way of looking at women.

Politics

The apparent use of deepfakes in political contexts is becoming seemingly notable and mundane. In March 2022, several media outlets reported that Meta and YouTube had detected and removed a deepfake video depicting Ukrainian president Volodymr Zelensky supposedly surrendering to Russia (Wakefield, 2022). Interestingly, the case elicited two predominant responses: on the one hand, many commentators noted the crudeness and poor quality of the deepfake video – a head that appeared too large, a voice that sounded too deep, different levels of pixel resolution within the video frame. The detection of the manipulated video was considered an 'easy win' for the social media platforms because it was easily debunkable as well as a 'childish provocation' by president Zelensky himself. On the other hand, it was seen by some commentators as emblematic of the continual 'eroding trust in authentic media' (Wakefield, 2022). The possibilities of using deepfakes in political contexts have instigated an emergent sense of ambivalence and anxiety. This is particularly salient in the context of elections. As Paul Scharre, Director of Technology and National Security at the US think tank Center for a New American Security (CNAS) stated, 'It is only a matter of time before deepfakes are used in an attempt to manipulate elections' (Scharre cited in Deeptrace, 2019). Indeed, just prior to the 2020 US presidential election, an election already marred by the ubiquitous and often strategic use of the term 'fake news', Diakopoulos and Johnson (2020) claimed that.

> Deepfakes contribute to the broader problem of 'fake news' by technically enabling the more widespread fabrication or manipulation of media that may be deliberately used for the purposes of disinformation and the introduction of uncertainty which can affect trust in news on social media … they have the potential to undermine the integrity of democratic elections. (p. 2)

Deepfakes, in other words, evoke political anxieties in that they are supposedly capable of disrupting democratic elections, damaging international relations, and undermining public trust in politicians as well as news outlets. In this view, deepfakes are 'a looming challenge for privacy, democracy, and national security' (Chesney & Citron, 2019). Some studies have claimed, however, that rather than directly misleading people deepfakes may

instead contribute towards a 'generalized indeterminacy and cynicism', which in turn intensifies the 'challenges to online civic culture in democratic societies' (Vaccari & Chadwick, 2020, p. 10).

Yet, there are also indications that there is a disjoint between public perception and the actual use of deepfakes in the political domain. For instance, it was noted that deepfake technology did not actually feature widely in the context of the 2020 US presidential election. As Cooper Raterink (2021) claims, the election instead saw 'verified political accounts and largely authentic grassroots behavior (such as well-intentioned and friend-of-a-friend misinformation) were most responsible for the spread of misleading narratives.' Whilst framed as harbingers of misinformation, eroding people's faith in political institutions, deepfakes have become an imaginary concentration point where various political, cultural, social, and technological anxieties are aggregated, reified, and amplified.

Crucially, we claim, the perceived capacity of deepfakes to rupture the political fabric is not so much a call to mitigate or anticipate possible harm to capitalist-liberal institutions, but rather an attempt to return to stable and fixed standards of political discourse. A nostalgia for stability and fixity. This is well illustrated by the claims put forward by law scholars Chesney and Citron (2019). They claim that 'democratic discourse is most functional when debates build from a foundation of shared facts and truths supported by empirical evidence' (p. 1777). What are the implications of deepfakes, then, for such foundations of shared truths? They respond:

> Deep fakes will allow individuals to live in their own subjective realities, where beliefs can be supported by manufactured 'facts'. When basic empirical insights provoke heated contestation, democratic discourse has difficulty proceeding. In a marketplace of ideas flooded with deep-fake videos and audio, truthful facts will have difficulty emerging from the scrum. (Chesney & Citron, 2019, p. 1778)

Deepfakes, in their view, propel a mass shift from objectivity to subjectivity and relativity, from truthful facts to continual contestation. From the fixed boundaries of truth to the proliferation of algorithmic fakes. This view is well embodied by the recent British thriller series *The Capture*. In the series, deepfakes are not only in focus but are ubiquitous, leading to a widespread rupture in trust. The manipulation and generation of fake voices, images, and video feeds – often happening in real time – repeatedly creates a sense for the viewer of a chaotic scrum as Chesney and Citron suggested. The world of deepfakes, in other words, is a world where nothing can ever be trusted, nothing is ever as it seems, where even the possibility of algorithmic manipulation has firmly eroded any chance of a foundation of shared facts and truths.

Yet, over-emphasising the potential power of deepfakes, as well as their impact on society, obfuscates the fact that politics is always already the realm of the difficult, problematic, and contested. This is what political theorist Chantal Mouffe (1993) argues in *The Return of the Political*. For Mouffe, liberal societies are incapable of grasping 'the irreducible character of antagonism' (p. 2). Instead of thinking of democracy in terms of rationalist, universalist, and individualist logics, Mouffe emphasises the ineluctable and constitutive role of antagonism, suggesting 'There will always be competing interpretations of the political principle of liberal democracy, and the meanings of liberty and equality will never cease to be contested' (p. 7). Therefore, the danger of over-

emphasising the ruptures of deepfakes can give rise to a nostalgia for what Mouffe calls 'the illusion of consensus and unanimity' (p. 5), an illusion which promises that 'a universal rational consensus could be produced by an undistorted dialogue, and that free public reason could guarantee the impartiality of the state' (p. 140).

Attending to the tensions of deepfakes, in other words, is to be attentive to the ways that these are entangled with the political, which has always already been a question of difficulty and contestation. Whilst deepfakes may constitute a novel mode of disrupting the political domain, it by no means erodes what Chesney and Citron (2019) called 'the foundation of shared facts and truths supported by empirical evidence'. On the contrary, the emergence of deepfakes should be a reminder that ideas of democracy and political discourse have never been settled, and that conflict and antagonism are precisely their 'condition of possibility and the condition of impossibility of its full realization' (Mouffe, 1993, p. 8). Rather than a lament for a bygone past of consensus and shared truths, the emergence of deepfakes should remind us that 'agonistic democratic practice is not a strategic design for control, but an admission of the contestability of one's own politics' (Heemsbergen et al., 2022, p. 4). The emphasis on the destabilising effect of deepfakes, in other words, obfuscates how politics has always been a question of contestability, intractability, and difficulty. Politics – as well as our ideas of objectivity – have therefore never been settled. It is precisely this question of objectivity we turn to in the following section.

Objectivity

As we have pointed out throughout this paper, deepfakes are said to accelerate a widespread distrust in images, videos, and texts. They arguably undermine the legitimacy of imagic and video representations by generating artificial and, by implication, misleading discourses and scenarios. The creation of fake news, exploitative porn, and fake faces via GANs. Anxieties surrounding deepfakes have therefore resulted in persistent calls for transparency and regulation as well as an 'arms race between deepfake technology and its detection techniques' (De Vries, 2020, p. 2110). Indeed, the political and technical response to deepfakes is predominantly instantiated through a framework of 'detection': the development of computational techniques and algorithmic models that detect whether an image or video is fake or not (van der Nagel, 2020). In February 2022, Google released an algorithmic model called 'Assembler', which combines different deepfake detection techniques. When encountering an image or video, the model will provide a probabilistic score indicating the likelihood that it has been algorithmically generated or manipulated (Hao, 2022). This echoes what Mike Ananny (2020) has called social media's framework of 'regulation-through-probability', whereby the algorithms makes decisions regarding the validity of material on the basis of ever-changing probabilities, thresholds, and confidence intervals.

In 2019, Mike Schroepfer, Chief Technology Officer at Meta, announced the start of the Deepfake Detection Challenge (DFDC). This challenge, built in collaboration with a variety of companies and universities such as Microsoft, MIT, and University of Oxford, aimed to 'produce technology that everyone can use to better detect when AI has been used to alter a video in order to mislead the viewer' (Schroepfer, 2019). Furthermore, Schroepfer states that 'the Deepfake Detection Challenge will include a dataset and

leaderboard, as well as grants and awards' as a way to 'spur the industry to create new ways of detecting and preventing media manipulated via AI from being used to mislead others.' Nick Clegg, current VP of Global Affairs and Communications at Meta, commented on the initiative, stating 'We must and we will get better at identifying lightly manipulated content before it goes viral and provide users with much more forceful information when they do see it' (Clegg, 2019). In these instances, the ethicopolitics of deepfakes is reducible to the development of sufficiently sophisticated tools, benchmarks, algorithms, and other countermeasures that are capable of capturing and preventing a wide variety of deepfakes. Or as Mike Ananny (2019) has put it, 'makers and detectors of 'deep fake' media play continual cat-and-mouse games to create and catch fabricated images, audio, and videos'. In short, it is a question of detection, verification, and legitimation.

Deepfakes do indeed constitute a rupture in terms of the unprecedented scope, scale, availability, and capacity to generate photorealistic and misleading outputs. The accessibility of open-source algorithms such as StyleGAN has made it possible to easily manipulate or generate fake images, texts, and videos. Yet, deepfakes, as well as the logic of detection which accompanies them, are problematic in a further sense: they reinforce a certain boundary between the real and fake. They participate in establishing and stabilising a certain conception of the 'real' as well as obfuscating the necessarily fraught nature underpinning all visual representations of reality. The idea of the truthful and fixed visual representation of reality is an unattainable dream, an oxymoron. Images and videos have never been 'objective'. They are always always predicated on and constituted through a particular technical apparatus, point of view, choice of subject matter, as well as different webs of power relations. In this view, the attempt to 'solve' the problem of deepfakes through a framework of detection constitutes a desire to stabilise the image, stabilise and fix its relation to reality. It is an attempt to establish a new form of what Daston and Galison (1992) call 'mechanical objectivity' whereby all traces of ambiguity and subjectivity are eradicated from the equation of verification. The assumption here is that the technical apparatus, as opposed to humans, can be distant, detached, and impartial in a given domain. The capture of deepfakes represents a dream to create the conditions for a realm of clear distinctions between truth and falsity as well as the capacity to accurately differentiate between what is real and what is artificial.

The relation between image and reality has never been fixed nor settled. In fact, the tension underpinning the relation between image and reality becomes even more apparent when seen from a historical perspective. From the 1810s to 1840s, for instance, Jonathan Crary (1990) argues that there emerged a 'new valuation of visual experience'. Within this particular framework, visual experience was given 'an unprecedented mobility and exchangeability, abstracted from any founding site or referent' (p. 14). Crary associates this new valuation of visual experience with the rise of modernity in the nineteenth century and the increasing 'uprooting of vision from the stable and fixed relations' introduced through novel visual techniques as well as emerging cultural and economic power relations (p. 14). In one sense, therefore, deep fakes can be understood as both a continuation and an intensification of such earlier processes of modernisation, insofar as they perpetuate a radical abstraction of the visual from any founding site or referent through algorithmic means. Here, deepfakes serve as a reminder of the necessary

contingency of vision as well as the slipperiness between the real and fake, which runs through all forms of representation.

On another level, however, deepfakes also point to the tensions inherent in attempts to erect clear and stable distinctions between the real and the fake. In her analysis of deepfakes, Katja De Vries (2020) acknowledges the usefulness of frameworks that divide fact from fiction. However, she argues that 'detection technology of deepfakes can be very useful, as long as it is not a distraction of the bigger picture: that facts can fall short of reality, and fabricated realities can be more representative of reality than so-called 'real' realities' (p. 2119). Instead of asking whether an image is real or a deepfake, the question for de Vries is 'whether it is fabricated well!' (p. 2119). Still, this evaluation of deepfakes must be grounded in a certain notion of objectivity. De Vries argues that the evaluation of deepfakes needs to be based on 'a set of professional standards' derived from 'statistical measures of reliability, representativeness, significance' (p. 2119). In our view, the use of statistical measures may prove a fruitful avenue in this context, but it nonetheless highlights the problematic attempt to establish a new form of objectivity by which the line between the real and fake can be clearly drawn. In this view, the relation between image and reality can be stabilised. This is an example of what Alain Badiou (2007) calls 'the passion for the real', that is, an attempt 'to grasp real identity, to unmask its copies, to discredit fakes' (p. 56). Yet, only attending to methods and standards of evaluation obfuscates the underlying politics that inheres in the practices of setting such standards as well as the particular actors involved. Does this mean, then, that no measures of detecting deepfakes should be put in place, that no measures are possible? Not quite. As Jacques Derrida (2002) aptly put it, 'we have to impose rules, but we also have to distrust them' (p. 50). Similarly, we have to detect deepfakes, but we also have to think critically about the practice of 'detecting' and how we define the parameters of what is being detected. As such, deepfakes embody an irreducible tension – on the one hand, between their potential to endanger and mislead and, on the other, their potential to reinforce problematic dichotomies between the real and artificial.

When companies such as Facebook and Google develop algorithmic detection techniques they not only participate in 'solving' the problem of deepfakes; they also reinforce their role as powerful 'arbiters of recognisability' (Amoore, 2020; Jacobsen, 2021), where they increasingly determine the boundaries between the real and the fake, what can be recognised and what is ignored. In other words, their efforts to detect and verify are entangled with the reinforcement of a certain notion of objectivity, one which benefits the tech companies themselves. Claims to 'solve' a problem are often made by those who also demarcated the boundaries of the problem space in question. Seen from this view, detection techniques should not be conceptualised simply as the answer to the problem of deepfakes. They should always be understood in relation to the companies developing or deploying them. They should be seen to reaffirm a dual dream: the capacity to distinguish between the real and the artificial as well as the possibility of producing objective representations of the world. In fact, the development and rolling out of various detection countermeasures raises the crucial question, who or what gets to decide what counts as a deepfake? Where is the threshold between real and fake being drawn? The 'problem' of deepfakes is therefore not easily solvable. While they have the potential to undermine the assumed legitimacy of information provided online, they also foreground deeper issues, such as the power to fix the boundaries and

thresholds between real and fake. These issues cannot be solved through deepfake detection techniques and algorithms alone. Instead, they demand that we examine more closely the role of social media platforms and tech companies in reinforcing a particular demarcation between the real and the fake underpinning this problem space. This power to demarcate and fix is one which also reinforces their position as powerful arbiters of recognisability in contemporary society.

Conclusion

As we have shown in this paper, deepfakes occupy a complex and contested space in society. They have the potential to introduce novel modes of disruption (in terms of their impact on bodies, politics, and ideas of objectivity). Yet, attending solely to their novelness risks obfuscating already-existing, socially ingrained norms, ideas, and struggles in the digital landscape. We therefore conceptualise deepfakes as assemblages of differential tensions, comprised of interlinking and overlapping ruptures and continuities. Drawing on David Beer's (2023) 'tensions' around algorithmic thinking – 'the competing forces at play in the advancement and application of automation' (p. 4) – we have highlighted in this paper how deepfakes always already reflect and embody multiple competing forces in society: ruptures, perpetuations, continuities and discontinuities. We further argued that deepfakes cannot be reduced to questions of recognition and detectability, that is, developing standards and algorithmic tools for the detection of deepfake content. For, as John Berger (1972) reminds us, 'the relation between what we see and what we know is never settled' (p. 7). The notion of the deepfake is never settled. Attending to the differential tensions of deepfakes is therefore to be dually attentive: firstly, to the ways in which deepfake technologies can and are being used to disrupt and harm, but secondly, to be critical of the performative effects of demarcating some digital content as being 'deepfake'. This act of labelling, which involves a complex network of human and nonhuman actors, transforms the ways in which information and content are perceived, analysed, and managed.

More specifically, we have explored the idea of the tensions of deepfakes through three separate case studies: bodies, politics, and objectivity. In the first section, bodies, we showcase how deepfakes have the capacity to intensify the exploitation, abuse, and objectification of women in digital spaces through the emergence of non-consensual deepfake pornography. Yet, through a reading of Laura Mulvey's (1989) notion of 'male gaze', we have also emphasised how objectification is not something new. Rather, deepfakes therefore occupy a contested space between rupture and continuity, between female objectification as always emerging and always already having emerged. In the second section, we showcased how the potential impact of deepfakes has been conceived for the political domain. Yet, drawing on Chantal Mouffe's (1993) work, we stressed that not only is history filled with altered images, but how anxieties about deepfakes can occlude the ways in which politics is always already fundamentally antagonistic and difficult. Lastly, we problematise the idea that deepfakes can and should be conceived as a problem that is solvable through an array of algorithmic detection techniques. We argue that the attempt to 'solve' the problem of deepfakes through a framework of detection constitutes a problematic attempt to erect clear boundaries between the real and the fake as well as to

establish a new form of what Daston and Galison (1992) call 'mechanical objectivity' whereby all traces of ambiguity and subjectivity are eradicated from the equation of verification.

Yet, there is still a need to ask, what are the wider implications of understanding deepfakes as an assemblage of differential tensions, composed of intermingling ruptures and continuities? We argue it is an antidote to the possible and probable fetishisation of deepfakes in society. In other words, they should not be fetishised as the sole driving force of an impending 'infocalypse' (see for instance Schick, 2020). Critical data scholars and theorists have long warned against the fetishisation of technical objects such as source code or algorithms, arguing that it paints a purely technical picture that, in turn, obfuscates the fact that it is always already permeated with socio-cultural and political practices and assumptions (Chun, 2008; Crawford, 2016; Dourish, 2016). In other words, to be attentive to the tensions of deepfakes is to be wary of accounts, scholarly or otherwise, that imbue deepfakes with a special kind of disruptive power without also attending to the societal issues, anxieties, and other modes of disruption of which they necessarily are a continuation. In short, to look beyond deepfakes as fetishised objects. In addition to their sociopolitical impact, we need to critically consider the means and mechanisms by which deepfakes are named, detected, arranged, presented, and managed. More importantly, we need to remain highly critical of the politics of contemporary visuality, how the line between real and fake is never settled nor can ever be settled – even in an age of deepfakes. This means that we cannot reduce the tensions of deepfakes to a solvable problem via detection frameworks. Borrowing from Donna Haraway (2016), attending to the tensions of deepfakes is to 'stay with the trouble' of deepfakes.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Dr Benjamin N. Jacobsen is a Postdoctoral Research Associate on Professor Louise Amoore's 'Algorithmic Societies' project at Durham University. His research currently explores the sociopolitical implications of generative modelling and synthetic data on society and culture.

Dr Jill Simpson is an associate lecturer in Sociology at University of York. Her research interests combine critical data studies, interdisciplinary social research and public engagement through creative practice. She has published work on the power and politics of data visualisations.

References

Amoore, L. (2020). *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Duke University Press.

Ananny, M. (2019). Probably speech, maybe free: Toward a Probabilistic understanding of online expression and platform governance. *Knight First Amendment Institute.* https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance

Ananny, M. (2020). Making political people: How social media create the ideals, definitions, and probabilities of political speech. *Georgetown Law Technology Review*, *4*, 351–366. https://georgetownlawtechreview.org/wp-content/uploads/2020/07/4.2-p351-366-Ananny.pdf

Badiou, A. (2007). *The century*. Polity Press.
Beer, D. (2017). The social power of algorithms. *Information, Communication & Society*, *20*(1), 1–13. https://doi.org/10.1080/1369118X.2016.1216147
Beer, D. (2023). *The tensions of algorithmic thinking: Automation, intelligence and the politics of knowing*. Bristol University Press.
Bennett, J. (2010). *Vibrant matter: A political ecology of things*. Duke University Press.
Berger, J. (1972). *Ways of seeing*. Penguin Books.
Brown, W., & Fleming, D. H. (2020). Celebrity headjobs: Or oozing squid sex with a framed-up leaky {Schar-JØ}. *Porn Studies*, *7*(4), 357–366. http://dx.doi.org/10.1080/23268743.2020.1815570.
Bucher, T. (2018). *If… then: Algorithmic power and politics*. Oxford University Press.
Chesney, R., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*, 1755–1820. https://doi.org/10.15779/Z38RV0D15J
Chun, W. (2008). On 'sourcery,' or code as fetish. *Configurations*, *16*(3), 299–324. https://doi.org/10.1353/con.0.0064
Chun, W. H. K. (2021). *Discriminating data: Correlation, neighborhoods, and the new politics of recognition*. MIT Press.
Clegg, N. (2019). Facebook, elections and political speech. *Meta Newsroom*. https://about.fb.com/news/2019/09/elections-and-political-speech/
Compton, S. (2021). More women are facing the reality of deepfakes, and they're ruining lives. *Vogue*. https://www.vogue.co.uk/news/article/stop-deepfakes-campaign
Crary, J. (1990). *Techniques of the observer: On vision and modernity in the nineteenth century*. MIT Press.
Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology & Human Values*, *41*(1), 77–92. https://doi.org/10.1177/0162243915589635
Daston, L., & Galison, P. (1992). The image of objectivity. *Representations*, *40*, 81–128. https://doi.org/10.2307/2928741
Deeptrace. (2019). The state of deepfakes: Landscape, threats, and impact. *Deeptrace Labs*. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf
Deng, J., Socher, R., Li, L. J., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848
Denson, S. (2020). *Discorrelated images*. Duke University Press.
Derrida, J., & Stiegler, B. (2002). *Echographies of television: Filmed interviews*. Polity Press.
De Vries, K. (2020). You never fake alone: Creative AI in action. *Information, Communication & Society*, *23*(14), 2110–2127. https://doi.org/10.1080/1369118X.2020.1754877
Diakopoulos, N., & Johnson, D. (2020). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*(7), 1–27. https://doi.org/10.1177/1461444820925811.
Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3(2)*, 1–11. https://doi.org/10.1177/2053951716665128
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 1–9.
Gosse, C., & Burkell, J. (2020). Politics and porn: How news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, *37*(5), 497–511. https://doi.org/10.1080/15295036.2020.1832697
Hallinan, B., & Striphas, T. (2016). Recommended for you: The netflix prize and the production of algorithmic culture. *New Media & Society*, *18*(1), 117–137. https://doi.org/10.1177/1461444814538646

Hao, K. (2022). Google has released a tool to spot faked and doctored images. *MIT Technology Review*. https://www.technologyreview.com/2020/02/05/349126/google-ai-deepfakes-manipulated-images-jigsaw-assembler/
Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599. http://dx.doi.org/10.2307/3178066.
Haraway, D. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.
Heemsbergen, L., Trere, E., & Pereira, G. (2022). Introduction to algorithmic antagonisms: Resistance, reconfiguration, and renaissance for computational life. *Media International Australia*, *183*(1), 1–13. https://doi.org/10.1177/1329878X221086042
Jacobsen, B. N. (2021). Regimes of recognition on algorithmic media. *New Media & Society*, 1–16. Online first. https://journals.sagepub.com/doi/pdf/10.11771461444821105355 5
Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *ArXiv*, 1-12.
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. *ArXiv*, 1-21.
Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087
Meikle, G. (2023). *Deepfakes.* Polity Press.
Mitchell, W. J. T. (2005). *What do pictures want? The lives and loves of images*. The University of Chicago Press.
Mouffe, C. (1993). *The return of the political*. Verso.
Mulvey, L. (1989). *Visual and other pleasures*. Palgrave.
Offert, F. (2021). Latent deep space: Generative adversarial networks in the sciences. *Media + Environment*, *3*(2), 1–20. https://doi.org/10.1525/001c.29905
Raterink, C. (2021). Assessing the risks of language model 'deepfakes' to democracy. *Tech Policy Review*. https://techpolicy.press/assessing-the-risks-of-language-model-deepfakes-to-democracy/
Schick, N. (2020). *Deepfakes: The coming infocapolypse*. Grand Central Publishing.
Schroepfer, M. (2019). Creating a dataset and a challenge for deepfakes. *Meta AI*. https://ai.facebook.com/blog/deepfake-detection-challenge?utm_source = hp
Striphas, T. (2015). Algorithmic culture. *European Journal of Cultural Studies*, *18*(4-5), 395–412. https://doi.org/10.1177/1367549415577392
Uricchio, W. (2011). The algorithmic turn: Photosynth, augmented reality and the changing implications of the image. *Visual Studies*, *26*(1), 25–35. https://doi.org/10.1080/1472586X.2011.548486
Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, (1), 1–13. https://doi.org/10.1177/2056305120903408
van der Nagel, E. (2020). Verifying images: Deepfakes, control, and consent. *Porn Studies*, (4), 424–429. https://doi.org/10.1080/23268743.2020.1741434.
Wakefield, J. (2022). Deepfake presidents used in Russia-Ukraine war. *BBC*. https://www.bbc.co.uk/news/technology-60780142
Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137–150. https://doi.org/10.1080/1369118X.2016.1200645
Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence: The International Journal of Research Into New Media Technologies*, *27*(1), 36–51. https://doi.org/10.1177/1354856520923963
Zeilinger, M. (2021). Generative adversarial copy machines. *Culture Machine*, *20*, 1–23. https://culturemachine.net/wp-content/uploads/2021/09/Martin-Zeilinger.pdf
Zylinska, J. (2020). *Ai art: Machine visions and warped dreams.* Open Humanities Press.