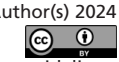
EUROPEAN JOURNAL OF Cultural Studies

Article

European Journal of Cultural Studies
1–17



DOI: 10.1177/13675494241240028
journals.sagepub.com/home/ecs


Deepfakes and the promise of algorithmic detectability

Benjamin N Jacobsen
University of York, UK

Abstract

Deepfakes, as a sociocultural and technical phenomenon, have engendered two distinct yet intimately interwoven set of responses: on one hand, they have created widespread anxieties concerning the potential and harmful impact of deepfakes. On the other hand, they have also given rise to a new regime of detection: tools, models, and methods that are developed and used to detect whether something is a deepfake or not. However, the ways in which machine learning algorithms are being framed as the solution to the problem of deepfake detection have not received sufficient critical attention. Drawing on the 2019 Deepfake Detection Challenge organised by Meta as well as finding resonances in the work of Eyal Weizman, this article seeks to problematise and unsettle what I call *the promise of algorithmic detectability*. That is, the claim that machine learning algorithms render the issue of deepfake detection knowable, tractable, and resolvable. Examining the themes of training data, thresholds, and certainty, I emphasise the inherent difficulties, intractabilities and contingencies of deepfake detection models. Ultimately, I argue that the promise of algorithmic detectability falls short and that the ethico-politics of deepfakes cannot be reduced solely to a framework of detection algorithms.

Keywords

Algorithms, data, deepfakes, detection, machine learning, meta

Introduction

In March 2023, an image of Pope Francis wearing a stylish-white puffer jacket went viral. The image was fake. It had been generated through prompts inputted to the popular image diffusion model called Midjourney. The image fuelled already-existing anxieties

Corresponding author:
Benjamin N. Jacobsen, Department of Sociology, University of York, York, Yo10 5GD, UK.
Email: benjamin.jacobsen@york.ac.uk

about the dangers and harms of so-called deepfakes (Stokel-Walker, 2023) – content such as images or videos that have been algorithmically altered or manipulated. Deepfakes, as a sociocultural and technical phenomenon, have become an intimate and mundane part of contemporary algorithmic culture (Hallinan and Striphas, 2016; Striphas, 2015).[1] They have engendered two distinct yet intimately interwoven set of responses: they have created widespread anxieties concerning the impact of deepfakes for political discourse, journalistic integrity, and trustworthiness (Chesney and Citron, 2018; Johnson and Diakopoulos, 2021; Vaccari and Chadwick, 2020), as well as how they are used to personalise and exacerbate gendered harms and exploitation (Compton, 2021; van der Nagel, 2020).[2] Inevitably, some of the popular discourses surrounding deepfakes have approximated moral panics, with some declaring an impending 'infocalypse' (Schick, 2020) and others claiming that 'manipulated video will ultimately destroy faith in our strongest remaining tether to the idea of common reality' (Foer, 2018).[3]

However, deepfakes have also given rise to a new regime of detection: tools, models, and methodologies that are used to detect whether something is a deepfake or not. This has created an ongoing situation that some have characterised as an 'arms race between deepfake technology and its detection techniques' (De Vries, 2020: 2110) or as a set of 'continual cat-and-mouse games to create and catch fabricated images, audio, and videos' (Ananny, 2019). Some of these methods emphasise the centrality of the human as the eye of detection. For instance, companies such as Reuters, in collaboration with Meta, have introduced an online training course to teach their journalists to detect deepfakes on social media (Ronik, 2022). Anxieties surrounding artificial intelligence (AI)-generated content, however, go beyond the domain of deepfakes. They have become endemic to contemporary society more broadly. There have been worries, for example, that students will increasingly rely on models such as ChatGPT to generate essays that would be difficult for teachers and markers to detect. In response, companies such as Winston AI and Turnitin have developed AI detectors aimed at uncovering written assignments that have been generated by ChatGPT (Marshall, 2023). Yet, these detection models are highly problematic, for as research scientists from Stanford University have shown that Generative Pre-trained Transformer (GPT) detectors consistently misclassify samples written by non-native English speakers as 'nonhuman' or 'AI generated' (Liang et al., 2023: 1). The fears over deepfakes therefore are, in many ways, symptomatic of a much broader set of societal anxieties about the implications of AI-generated texts and images.

It also remains the case that, as one computer science paper observed, 'most of the current DeepFake detection methods use data-driven neural networks as backbone' (Li et al., 2020: 2).[4] It is claimed that while many deepfakes are obvious to the naked human eye, 'deepfakes often generate artifacts which may be subtle to humans, but can be easily detected using machine learning' (Mirsky and Lee, 2020: 26). In other words, machine learning algorithms are framed as the solution to the problem of deepfakes, especially if and when this problem grows in scale.[5] In contrast to humans, algorithms are imagined to operate on a different, more granular level of perception, which is supposedly more conducive to deepfake detection. As such, they promise to solve the issue of algorithmically generated content and consequently reduce the ethico-politics of deepfakes to a framework of detection and the development of increasingly

sophisticated computational tools, data benchmarks, and model architectures. It becomes a question of detection, verification, and legitimation. Put differently, just as dangerous as deepfakes and other AI-generated content appear, just as seductive appears the promise of algorithmic detectability.

This article explores how machine learning algorithms are being imagined, developed, and implemented as a way to detect deepfakes. While there is already a substantial critical scholarship that has examined the sociopolitical impact of deepfakes, there is little research that critically explores the politics of deepfake detection and what can be called *the promise of algorithmic detectability*.[6] Following the work of Eyal Weizman (2017), I evoke the notion of 'detectability' as a way to problematise ideas about what it means to detect deepfakes. For Weizman, detectability is not simply how and when things are detected. Rather, it refers to the interlacing social, political, legal, and technical operations that render things detectable as such. For instance, he shows that while the pixel resolution of satellite images improved gradually from the 1970s and onwards, this development was halted in the early 2000s. The reason being that the pixel size was now small enough for human rights violations to begin to be recognisable within satellite images, causing geopolitical implications for countries such as the United States that continue to conduct drone warfare in the Middle East. It is therefore apparent, Weizman (2017) argue, that 'the pixel resolution of contemporary, publically satellite images is not only a product of optics, data storage, or bandwidth capacity, but of legal regulations that bear upon political and even geopolitical rationales' (pp. 27–28). Making something detectable, therefore, is a way of organising the world, making it matter in specific ways and not others. Taking inspiration from Weizman's analysis, this article does not consider deepfake detection as a question of either/or (either something is accurately detected or not). Instead, it considers the politics of how something falls within the 'thresholds of detectability' (Weizman, 2017). That is, how machine learning algorithms promise to make deepfakes detectable and thus manageable.

The article pays specific attention to algorithmic models that have been developed as part of this new detection regime. It also explores the tensions that surround them, focusing on how they are created, what assumptions underpin them, and how they actively impinge on the content they are supposed to merely detect. As such, the analysis draws on the case of the 2019 Deepfake Detection Challenge (DFDC) organised by Meta. Through the DFDC, Meta actively sought to make the algorithmic detection of deepfakes possible, desirable, and necessary. After outlining the history and main points of interest of the DFDC, I will use this case as a way to critically explore some of the broader issues related to deepfakes and the promise of algorithmic detectability. In what follows, I will focus on the following three core areas: training data, thresholds, and uncertainty. These three areas will help to foreground both the sociotechnicality of all efforts at deepfakes detection, but also how the promise of algorithmic detectability obscures its own inherent difficulties, contingencies, and intractabilities. In this way, I seek to problematise the promise and grounds of algorithmic detectability, ultimately arguing that the ethico-politics of deepfakes cannot be made resolvable simply through a framework of algorithmic detection. In the conclusion, I will also reflect on the extent to which deepfakes indicate a wider societal obsession with detection as such, claiming that deepfakes participate not

only in the potential erosion of objectivity and truth but also in accelerating the emergence of a so-called 'detection society'.

Benchmarking the detection of deepfakes

The detection of deepfakes did not emerge as the natural and inevitable response to the potential threats of deepfakes. Rather, it was made to happen in a particular social context. In September 2019, Meta organised the so-called Deepfake Detection Challenge (DFDC) in collaboration with different tech companies like Microsoft as well as universities such as MIT and Oxford University.[7] The DFDC was pitched as both a large-scale collaborative forum as well as a competitive space (the prize was US$500,000) where researchers and engineers could benchmark their deepfake detection models, comparing results and approaches with that of other competitors (Ferrer et al., 2020). As such, Meta attempted to create a deepfake detection benchmark equivalent to ImageNet, a training dataset which has remained a dominant computer vision benchmark for over a decade and has led to ground-breaking innovations in deep learning, most notably the AlexNet algorithm in 2012 (Krizhevsky et al., 2012). According to Mike Schroepfer (2019), former Chief Technology Officer at Meta, 'the industry doesn't have a great dataset or benchmark for detecting them', and therefore 'we want to catalyze more research and development in this area and ensure that there are better open source tools to detect deepfakes'.

Yet, the competition was also framed as a normative project. Sir Nick Clegg, former UK politician and current President of Global Affairs at Meta, stated that '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 Clegg's view, the dangers of deepfakes are not strictly localised in AI-generated images, but rather are symptomatic of a wider set of societal problems regarding disinformation and a crisis of legitimacy. For Schroepfer (2019), the primary aim of the challenge was to democratise the production and distribution of deepfake technology, to produce algorithmic tools that 'everyone can use to better detect when AI has been used to alter a video in order to mislead the viewer'. From the outset, therefore, Meta's deepfake detection challenge embodied a normative commitment to transparency, openness, and the public good. It embodied the promise that the algorithmic detection of deepfakes is the solution to the problem of deepfakes.

The DFDC competition launched 5th September 2019 and ended in June 2020. At the start, participants got access to a bespoke dataset comprised of 115,000 deepfake videos. This dataset had been created specifically for the challenge, and all algorithmic models were to be trained on this benchmark. A leadership board hosted by the online platform Kaggle was also set up where people could see the rankings of the different detection models. The participants' models were then assessed based on how they performed on this dataset as well as a previously unseen test dataset, which DFDC representatives called 'a black box dataset' (Meta, 2020). The reason for this, they stated, was that 'by using a distinct dataset, we were able to replicate real-world challenges, where models must be accurate even when tasked with new or unfamiliar techniques for creating deepfakes' (Ferrer et al., 2020).

By the end of the challenge, 2114 participants had submitted more than 35,000 algorithmic models to the competition. 'The most successful models', it was noted after the competition, 'all found ways to innovate in the task of deepfake detection' (Ferrer et al., 2020). Some of these innovations included novel uses of augmentation methods (such as blending real and fake faces, or dropping portions of faces randomly in videos), interesting uses of model architectures (such as different arrangements of neural network layers) as well as the absence of forensics methods (i.e. techniques that operate at a pixel level of the image, like sensor noise fingerprints). Ultimately, the winner of the DFDC was Selim Seferbekov, a Senior Machine Learning Engineer at Mapbox, whose model achieved a 82.56 percent average precision rate. After the competition, computer scientists and AI researchers at Meta alongside their academic partners stress-tested the winning models in order to better understand any specific vulnerabilities in the models before they were open-sourced (Ferrer et al., 2020). The results of the stress testing as well as an overview of the competition were then presented at a Workshop on Media Forensics at the 2020 Conference on Computer Vision and Pattern Recognition (CVPR) in Seattle, the United States.[8]

However, Meta representatives were also quick to acknowledge that despite these findings and innovations 'the DFDC results also show that this is still very much an unsolved problem' (Ferrer et al., 2020). This was evidenced by the fact that while the top-ranking models had performed well on the public dataset of deepfake videos, none of them achieved over 70 percent accuracy on the previously unseen 'black box dataset'. As outlined in a statement from Meta reflecting on the results of the competition, 'the highest-performing entrant was a model entered by Selim Seferbekov. It achieved an average precision of 65.18 percent against the black box dataset. Using the public dataset, this model had been ranked fourth' (Meta, 2020). Indeed, one of the conclusions of the competition was that 'this outcome reinforces the importance of learning to generalize to unforeseen examples when addressing the challenges of deepfake detection' (Meta, 2020). While the aim of the DFDC had been to 'spur the industry to create news ways of detecting and preventing media manipulated via AI' (Schroepfer, 2019), the results of the competition foregrounded some of the inherent difficulties, intractabilities, frictions, and contingencies of algorithmic detectability.

Rather than make the promise of algorithmic detectability appear more seductive, the competition highlighted what Mike Savage (2013: 5) has called 'the social life of methods', that is, it facilitates critical engagement with research methods 'by resisting the instrumental framing in which they are simply seen to be technically "better or worse"' than other methods. Similar to the ImageNet competition (Russakovsky et al., 2015), what the DFDC made clear is that there is nothing obvious nor self-evident about the algorithmic detection of deepfakes. Rather than seeing the production of detection methods in purely instrumental terms, the conditions of their possibility have to be actively (and messily) made. And these efforts to create a benchmark for algorithmic deepfake detection showcase an array of tensions and frictions underpinning all attempts at detection. These tensions and frictions stem from issues related to the training data on which a detection model is trained.

Training data

For any machine learning algorithm to learn how to recognise people, objects or patterns more generally in data, it first needs to be exposed to a vast amount of training data. For instance, a computer vision model learning to recognise breeds of dogs must have been trained on a dataset of images or videos containing diverse instances of dogs (Kelleher, 2019). The distribution of diverse examples in the training data is how things in the world are made recognisable and intelligible to the algorithm (Amoore, 2020; Jacobsen, 2021). Similarly, for an algorithmic deepfake detection model to work it must have been exposed to and trained on a large dataset of deepfakes. Optimally, it has to be trained on a hybrid dataset that contains both real and deepfake images or videos in order for the model to be able to differentiate between these. The detection algorithm, in other words, learns to detect what a deepfake 'is' – its features, patterns, defects, differences – in order to be capable of distinguishing between deepfakes and real content. However, this presents a number of challenges, one of which is that the relationship between these contrasting sets of images and videos is not static. 'As the counterfeits become more and more realistic', some computer scientists have noted, 'the differences between real and fake ones will become more and more subtle', and this poses problems for the detection models, as they are forced to look for increasingly fine-grained features and defects (Lu et al., 2021: 2). Others have pointed out that 'a closer look at the DeepFake videos in existing datasets reveals stark contrasts in visual quality to the actual DeepFake videos circulated on the Internet' (Li et al., 2020: 1). Algorithmic detection models therefore rely on a vast volume of good-quality training data, from which they can learn what deepfakes look like as well as the subtle differences between deepfakes and real videos.

As it currently stands, one of the main limitations concerning deepfake detection is the sheer lack of good-quality and diverse training data that is available. The development of such a training dataset was therefore one of the driving forces behind the challenge. As researchers from Meta put it,

> Motivated primarily by the fact that many previously-released datasets contained few videos with few subjects and with a limited size and number of methods represented, we wanted to release a dataset with a large number of clips, of varying quality, and with a good representation of current state-of-the-art face swap methods. (Dolhansky et al., 2020: 2)

As such, the aim of Meta's Deepfake Detection Challenge was not simply to encourage the production of technologies that can be used to detect deepfake, as Mike Schroepfer (2019) put it. Rather, it was to create the conditions of possibility for algorithmic detectability in the future. To fuel its promise – that is, to make algorithmic detectability of deepfakes appear possible, tractable, and desirable. What this suggests is that deepfakes are not simply content that machine learning algorithms detect at varying levels of accuracy, but they must be *made* detectable through the procurement and organisation of a carefully curated training dataset.

In a 2020 computer science paper, machine learning engineer Brian Dolhansky and colleagues provide further insights into Meta's DFDC dataset. Echoing a claim made by Lisa Gitelman and Virginia Jackson (2013), the article highlights how training datasets

for machine learning are not simply out there and available for extraction, but must always be 'cooked' in some way, must be produced, must be made to matter. The DFDC dataset contained over 100,000 video clips of people's faces that had been algorithmically swapped and altered. As opposed to previous datasets, there is a continual emphasis in the article on the ways in which the dataset ensures 'diversity in gender, skin tone, ethnicity, age, and other characteristics' (Ferrer et al., 2020) in order to achieve extensive 'visual variability' (Dolhansky et al., 2019: 19; see also Jacobsen, 2023). This variability promises to make generalisation possible, making it feasible for algorithms to detect a wide range of deepfakes in different domains in the wild.

Yet, the question still remains: *how to procure a benchmark training dataset when there is insufficient data examples available?* In the context of DFDC, the answer was not data extraction – as was the case for ImageNet and most other benchmark datasets – but rather data commissioning. That is, the dataset was commissioned by Meta and comprised 'paid actors, with the required consent obtained' (Schroepfer, 2019). Over 3500 individual actors took part, producing the equivalent of 38.5 days of video footage. Another rationale for this approach was that previous datasets such as Celeb-DF, which upon release was magnitudes larger than previous ones, had issues with use restrictions: 'the subjects in these videos may not agree to have their faces manipulated' (Dolhansky et al., 2020: 4). They state that this is precisely why 'we did not construct our dataset from publicly-available videos', but instead 'we commissioned a set of videos to be taken of individuals who agreed to be filmed, to appear in a machine learning dataset, and to have their face images manipulated by machine learning models' (Dolhansky et al., 2020: 4). In other words, the commissioning of data from consenting, paid actors were seen as a fairer and more legally tractable way of creating a deepfake detection dataset as opposed to extracting data from publicly available videos.

While the DFDC dataset may not suffer the same legal constraints as Celeb-DF, the implications of this mode of producing a detection dataset nonetheless go beyond notions of consent and legality. Namely, they problematise any stark dichotomy between real and fake. Elizabeth Strickland (2019) put it best: 'the participants had to agree to become deepfake characters'. Already here, then, at the outset of the challenge, there is a certain politics of fakery at play. For algorithms to be able to detect fakes, the fakes had to be made, had to be acted out and performed. Indeed, this highlights the inherent *performativity* of algorithmic detection, both in terms of how the detectable is actively made but also how it relies on conditions of theatricality and artifice, of actors acting out what an example of a deepfake could look like. This shows that the promise of algorithmic detectability to maintain a clear distinction between real and fake is fundamentally unsettled, precisely because the real and the fake are already intimately entangled at the level of the training data. For the fake to be detected, the real has to be transformed into a performance, a fake made amenable to the training of machine learning algorithms.

Thresholds

The promise of algorithmic detectability is to render deepfakes algorithmically detectable and manageable. Yet, as Eyal Weizman (2017) has shown, any given framework of detection always reduces to a binary decision: either it is detected or not, either the

detection is accurate or not. But Weizman argues that things are not simply out there and ready to be detected – by a human, a drone, a facial recognition model – but instead, they must be *made detectable* through the setting of thresholds. Whether something ultimately becomes detectable is intimately interwoven with the politics of how these thresholds are produced and who produces them. Indeed, at 'the threshold of detectability', Weizman (2017) suggests, one finds 'things that hover between being identifiable and not' (p. 20). The case of deepfake detection, therefore, is not simply to detect but to establish and manage this threshold, making things detectable as such.

Consider again the outcome of the 2018 deepfake challenge. When reflecting on the outcomes of the DFDC, researchers from Meta expressed ambivalence concerning the varying accuracy rates of the top-performing detection models. Algorithmic detection tools provide a probability or likelihood score that an image or video is a deepfake. While the winning model achieved an average precision score of 82.56 percent on the known training dataset, it only achieved 65.18 percent on the unknown 'black box' dataset (Meta, 2020). In other words, when encountering examples of deepfakes, it had not previously been trained on, the accuracy of the winning model was only somewhat better than a random coin toss. The question this raises is one of thresholds. For as Amoore (2019) has argued, to train and fine-tune an algorithmic model is to 'tweak it to the level of detection that is useful to you' (p. 6). What is considered a good enough score for an algorithmic model to successfully detect a deepfake in the social world? Or, in other words, *what is the useful decisional threshold for a machine learning model detecting likely deepfakes?* This question is never answered by the organisers of the DFDC. Another difficulty is that the threshold of these models continuously shifts, depending on new data input the models are exposed to. In other words, there is no 'natural' or 'self-evident' way to determine the thresholds of algorithmic detectability. These are constantly made and re-made in and through the machine learning models used to detect deepfakes. It follows that deepfakes are not natural entities that exist out there in the real world. Rather, they hover on the 'threshold of detectability' (Weizman, 2017), as they are continuously *made* and *come to matter* as deepfakes through various detection models and tools.

The Meta researchers seem to indicate that an accuracy rate larger than 80 percent would approximate that of many computer vision models. Still, they do not actually answer the question what is 'good enough' for deepfake detection. As Louise Amoore (2020) has also argued, 'it is not the accuracy of the algorithm that matters so much as the sufficient proximity to a target' (p. 67). Similarly, the reason for Meta's strategic silence is that there is no common or ideal accuracy measure for deepfake detection models. There is no overarching, pre-existing benchmark which can authorise a detection threshold. When is something (confidently) considered a deepfake – when the model achieves 75 percent accuracy? 80 percent? 90 percent? While a higher accuracy is better than a lower, there is a certain level of arbitrariness underpinning these thresholds, which are negotiated by those building and implementing these 'good enough' detection models in specific sociopolitical contexts.

While this arbitrariness needs to be taken seriously, it is equally important to consider what these detection threshold produce in the world. That is, what responses, interventions, operations, and foreclosures do they authorise? As Ramon Amaro (2020) states, 'models provide a means by which data can be analyzed to achieve a quantitative

threshold value, which is often translated into a series of targeted actions'. Similarly, the threshold in our case authorises a particular drawing of the line between real and fake. It gives weight to the final decision, which renders something detectable as a 'deepfake', reified and organised as fake content. This foregrounds the importance of the context in which the threshold is made. Mike Ananny (2019) has shown how content moderation on social media platforms is often based on both spoken and unspoken 'thresholds of acceptability', that is, sociotechnical negotiations of where to draw the line in certain cases of hate speech and misinformation. This finds resonance in a 2020 case, when Meta announced a new policy banning deepfakes from being shared on their platforms. In their 'Misinformation' policy, it read that Meta will remove videos if 'the video is the product of artificial intelligence or machine learning, including deep learning techniques (e.g. a technical deepfake), that merges, combines, replaces and/or superimposes content onto a video, creating a video that appears authentic' (Meta, 2022). They also made it clear that media can be edited in a variety of different ways and for different purposes. Satire, parodies as well as cropping content or adding music for artistic reasons would therefore still be permissible on the platform. However, 'in other cases, the manipulation is not apparent and could mislead, particularly in the case of video content'. The result is that 'we remove this content because it can go viral quickly and experts advise that false beliefs regarding manipulated media often cannot be corrected through further discourse' (Meta, 2022).

As this example illustrates, thresholds shape what comes to matter to a machine learning detection algorithm as a deepfake. The threshold foregrounds the importance of context, or *where* the detection model is being deployed and for what end. For when thresholds are optimised according to what a certain institution or company such as Meta considers manipulation or potentially false, then the issue of deepfake detection is not simply an issue of either/or. It has a significant impact on where the threshold for detection and intervention is set. The threshold authorises a certain set of political interventions into the question of what gets removed as a deepfake and what does not. In the case of Meta's Misinformation policy, if something is suspected to be a deepfake and has the potential to mislead – but if it exists at the threshold of detectability, hovering between being identifiable and not – then it will be removed from the platform regardless whether or not this content is a deepfake. The threshold functions to settle the issue of deepfakes, even in cases where the line between, say, deepfake and parody is unclear. Again, echoing Amoore (2020),

> to adjust the threshold of what is 'good enough' is to decide the register of what kinds of political claims can be made in the world, who or what can appear on the horizon, who or what can count ethicopolitically. (p. 69)

The promise of algorithmic detectability is therefore a political claim, because it actively participates in the 'ordering and arranging different ways of being in the world' (Bucher, 2018: 3). Making deepfakes detectable is therefore not simply a question of the nature of the content nor the training data. Rather, it depends on the thresholds produced by interlacing social, political, technical, and economic operations that render things detectable as such.

Certainty

While the threshold of detection algorithms authorises certain lines of intervention and foreclosure, the development of increasingly accurate models for deepfake detection simultaneously produces a veneer of certainty. As Mike Ananny (2016) put it in the context of algorithmic categories, they 'signal certainty, discourage alternative explorations, and create coherence among disparate objects' (p. 103). As such, one of the fundamental aspect of the promise of algorithmic detectability is the production of certainty, that machine learning algorithms can operate as effective detectors of deepfakes. For instance, some of the AI image detectors that are open source on platforms such as Hugging Face have been known to output a 95 percent likelihood that an input image was made by a human when, in fact, it had been generated by a diffusion model.[9] The question here is not so much that models can be mistaken, but rather the certainty with which they are wrong. The certainty they produce in the face of error. In the specific context of deepfakes, this idea of certainty is well captured in a 2022 report published by the Europol Innovation Lab,

> Ideally, a system would scan any digital content and automatically report on its authenticity. Such a system will most likely never be perfect, but with increased sophistication of deepfake technology, a high degree of certainty from such a system could be worth more than the manual inspection. (Europol, 2022)[10]

There is an underlying assumption that there is a parallel between human certainty in detection outputs and the increasing sophistication of those models. The result is that it may give those working with and alongside the machine learning models a certain degree of certainty in the capacity of algorithmic systems in detecting deepfakes and accurately separating between the real and the digitally altered. Moreover, the algorithmic detection model becomes the bedrock on which human confidence in algorithmic detectability is based, perpetuated, and rendered increasingly desirable. The certainty that these models provide – in their probabilistic processing, in their final outputs – is what fuels the sense that they may be the solution to the problem of deepfakes.

Yet, the outcome of the DFDC crucially illustrates the fundamental uncertainty inherent in all algorithmic detectability. For even if the top-performing models had achieved a high accuracy rate in both the public and black box datasets, their accuracy or measure of certainty refer less to their capacity to detect deepfakes in the social world but rather their capacity to detect deepfakes *in a very specific training and testing data environment*. Therefore, a rich source of uncertainty and an enduring point of intractability in algorithmic detection models is the challenge of generalisation. Broadly stated, the notion of generalisation refers to 'how well the trained model performs on data it has never seen before' (Chollet, 2021: 130), and it is a problem widely discussed in the computer science literature. Yet, it is not simply a computational problem; it is also an ethical and political problem. For generalisation, 'does achieve a degree of slippage into the claim that a model can break free of all context, becoming generally useful across multiple domains of life' (Amoore et al., 2023: 10). As Louise Amoore (2020) has observed, drawing on the case of the 2015 riots following Freddy Gray's death in the

custody of the Baltimore Police Department, the algorithms used by the police prevented certain people from joining in the protests, because the algorithms had been trained on various social media data and had produced 'scored outputs of the incipient propensities of the assembled people protesting Gray's murder' (p. 2). In other words, what had been 'simply' social media data were generalised, achieving a degree of slippage into a political claim that algorithms had learned to recognise what a protest is and who could participate in one.

Yet, this question of generalisation in algorithms is not only politically problematic, but it is also marked by fundamental tensions and friction. The question of deepfake detection highlights this. For in the context of algorithmic deepfake detection, generalisation refers to the model's capacity to generalise what it has learned about deepfake data it has been trained on and applying it to unseen cases of deepfakes. But as one computer scientist put it, 'DeepFake detection methods trained using different DF datasets have trouble extending the performance to different datasets' (Lyu, 2020). This is also echoed by Tal Hassner, Applied Research Lead at Meta AI. In a keynote lecture presented at the 2020 Media Forensics CVPR, Hassner explains that 'different GANs have different distributions', which in turn means that 'detectors trained on a single GAN do not generalize well on other GANs' (Hassner, 2020). The implications of this, he continues, is that many state-of-the-art methods 'fail on unseen Deepfakes of slightly different distributions'. In other words, deepfakes generated with one algorithmic model may not be picked up by a detection model trained on a dataset of deepfakes generated by a different algorithm. It also points to the limit points of generalisation as something always achieveable.

This highlights the dangers of considering the model's accuracy as ground for people's certainty in the detection model. A high accuracy level, such as the one outputted by the open source detection software on Hugging Face, may actively obfuscate the fact that an algorithmic detection model could have been trained on an insufficiently difficult or challenging dataset. As Louise Amoore (2019) has emphasised, the single output of a machine learning model reduces the vast multiplicity of possible pathways, of possible lines of action, and effaces the space for doubt. As she puts it, 'it is this process of condensation and reduction to one from many that allows algorithmic decision systems to retain doubt within computation and yet to place the decision beyond doubt' (Amoore, 2019: 8). However, Hassner's observation emphasises that deepfakes are not singular, fixed or static objects waiting to be detected. Rather, they result from variegated distributions, emerging from diverse models. As such, they must be reduced and rendered detectable as a single deepfake. Therefore, the deepfake as a holding term for many, slightly different data distributions foregrounds the fundamental challenges and intractabilities of algorithmic detectability while also problematising its veneer of certainty.

Conclusion: the emergence of a detection society

This article has examined the promise of algorithmic detectability that has followed the emergence of deepfakes in society. It has emphasised, through an analysis of training data, thresholds, and certainty, that while algorithmic detection is a necessary feature of the emerging digital landscape it is also fundamentally insufficient: deepfakes are not

fixed but are the product of changing technologies, changing data distributions, as well as the social politics of setting normative thresholds. As such, the politics of deepfakes should not be reduced to such a singular framework. The development of various algorithmic deepfake detection tools and methods is a result of the anxieties regarding the possible sociopolitical implications of deepfakes. With new algorithmic models such as ChatGPT and Midjourney, these anxieties have been exacerbated. Most recently, Geoffrey Hinton, a well known computer scientist often regarded 'the godfather of deep learning', expressed worries about the impact of these large algorithmic models, stating that 'the internet will be flooded with false photos, videos and text, and the average person will "not be able to know what is true anymore"' (Hinton cited in Metz, 2023). In his view, deepfakes have participated in the making of a culture where there is increasing anxiety around the relativising effects of machine learning and AI, where the lines between real and fake are becoming hopelessly blurred. In fact, deepfakes have come to represent both a nostalgia for a bygone period of stable categories and distinctions as well as a contemporary moment where now gone are the safe grounds on which meaning, discourse, politics, visuality, and indeed common reality can be securely based.

While there is most certainly a need to both take seriously the threats of deepfakes and to prevent the continual erosion of truth and objectivity that they may bring about, deepfakes have also simultaneously fuelled the emergence and normalisation of a *detection society*, one increasingly preoccupied or even obsessed with detecting, demarcating, and reinforcing clear lines between the real and the fake, the true and the false. A detection fever. Here, developments in deepfake detection technologies can be seen as symptomatic of a broader development of an emergent culture of detection, verification, and veridiction. Fact-checkers have become an increasingly indelible part of the social media landscape in recent years (Ananny, 2020) as have algorithmic systems that distinguish between humans and bots (Aradau and Blanke, 2022). In many ways, the body has become a crucial means of biometric identification, such as facial or fingerprint recognition, with companies such as Amazon and MasterCard also utilising 'pay by selfie' as an option for people to authenticate financial transactions (Harvey, 2022). In her book *Posthuman Knowledge*, Rosi Braidotti (2019: 2) opens with some reflections on having to 'confirm you are human' when accessing certain websites by pressing a box which says 'I'm not a robot'. While Braidotti uses this observation as a springboard to explore who or what gets to count as (post)human, the example also illustrates a certain desire to *reinscribe* the human: you *are* a human, *not* a robot. Here, the ontological distinction between human and machine is not simply blurred, but instead is repeatedly confirmed and reinforced. In all these cases, the idea of a detection society includes the phenomena of deepfakes but also goes beyond it. It indicates a feverish trend towards stable and clear dichotomies between the real and the fake, truth and falsity, the human and the robot. Therefore, while the threats of deepfakes may greatly unsettle the boundaries between real and fake, they have also given rise to an obsession with reinscribing and reifying the line between real and fake, between human and nonhuman.[11]

There are therefore ethico-political implications at stake in the algorithmic detection of deepfakes that go beyond individual cases – including bias, gendered harm, accountability, legitimacy, the nature of detection, and what and who counts as human. On one level, deepfakes and their detection can be understood as a pathway into critically

exploring the multifaceted effects of a detection society. As I mentioned in the introduction, the emergence of large language models such as ChatGPT has resulted in a proliferation of AI-detection tools, especially in the domain of education, where anxieties abound concerning AI-generated assignments and essays. As such, deepfakes are not only an isolated problem, but function to foreground wider societal tensions concerning the implications of AI-generated texts and images. This means taking seriously the harms and abuse of deepfakes, the vast majority of which is still gendered (van der Nagel, 2020). But it also means acknowledging that setting the thresholds for the detection of deepfakes is a political and performative attempt to reinforce clear boundaries between real and fake. As such, it is a question of power, 'akin to the exercise and effects of domination' (Aradau and Perret, 2022: 422). Similarly, the use of algorithmic detection is symptomatic of an epistemological need to penetrate into the core of deepfakes and to manage their circulation (or, more precisely, to prevent their circulation). This means that the promise of algorithmic detectability does not simply concern the use of algorithms to detect deepfakes. Rather, it is a promise to clearly draw the line between real and fake as well as to reinforce the legitimacy and naturalisation of that line. To render detectable is to render knowable, legitmate, acceptable, natural, and actionable.

On another level, algorithmic detection, as a framework for approaching the question of deepfakes, has participated in a widespread reduction of the visual and the textual to matters of binary classification: true or false, real or fake. The issue here is that this binarism forecloses and hollows out dimensions of the visual and the textual that necessarily go beyond it: ambiguity, parody, satire, polysemy, aporia and so on. This raises questions such as: what will happen to parody and satire in an age of generative AI and deepfakes? How does it transform political and cultural resistance? Within this framework, all images and texts are emptied of their ambiguity and multiplicity, becoming subject to the tyranny of an *either/or* decisional logic. Yet, as this article has shown, algorithmic detection models struggle to detect deepfakes that fall outside of the data distribution on which they were trained, they do not take into consideration the context in which an image or text emerged, the intentions behind their production, and so on. While simply re-introducing a 'human in the loop' will not solve these challenges either, many of these challenges are actively obfuscated by a reductive binarism – true or false – that fundamentally underpins how algorithmic deepfake detection systems see and do work in the world.

Data availability statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Benjamin N. Jacobsen https://orcid.org/0000-0002-6656-8892

Notes

1. 'Deepfakes' are often understood as a specific use of deep learning algorithms to create 'realistic simulations of a person's face, voice, or body' while the notion of 'synthetic media' covers a broader range of 'media that is enabled or modified by Artificial Intelligence (AI)' (Adjer and Glick, 2021: 9). But as the technologies creating deepfakes are constantly changing as is the popular media coverage, I have chosen to refer to deepfakes throughout this article (but in a loose sense).
2. Deepfakes first came to public attention in 2017, when tech journalist Samantha Cole wrote a piece in *Vice* about an anonymous user on the popular discussion forum Reddit posting pornographic video content that seemed to feature famous female actors such as Gal Gadot and Daisy Ridley. While the content was, in a sense, real (they were indeed real pornos), the faces of Gadot and Ridley had nonetheless been transposed unto the faces of the female adult movie stars, using a generative model akin to the StyleGAN algorithm developed by NVIDIA researchers. The Reddit handle of the user was '@deepfake' (for more on this, see Cole, 2018 and Meikle, 2023).
3. Deepfakes are commonly produced through the use of generative adversarial networks (GANs): a generator network is tasked with generating synthetic outputs which are then fed to a discriminator network tasked with distinguishing between the synthetic outputs and a training dataset of real data. When the discriminator is unable to distinguish between the outputs of the generator and the real training data the GAN is considered ready for use (Goodfellow et al., 2014).
4. One notable example of algorithmic detection models is so-called 'forensic detectors' (Li et al., 2018) that focus on semantic inconsistencies in images and videos, such as skewed facial proportions as well as accessory cues such as 'mismatched earrings' (Corvey, 2021). For an overview of machine learning-based deepfake detection models, see Mirsky and Lee, 2020.
5. Of course, this relationship between deepfakes and the detection of deepfakes is by no means the only example where a solution to a problem is imagined to be immanent to the problem itself. For instance, Sheila Jasanoff (2016) has critiqued the (simplistic) notion that technology could help to prevent or at least ameliorate the problems of nuclear and environmental disaster, for technology continues to participate in the emergence of these problems in the first place.
6. For one notable exception, see Gregory et al. (2021).
7. The full list of collaborators is as follows: Facebook, the Partnership on AI, Microsoft, Cornell Tech, MIT, University of Oxford, UC Berkeley, University of Maryland, College Park, and University at Albany-SUNY (Schroepfer, 2019). On the question of overseeing the competition, Schroepfer (2019) adds that

   > the governance of the challenge will be facilitated and overseen by the Partnership on AI's new Steering Committee on AI and Media Integrity, which is made up of a broad cross-sector coalition of organizations including Facebook, WITNESS, Microsoft, and others in civil society and the technology, media, and academic communities.

8. https://sites.google.com/view/wmediaforensics2020/.
9. See Maybe's AI Art Detector as an example: https://huggingface.co/spaces/umm-maybe/AI-image-detector.
10. The Europol Innovation Lab, mandated by EU Justice and Home Affairs ministers in 2019, aims to 'support the law enforcement community in the area of innovation' and 'identify, promote and develop concrete innovative solutions in support of the EU Member States' operational work' (https://www.europol.europa.eu/operations-services-and-innovation/innovation-lab).

11. Of course, it is worth stating that the need for detection, verification, and fact-checking of information is not unique to this historical moment. As has been noted by Paris and Donovan (2019), for instance, 'the treatment of visual media as an objective documentation of truth is a 19th century legal construct' (pp. 17–18).

References

Adjer H and Glick J (2021) Just joking! Deepfakes, satire, and the politics of synthetic media. *WITNESS*. Available at: https://www.witness.org/deepfakes-and-satire-report-released/

Amaro R (2020) Threshold value. *E-flux Architecture*. Available at: https://www.e-flux.com/architecture/education/322664/threshold-value/

Amoore L (2019) Doubt and the algorithm: On the partial accounts of machine learning. *Theory, Culture & Society* 36(6): 147–169.

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

Amoore L, Campolo A, Jacobsen B, et al. (2023) Machine learning, meaning making: On reading computer science texts. *Big Data & Society*. Epub ahead of print 6 December. DOI: 10.1177/20539517231166887.

Ananny M (2016) Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology & Human Values* 41(1): 93–117.

Ananny M (2019) *Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance*. Knight First Amendment Institute. Available at: https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance

Ananny M (2020) Making up political people: How social media create the ideals, definitions, and probabilities of political speech. *Georgetown Law Technology Review* 4(2): 351–365.

Aradau C and Blanke T (2022) *Algorithmic Reason: The New Government of Self and Other*. Oxford: Oxford University Press.

Aradau C and Perret S (2022) The politics of (non-)knowledge at Europe's borders: Errors, fakes, and subjectivity. *Review of International Studies* 48(3): 405–424.

Braidotti R (2019) *Posthuman Knowledge*. Cambridge: Polity Press.

Bucher T (2018) *If...Then: Algorithmic Power and Politics*. Oxford: Oxford University Press.

Chesney R and Citron D (2018) Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review* 107: 1755–1820.

Chollet F (2021) *Deep Learning with Python*. Shelter Island, NY: Manning Publications.

Clegg N (2019) Facebook, elections and political speech. *Meta Newsroom*. Available at: https://about.fb.com/news/2019/09/elections-and-political-speech/

Cole S (2018) We are truly fucked: Everyone is making AI-generated fake porn now. *Vice*. Available at: https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley

Compton S (2021) More women are facing the reality of deepfakes, and they're ruining lives. *Vogue*. Available at: https://www.vogue.co.uk/news/article/stop-deepfakes-campaign

Corvey W (2021) *Semantic Forensics (SemaFor)*. Defense Advanced Research Projects Agency. Available at: https://www.darpa.mil/program/semantic-forensics

De Vries K (2020) You never fake alone: Creative AI in action. *Information, Communication & Society* 23(14): 2110–2127.

Dolhansky B, Bitton J, Pflaum B, et al. (2020) The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv*. Available at: https://arxiv.org/abs/2006.07397

Dolhansky B, Howes R, Pflaum B, et al. (2019) The Deepfake Detection Challenge (DFDC) Preview Dataset. *ArXiv*. Available at: https://arxiv.org/abs/1910.08854

Europol (2022) Facing Reality? Law Enforcement and the Challenge of Deepfakes. Luxembourg: Europol Innovation Lab; Publications Office of the European Union. Available at: https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes

Ferrer CC, Dolhansky B, Pflaum B, et al. (2020) Deepfake detection challenge results: An open initiative to advance AI. *Meta AI*. Available at: https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

Foer F (2018) The era of fake video begins. *The Atlantic*. Available at: https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/

Gitelman L and Jackson V (2013) Introduction. In: Gitelman L and Jackson V (eds) *Raw Data Is an Oxymoron*. Cambridge, MA; London: The MIT Press, pp.1–15.

Goodfellow IJ, Pouget-Avadie J, Mirza M, et al. (2014) *Generative Adversarial Nets*. ArXiv 1-9.

Gregory S, Leibowicz C, Ovadya A, et al. (2021) *Governing Access to Synthetic Media Detection Technology*. Tech Policy Press. Available at: https://techpolicy.press/governing-access-to-synthetic-media-detection-technology/

Hallinan B and Striphas T (2016) Recommended for you: The Netflix Prize and the production of algorithmic culture. *New Media & Society* 18(1): 117–137.

Harvey A (2022) Today's selfie is tomorrow's biometric profile. *Essay commissioned by Hartware MedienKunstVerein for the House of Mirrors exhibition*. Available at: https://ahprojects.com/todays-selfie

Hassner T (2020) Face understanding at Facebook. keynote at. Media Forensics CVPR 2020. Available at: https://www.facebook.com/watch/?v=883064565545957

Jacobsen BN (2021) Regimes of recognition on algorithmic media. *New Media & Society* 25(12): 3641–3656.

Jacobsen BN (2023) Machine learning and the politics of synthetic data. *Big Data & Society* 1–12.

Jasanoff S (2016) *The Ethics of invention: Technology and the Human Future*. WW Norton & Co.

Johnson D and Diakopoulos N (2021) What to do about deepfakes. *Communications of the ACM* 64(3): 33–35.

Kelleher JD (2019) *Deep Learning*. Cambridge, MA; London: The MIT Press.

Krizhevsky A, Sutskever I and Hinton G (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 2: 1097–1105.

Li H, Chen H, Li B, et al. (2018) Can forensic detectors identify GAN generated images? In: APSIPA Annual Summit and Conference, Hawaii, 12–15 November, pp.722–727. New York: IEEE.

Li Y, Yang X, Sun P, et al. (2020) Celeb-DF: A large-scale challenging dataset for deepfake forensics. *ArXiv*. Available at: https://arxiv.org/abs/1909.12962

Liang W, Yuksekgonul M, Mao Y, et al. (2023) GPT detectors are biased against non-native English writers. *Patterns* 4: 1–4.

Lu W, Liu L, Luo J, et al. (2021) Detection of deepfake videos using long distance attention. *ArXiv*. Available at: https://arxiv.org/abs/2106.12832

Lyu S (2020) Deepfake detection: Current challenges and next steps. *ArXiv*. Available at: https://arxiv.org/abs/2003.09234

Marshall J (2023) As AI cheating booms, so does the industry detecting it: 'We couldn't keep up with the demand'. *The Guardian*. Available at: https://www.theguardian.com/technology/2023/jul/05/as-ai-cheating-booms-so-does-the-industry-detecting-it-we-couldnt-keep-up-with-demand

Meikle G (2023) *Deepfakes*. Cambridge: Polity Press.

Meta (2020) Deepfake detection challenge dataset. *Meta AI*. Available at: https://ai.facebook.com/datasets/dfdc/

Meta (2022) *Facebook Community Standards: Misinformation*. Meta Transparency Center. Available at: https://transparency.fb.com/en-gb/policies/community-standards/misinformation

Metz C (2023) 'The Godfather of A.I' leaves Google and warns of danger ahead. *The New York Times*. Available at: https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

Mirsky Y and Lee W (2020) The creation and detection of deepfakes: A survey. *ACM Computing Surveys* 1(1): 1–38.

Paris B and Donovan J (2019) Deepfakes and cheap fakes: The manipulation of audio and visual evidence. *Data & Society*. Available at: https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf

Ronik (2022) Identifying and tackling manipulated media: *Providing clarity in an age of misinformation with Reuters*. Available at: https://www.ronikdesign.com/project/manipulated-media

Russakovsky O, Deng J, Su H, et al. (2015) ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*: 115: 211–252.

Savage M (2013) The 'social life of methods': A critical introduction. *Theory, Culture & Society* 30(4): 3–21.

Schick N (2020) *Deepfakes: The Coming Infocapolypse*. New York: Grand Central Publishing.

Schroepfer M (2019) Creating a dataset and a challenge for deepfakes. *Meta AI*. Available at: https://ai.facebook.com/blog/deepfake-detection-challenge?utm_source=hp

Stokel-Walker C (2023) Should you be worried that an AI picture of the pope went viral? *New Scientist*, 27 March. Available at: https://www.newscientist.com/article/2366312-should-you-be-worried-that-an-ai-picture-of-the-pope-went-viral/

Strickland E (2019) Facebook AI launches its deepfake detection challenge. *IEEE Spectrum*. Available at: https://spectrum.ieee.org/facebook-ai-launches-its-deepfake-detection-challenge

Striphas T (2015) Algorithmic culture. *European Journal of Cultural Studies* 18(4–5): 395–412.

Vaccari C and Chadwick A (2020) Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* 6(1): 1–13.

van der Nagel E (2020) Verifying images: Deepfakes, control, and consent. *Porn Studies* 7: 424–429.

Weizman E (2017) *Forensic Architecture: Violence at the Threshold of Detectability*. New York: Zone Books.

Biographical note

Benjamin N Jacobsen is a Lecturer in Sociology at the University of York as well as a Visiting Fellow on Professor Louise Amoore's 'Algorithmic Societies' project at Durham University. His current research explores the ethico-political implications of generative modelling and synthetic data on society and culture.