GANs Gone Wild:

Public Perceptions of Deepfake Technologies on YouTube

2471280P
19108486
90164042

Presented in partial fulfilment of the requirements for the Degree of International Master in Security, Intelligence and Strategic Studies

Word Count: 21,836
Supervisor: Dr. Petr Špelda
Date of Submission: 23 July 2021

Table of Contents

INTRODUCTION

PART ONE

PART TWO

CONCLUSION

Reality is a very subjective affair. I can only define it as a kind of gradual accumulation of information...You can get nearer and nearer, so to speak, to reality; but you never get near enough because reality is an infinite succession of steps, levels of perception, false bottoms, and hence unquenchable, unattainable...So that we live surrounded by more or less ghostly objects.

– Vladimir Nabokov (1962: 10)

Abstract

Deepfake technologies are a form of artificial intelligence (AI) which are based on generative adversarial networks (GANs), a development which has emerged out of deep learning (DL) and machine learning (ML) models. Using a data range which spans the years 2018 – 2021, this research explores public perceptions of deepfake technologies at scale by closely examining commentary found on the social video-sharing platform, YouTube. This open source, ground-level data documents civilian responses to a selection of user-produced, labelled deepfake content. This research fills a gap regarding public perception of this emerging technology at scale. It gauges an underrepresented set of responses in discourse to find that users demonstrate a spectrum of responses which veer between irony and concern, with greater volumes of commentary skewed towards the former. This study of user commentary also finds that YouTube as a wild space ultimately affords reflexive and critical thinking around the subject of deepfake technologies and could prove to be effective as a form of inoculation against disinformation.

Keywords: Deepfake, Artificial Intelligence, Technology, Politics, Discourse.

Acknowledgments

In the aftermath of the events of this year and the last, some of the most reliable systems we had come to trust were disrupted, unveiling the fragilities of global order. In the years to come, there is little doubt that the innovation landscape will continue to accelerate the pace of change, driven in part by some of the emerging technologies discussed here. Going forwards we can say with certainty that the dichotomies we are confronted with, will more often than not, be more complex than they appear at the surface.

With this in mind, I would like to express my deepest gratitude to the inimitable Dr. Petr Špelda for shining a light on the black box of AI; his expertise, recommendations and sense of perspective remain unparalleled – he has been a beacon of brightness throughout this research journey. Along the way, I have been grateful for the conversations that have come about from presenting initial findings of this research at IFJP, Nottingham Postgraduate Conference in Politics and International Relations and RightsCon. I am also indebted to the staff of the FSV faculty at Charles University, most notably Dr. Vit Stritecky and Dr. Filip Vostal, whose feedback proved invaluable in the scoping process. Many thanks also to Dr. Nicola Bozzi, Dr. Markus Bresinsky and Dr. Benoît Pelopidas for their generous offers to proofread early drafts.

And finally to C, to whom this paper is dedicated to – thanks for being my North Star, in absentia.

List of Figures

List of Tables

INTRODUCTION

1.1 Background

“Deepfake” is often used as a colloquial term for GANs; at the most basic level, GANs make use of two competing artificial neural networks, one of which works to produce a falsified interpretation of a likeness whilst the other attempts to detect it (Porup, 2019). As a technology which emerges out of DL, a sub-branch of AI, GANs’ applications span a wide array of different contexts, with the output of these adversarial models being as varied as its training inputs; from text to audio and audiovisual data.

It is notable that this AI sub-branch, which began as a pet project of ML scientists designed to push the limits of neural networks, should enter the mainstream and capture public imagination to the degree of advancing the dialectics of disinformation. The genesis of adversarial networks emerges from the research of computer scientist Ian Goodfellow, whose pioneering work with adversarial examples have scaled with such alacrity and technological sophistication that its latest iterations are able to fool humans in real-world security settings (Goodfellow, McDaniel and Papernot, 2018). As a relatively recent development in the realm of AI, GANs as models for training DL systems might have been relegated to the obscurity of the laboratory if not for their applications in popular culture, with the term “deepfake" first popularised by English-language social networks (Cole, 2018). The original use of the term to describe audiovisual content trained using DL models is attributed to Reddit fora from 2017 onwards (Brundage et al., 2018: 49). From this period, DL technologies have become relatively commonplace, with an explosion of user-generated content spreading across social networking sites since, particularly on video-hosting platforms (Westerlund, 2019). The progressive mainstreaming of this branch of ML, out of a controlled lab environment and into complex real-world scenarios, is commonly referred to as entering the “wild” (Prabhu and Birhane, 2020).

Deepfakes have thus come to be seen as a disruptive force at a political and industrial level, and in some instances this has led to deepfakes being cast as part of a toolbox of disinformation in a “post-truth” age (Fazio, 2020a; Helbing et al., 2017; Jasonoff and Simmet, 2017). However, this research takes the use of “post-truth” to signify an accelerated use of disinformation tools at scale rather than framing this current security environment as an entirely new phenomenon or a wholesale era of untruth. The pervasiveness of deepfake technologies in the public domain poses a pertinent topic for study given the massive growth of such content in the years since 2017 – a trend which is still ongoing (Calderone, 2021). Deepfakes have garnered notoriety for their deployment in several high profile criminal cases, including the use of audio deepfake technology in corporate theft (Ajder et al., 2019b) and the en masse creation of falsified profiles featuring GAN-based profile photos (Nimmo, et al., 2019). Deepfake AI has otherwise been dominantly used to produce non-consensual pornographic content (Gosse and Burkell, 2020; Maddocks, 2020; Osipian, 2020). The slipperiness of this topic is partly rooted in the fast-democratisation and accessibility of these still-nascent technologies, which has seen progressively sophisticated results with little training data (Lewis and Nelson, 2019). These accelerated growth factors have posed problems for recent literature in this area which still comes from largely short-term observation and forecasts which run into fast-obsolescence issues (Harwell, 2019 and is sometimes aptly dubbed, in adversarial terms, as a “race” against the odds (Murphy, 2019; O’Sullivan, 2018).

Given the status of deepfake technology as a moving target, the primary research objective here lies in exploring public reactions to deepfake content. This research specifically addresses the following questions in detail:

- RQ1: How do public perceptions of deepfake technology play out in ground-level discourse?
- RQ2: How does ground-level discourse map onto wider discussions of disinformation?

The first of these questions addresses the dimensions of public perception through user-generated commentary on a sample of recently published deepfake videos. This is sampled from a selection of publicly accessible, labelled content. The selection of videos here have been chosen for featuring explicitly political content and for the ample engagement they have attracted in the form of user responses. These allow for a timely discussion of the sociotechnical dimensions of this topic, being drawn from YouTube, a popular video-sharing social network. The approach here establishes a set of standards for examining the longer term outlook of this phenomenon and is useful for benchmarking similarities between other tools that could be co-opted as part of disinformation campaigns.

The second of these questions explores how public perception applies to issues of real-world deployment. This paints a more established picture of the risks and opportunities for broader developments at the nexus of AI and disinformation. It queries issues of robustness within a growing body of safety literature, both within the technical domain of AI and the wider realm of security studies (Yampolskiy and Fox, 2013; Amodei et al., 2016). Robustness here is taken in the traditional sense to refer to the technical stability of models, particularly in real-world settings, but is also applied in issues of replicability in adjacent contexts (Rudner and Toner, 2021a). For instance, while AI and its sub-branches are themselves subject to a number of issues concerning technical robustness, the methods used to explore the phenomenon so far are subject to their own issues of sociotechnical robustness which are ill-defined (Rudner and Toner, 2021b). This research hence explores these technical and methodological fault lines and unpacks the sociotechnical dimensions by reconciling these with a more critical set of insights into the complex environment of deepfake deployment.

This research subsequently makes a novel contribution in two ways. Firstly, by offering an empirical insight into the ground-level discourse around deepfake deployment; an area which risks being overlooked by policymakers. It evaluates the human issues at the core of these AI-based applications by sampling a cross-section of public discourse from relevant user-produced content.

Secondly, this research makes a measurable and timely contribution to the growing body of safety literature which is so far preoccupied with a narrow set of themes and sources. This research complements academic literature on the technical foundations of DL models (Evtimov et al., 2020) by broadening the sociotechnical dialectics on safety and disinformation literature and providing a more robust approach for exploring user-generated deepfake content in the wild.

Jointly, these contributions shed greater light on the issue of audiovisual disinformation at scale and help to better understand the place of deepfake technologies as an emerging phenomenon, in order to manage their spread and sources.

1.2 Research Structure

On the structure of the following sections, the first part of this research will discuss the origins of DL technologies and the theoretical considerations of post-truth framing. This first part provides context for the research design and methodology section in the second part of this paper.

This first set of chapters effectively acts as a literature review, scoping out the major debates in deepfake disinformation and the state of play regarding these; specifically the technical and theoretical considerations which feed into the limits and potential of deepfake technologies, as well as how the dialectics of public discourse affects the perception of this nascent but fast-growing phenomenon.

The second part of this research paper provides an insight into the hows and whys of the research, detailing the process and the findings accordingly. These later chapters contain an insight into methodology and the research design used to empirically ground public perception in user commentary, as well as documenting the data collection and selection process. This is followed by a discussion of the findings and the study's limitations of the outcomes, followed by further directions for research and policy-orientated next steps.

PART ONE

"All things exist as they are perceived: at least in relation to the percipient...The mind is its own place, and of itself can make a heaven of hell, a hell of heaven."

— Percy Bysshe Shelley (1890: 65)

2. Literature Review

This section provides an overview and a critical discussion of the main scholarly literature on the nexus between disinformation, deepfakes and public perception. This features a discussion of the scholarly literature in tandem with grey literature; the pieces here have been selected to reflect upon arguments about how the democratisation of AI technologies and how they might augment the scale and speed of disinformation. This section effectively looks to advance these discussions and ground the argument by debunking myths around the operational arena of deepfake deployment.

As the latest tool to be absorbed into a growing canon of concerns around disinformation, deepfake technologies have been positioned as a force for total democratic disruption as part of the fabric of a post-truth age (Watts and Hwang, 2020). However this section argues that post-truth when it is taken as a term encompassing current operating conditions, or the 'wild,' of deepfake deployment is an oversimplification. This dulls the environmental complexity and full scope of the issues at stake. It is important to consider a more comprehensive history of audiovisual disinformation to balance the positive values of deepfakes as an emerging technology alongside its more markedly negative applications. As such, we also need to understand how the historically contested status of truth contextualises the current status of this technology. This requires us to turn to the sociotechnical roots of deepfake technology to better understand the origins and development of these technologies and the implications they have in the public domain.

The following chapters prime the methodological exploration of deepfake disinformation, building on the current definitional quandaries in the field. This looks at the wider political dimensions of disinformation and interrogates the role of public perceptions of this technology, providing a foundational basis for the discussion later on. Together, these research questions establish a more robust critical chronology of developments in a still-nascent field which is fraught with speculation and uncertainty.

2.1 Into the Wild: Deepfake Disinformation and the State of Play

2.1.1 Gab to GAN: Text to Audiovisual Models

Deepfake technology, as a newcomer to the pantheon of disinformation tools, finds itself branded not merely a propagator of untruths, but an arbiter of it (Villasenor, 2019b; Weiss, 2020). The nature of truth as it is mediated through mass media has a contentious history but despite this, contemporary discourse around post-truth framing tends to defer to the current information environment as a new, overarching condition (Citron and Chesney, 2019; Min-Yeong, 2020; Spicer, 2018). But arguably, the current information environment carries similar precarities around mediated truth as previous eras rather than breaking from structurally created conditions entirely.

Deepfakes as the latest medium to spark frisson in the complex, real-world environment are positioned as an arbiter of a perceived shift in reality, with many researchers and journalists suggesting that deepfakes blur the line between fact and fiction itself (Choi, Oh and Lee, 2019; Kalpokas, 2020; McBeth, 2018). However, this purported blurring of boundaries fails to consider the fact that the pursuit and mediation of truth through media has always been fragile (Metzger and Flanagin, 2013). Danielle Citron and Robert Chesney's (2019) article titled *Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics* posits that deepfakes arrive at a point in history where fact is inseparable from fiction, suggesting that deepfakes spur an unavoidable "descent into a post-truth world," for the indirect effects they will provoke in a widely-diffused set of sectors. Citron and Chesney's work is prolifically cited across recent writing on deepfake disinformation for the salient points made about the always-on newscycle and social media ecosystem and how these have given rise to post-truth as a framework. The opinions expressed by Citron and Chesney balance the proliferation of this technology with its net positives, and see the aforementioned effects of disinformation supported by wider indicators on some of the more noticeable permutations of deepfake content, for instance across social media (Mericle, 2020). However, there is no evidence beyond speculation that the greater circulation of deepfake content on the whole is enough to warrant a complete fragmentation of societal norms, nor is there robust justification given to how a post-truth framework should be defined. It is difficult to vouch for a break between a 'truthful' period which was inherently verifiable and a sudden blurring of reality by disinformation, and as such, the current state of play should be seen instead as one largely altered by access and democratisation of these AI technologies and supporting factors such as the speed and scale of content production.

The hasty and speculative judgements made of deepfake technology follow from prior historical precedents. New media in every period has been treated with some suspicion, in part due to disparities in media literacy, as a matter of select groups being privileged to particular types of knowledge above others (Ștefăniță, Corbu and Buturoiu, 2018). But it might be suggested, controversially, that there is nothing novel about deepfake technology itself, having developed out of familiar innovations in recent history. Deepfake technology is neither a siloed development in the realm of DL nor does it break with the conventions of disinformation's visual forms which often need very little technical effort to sway public opinion (Ignatidou et al., 2019). Deepfake technology is iterated out of developments in AI and its sub-branches with its own "deep" issues (Marcus, 2018) and sociotechnical baggage (Martinez and Castillo, 2019; Yadlin-Segal and Oppenheim, 2021). And so in regards to approaching disinformation in the round and understanding the operational conditions of deepfake technologies and their role within the disinformation ecosystem, we should first turn to text models as a precursor to developments in audiovisual models, their shortcomings and the risks they are perceived to carry.

Text models such as Generative Pre-trained Transformer (GPT-2), released in 2019, share several synergies with the system architecture and capabilities of their basic visual counterparts. These GPT models designed by OpenAI are able to generate text and answer queries at a near-human level (Toner, 2020). This has prompted public fears around their usage and proliferation into fake news and its role in creating realistic botfarms which bypass conventional detection methods (Thornhill, 2020). This set of anxieties is further provoked by the accessibility and distribution of the model as open source software (McCain, 2020). But despite these fears its latest iteration, GPT-3, was publicly released as a beta version a year later in 2020 (Thornhill, 2020). This ability of text models to superficially reflect "human" qualities has been viewed as a possible security threat, but despite being lauded for performing statistically close to benchmarks set to human standards, they are largely unsustainable for devising longer passages, producing repetitive or nonsensical passages in the process (Geirhos et al., 2020: 10).

These issues are parallel to the similarly shallow assessments made of visual models and by extension, deepfake technologies at large. This is especially if we consider that the most pervasive AI models in the wild are often “cheapfakes” (Paris, 2019) or “shallowfakes” (Ignatidou et al., 2019), which are disseminated even without the exceptional formal qualities or technical sophistication of their deepfake counterparts. These shallowfakes are as easily spread as they are created, given that they do not require an abundance of source code (Schick, 2021a).[1] This shallowness is a key theme which will be explored throughout this research, with the earliest iterations of DL technology in the public realm eliciting a wide spectrum of responses, from the utopian to the dystopian. The established precedent of text models and their neural networks exemplifies this well, with public perception veering broadly between fear of potential consequences (Chakhoyan, 2018), acceptance of usefulness (Hao, 2020) and disillusionment with limitations (Knight, 2020). Similar attitudinal transformations are also evidenced in visual processing models, which follow closely with a comparable range of commentary (Kietzmann, Mills and Plangger, 2020). But owing to the increasing accessibility of this technology, manipulated media’s scale and speed of deployment is set to increase (Gillespie, 2020). The concerns for the capabilities of deepfake technologies centre on audiovisual content as the most potent outlet for disruption. This is especially true when we account for their possible

[1] Web apps such as TalkToTransformer.com and mobile apps such as FaceApp, which allows users to alter photos of human subjects by de-aging, aging and gender swapping them. Later iterations of these apps such as Faceswap feature the ability to splice faces between celebrities and input user photos. See Tolosana et al., 2020 for more on this. These are both designed for static visual content, but a recent app release in Wombo.ai, allows for short lipsync videos to be generated. All the aforementioned apps usually require only a single piece of training data. But despite being largely entertainment apps for public consumption, these have been attributed to state-level actors, such as FaceApp which gained notoriety for its use of user data and Russian state affiliation. For more on this see Luca Guarnera, Oliver Giudice and Sebastiano Battiato’s (2020) work on fighting deepfakes by exposing convolutional traces on images.

deployment in formal campaigns and subsequent perception by the public in high stakes events such as elections (Polyakova and Fried, 2019).

Audio, visual and audiovisual media have long been held up as a form of verification but have also been historically vulnerable to manipulation and use for the purposes of deliberate deception. This deliberate deception takes on a range of multimedia formats, from photoshopped imagery (Vaccari and Chadwick, 2020) or “fauxtography” (Fazio, 2020b; Wang et al., 2020), to the creation of convincing audio mimicry (Chintha et al., 2020).[2] The picture’s status, or “pictorial politics,” has been largely rooted in specific qualities of the image, namely its “duality, its vagueness, and its temporality” (Grave, 2019) which are all regarded as vulnerable to exploitation by malicious actors and used in disinformation campaigns. The nexus between emerging technologies such as DL and audiovisual deception is a significant preoccupation for the realm of AI safety (Verdoliva, 2020: 1). As such, the suggestion that audiovisual formats of DL technologies might be used for disinformation has been regarded with particular potency. Audiovisual content’s long-held status as an incorruptible unit of objectivity (Jurgenson, 2019: 32) has seen it used in both lay and administrative contexts for proof and prosecution (Grave, 2019; Woolley, 2020). The generic GAN comprises generator and discriminator networks which play out a “minimax game” and compete to distinguish progressively convincing approximations until an impasse is reached between generator and discriminator (Špelda and Stritecky, 2021: 94). This has resulted in the real-time manipulability of deepfake technologies in the public domain, which has roused new security concerns for the verification of information as it is presented in public contexts (Thies et al., 2018).

[2] For the longer history of deliberate multimedia deception refer to Mika Westerlund’s (2019) review of deepfake technology, which provides an excellent summary of the evolution of disinformation.

But the malicious use of both text and audiovisual models have been detected with relative speed so far, with disproportionate amounts of publicity afforded to single incidents. These are not dissimilar to the patterns of creation, detection and debunking in the media preceding DL technologies. Evidently, advances in the field of AI ethics and safety develop alongside public perception of these technologies, which need to be accounted for in the broader discourse surrounding deepfake deployment. So despite these more disruptive elements of AI innovation, deepfake technology needs to be framed less as a “new frontier of AI trickery” (Venkataramakrishnan, 2020), and considered equally for its shallowness of audiovisual mimicry. Root causes of how deepfake technology feeds into disinformation can be inferred partly from the public reception of these technologies. Tracking the reception of such technologies could potentially form a set of critical countermeasures to controlling or mitigating the worst effects of deepfake technology alongside existing safety measures. Understanding how deepfake content could be disruptive requires us to probe its audience beyond the assumptions and risk assessments born out of a precarious post-truth framework.

2.1.2 Disentangling Superficial “Post-Truth” Falsehoods

The dialectics of a post-truth framework suggest that deepfakes are a novel medium and that this novelty is enough to disrupt democratic principles beyond the scope of other formats which have come before it (Min-Yeong, 2020). These dialectics put forward a set of assumptions which fail to balance the foundational shallowness of deepfake technology with the broader mechanics of disinformation. This understanding of shallowness is critical for making sense of the complex environment which contextualises the ‘wild’ in which deepfakes are deployed. This complex environment comprises other democratically disruptive forces such as fake news, disinformation and misinformation. And though these three terms are definitionally distinct from one another, their convergence comprises the complex environment of deepfake disruption.

It should be explicitly noted here that disinformation and misinformation are separate forms of disruption; while the former carries deliberate intent to mislead and is often conflated with misinformation, the latter is more appropriately distinguished as simply being “incorrect information” (Landon-Murray, Mujkic and Nussbaum, 2019). Fake news has otherwise been defined as a term designated “by political actors to describe content from mainstream news outlets that contradicted particular political agendas, with the intention of discrediting that content,” (Tredinnick and Laybats, 2019: 92) and has notably transformed from a label for misinformation spread by social networks into one which absorbed into the wider web of disinformation. This has been fed by the emergence of personalised and increasingly targeted media around the 2010s (Florea, 2013), also known as microtargeting, which has in turn spurred burgeoning discourse around filter bubbles and echo chambers (Krafft and Donovan, 2020; Manor, 2019). And though the significance of echo chambers has been disputed by some (Dubois and Blank, 2018), their role in bolstering fake news, most notably in the wake of Brexit and the 2016 Trump campaign, is more difficult to deny (Cadwalladr, 2017; Spicer, 2018). It is important here to earmark these several distinctions to acknowledge how layered and multifaceted the development of the information environment has become, because collectively the nexus between these forms of disruption constitutes a shift in the conditions of contemporary content production.

These forms of disruption have been shown to be as potent as one another in their ability to accelerate and amplify the flow of disinformation. The convergence between fake news and misinformation (Rubin, 2019) are attributed with the decline of public opinion, the amplification of the echo chambers and unconscious consensus building (Kawahata, 2019). Deepfakes, when contextualised by the preoccupations of their use as a tool of disinformation plays witness to a similar dynamic. Falsified audiovisual content, despite its variable qualities and formal shallowness across the spectrum of deep to shallowfakes, is perceived by policymakers as an entity which can potentially cloud public thinking. This is also due to the fact that both these forms are equally likely to be disseminated (Venkataramakrishnan, 2019) though the distinction between deepfakes, which make use of algorithmic models and formal training data and its less technically intensive shallowfake counterparts, are markedly different in the intensities of their training processes. The potential of audiovisual disinformation to confound the information environment at scale grows proportionally to the continued democratisation of these technologies. And so, with deepfake technology becoming more frictionless to access and less labour intensive to use, there is some concern that audiovisual forms of disinformation are likely to keep proliferating into public spaces faster than they can be controlled or detected (Harwell, 2019; Urbani, 2020).

But to some extent, the current assessment of deepfake technologies is not as robust in justifying appropriate ends for its means in line with the concerns facing deepfake technologies. Current assessments amount to no more than mirroring attempts to police strong AI or AGI/ASI (Baum, 2018a), which are speculative at best. Audiovisual disinformation largely sits within the scope of weak AI, but just like AGI/ASI, deepfake technology has been positioned as a political tool which requires a fundamental rethinking of security mechanisms which range from policing to lawmaking and technical counters (Bechmann, 2020; Kalpokas, 2020) despite the relative technical and theoretical shallowness of this technology at present. These security assessments for the most part lack an empirical basis and are based in anticipatory ethics (Ahmed, 2020), forming an apt parallel to similarly preemptive narratives within AI safety (Everitt, Lea and Hutter, 2018). These preemptive responses tend to rely on a deductive approach to historical precedents, such as security dilemmas, and champion zero sum games of AI superiority (Yampolskiy and Fox, 2013). These often fail to take into account key tenets of disinformation's spread, such as the role of the audience (Shu et al., 2020; Starbird, 2019). The speculative state that deepfake technology finds itself mired in is more akin to a moral panic without recourse, and tends to observe the symptomatic elements rather than unpack the structural issues behind the phenomenon. And while safety is key to the study of AI, the misapplication of safety concepts in this instance only serves to trigger moral panic. This obscures the broader dialectics around DL technologies, and worse, risks hindering the positive developments and other progress made in this area. The next section hopes to explore these themes to dissect these moral panics to form a more robust assessment of the technology.

2.1.3 The Dangerzone of Egregious Content: Moral Panics and the Funny-Bad Nexus

It is estimated that 14,678 deepfake videos were hosted online in 2019; nearly a 100% increase from the year previous (7,964). A recent search yielded 10,200,000 results, demonstrating an exponential increase of around 1,218%.[3] The significant majority of deepfake videos (96%) are estimated to be pornographic in content (Ajder et al., 2019b) whilst the remaining non-pornographic content that can be found on the surface web is entertainment-based content featuring public figures such as celebrities and politicians (Li et al., 2020). This is due to the widely-available training

[3] It is noted that the true extent of these figures are likely to exceed the numbers given here which only account for surface web numbers taken from an aggregate of English-language search engines (Google, Bing, Yahoo, Duckduckgo, Ask.com) as of March 2021.

data available to deepfake video creators in formal repositories (Xie et al., 2020).

Since 2019, the security discourse surrounding audiovisual disinformation has been bolstered by the greater visibility afforded to deepfake technology's use in blockbuster Hollywood productions (Bradshaw, 2019), prolific legal cases (Ryan and Hii, 2021) and shifts in policy-making with the signing of federal legislation on deepfakes into law under Section 5709 of the National Defense Authorization Act (NDAA) (Hale, 2019).[4] Deepfakes have been dubbed a "new weapon of choice" (Gieseke, 2020), "disinformation on steroids" (Citron and Chesney, 2020) and are otherwise seen as disruptive to democracy for their erosion of public trust and faith in established institutions (Whyte, 2019) when used as part of disinformation campaigns. These views are echoed by wider grey literature and in journalism (Grothaus, 2021; Woolley, 2020) which posits "infocalypse" scenarios (Schick, 2020b) to fuel an escalatory narrative of moral panic which concerns itself with high-level political disruption.

Even though successful high-level deployment of algorithmically-trained audiovisual disinformation is currently rare, deepfake technologies have in some instances been used to achieve criminal objectives (Tammekänd et al., 2020) and have otherwise gained notoriety for their role in creating non-consensual pornography (Maddocks, 2020). These activities have been largely carried out by malicious actors such as organised criminal groups (Hartmann and Giles, 2020; Bateman, 2020), but are also susceptible to be spread by those who encounter it most often, namely laypeople. In the instance of non-pornographic video content, these range from the cautionary to the educational and the satirical, with some of the most pertinent examples featuring actors such as Bill Hader as Tom

[4] Prior to this, several state-wide bills passed before this, criminalising the production of deepfake content. For more on this, see Kelsey Farish's (2020) work on whether English law should adopt similar intellectual property laws to those passed in California.

Cruise, with favoured tropes carrying out faceswaps of Steve Buscemi and Nicolas Cage on the bodies of various other celebrities from popular film scenes (Figure 1). Some of the first pieces of deepfake content to become mainstream examples featured explicitly political subjects from the likenesses of figures from Barack Obama to Donald Trump. Returning to issues of scale and speed, the lay sophistication and personalisation of these media sees an imminent trend towards more targeted or “microfake" content, as per ongoing developments in microtargeting (Ascott, 2020). This more targeted content could potentially see the even-faster acceleration and upscaling of such technology, in accordance with what we have seen with the information ecosystem so far.

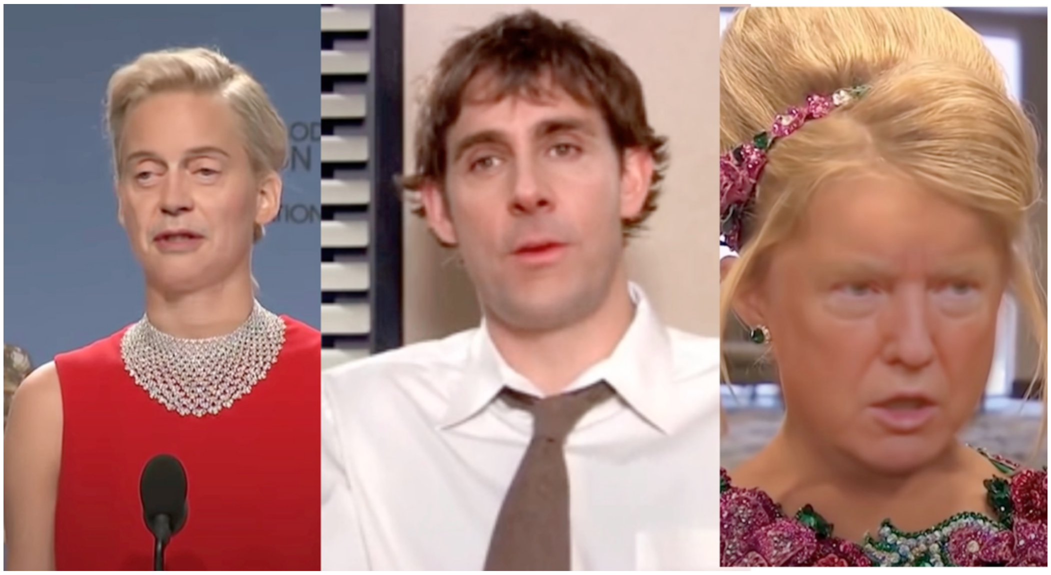


Figure 1. From left to right: Steve Buscemi, Steve Carrell and Donald Trump deepfaked onto the bodies of other public figures (2020). Source: Mashable

This moral panic narrative has garnered a scale of contrary reactions, with claims that policy and legislation are not going far enough to regulate these emerging technologies (Nonnecke, 2019). And that efforts towards accountability have been similarly shallow, serving to overencumber and over-broaden the scope of how deepfakes are defined in law thereby burdening legitimate users, such as commercial ventures (Schapiro, 2020). It has also been suggested that regulation at this relatively early stage is reductive (Dowdeswell and Goltz, 2020). These assessments are predominantly steeped in US-centric interests about terrorism, corporate loss and the diplomatic disruption of power in the West (Braddock, 2020; Antinori, 2019).

But even prior to concerns around deepfake technology itself, the entertainment sphere saw the emergence of an adjacent phenomenon, with the sophisticated ability to map features onto bodies and faces. Rather than intensive training through DL algorithms, Hollywood blockbusters have been the sandbox of sophisticated audiovisual mimicry, with decades-worth of labour intensive efforts used to manually map computer-generated effects across a diverse spectrum of actors and scenarios to believable effect. From *Forrest Gump* (1994) and *Gladiator* (2000) which make use of the post-editing process to anachronistically insert actors into historical footage, to more recent films such as *Gemini Man* (2019) and *The Irishman* (2019) where the sophistication of facial swaps are painstakingly mapped onto actors' bodies in real-time, allowing actors to play a de-aged version of themselves (Figure 2) (Hall, 2018: 58). Though deepfake content increasingly matches these same cinematic capabilities, with effects traditionally achieved with professional staff and equipment being parsed with fewer resources, Hollywood's long-term cultivation of a cinematic uncanny also serves to prime an audience in many ways for the rapid manufacture and increasingly democratised release of deepfakes into the wild (Foer, 2018).

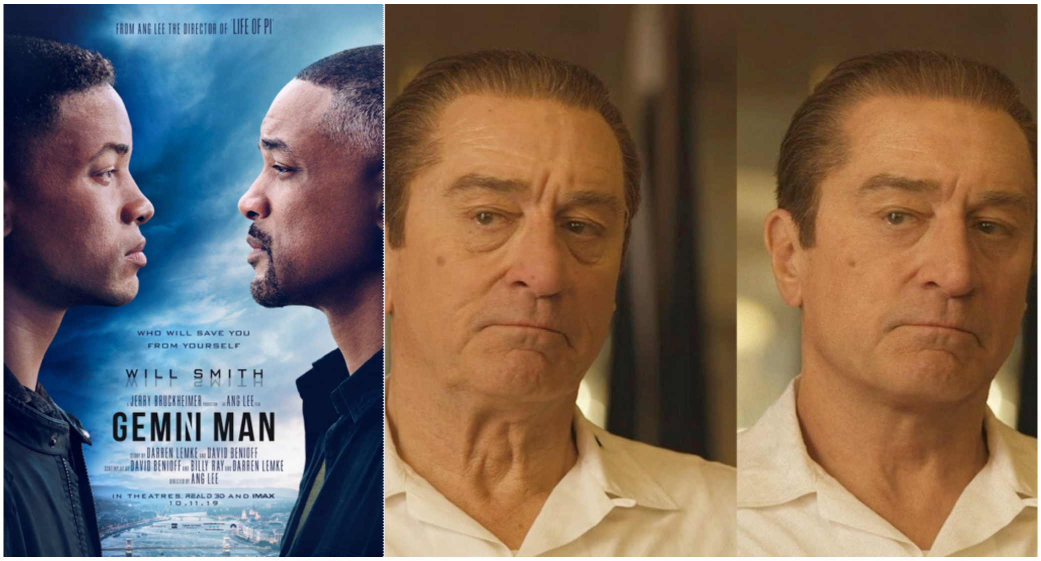

FROM ANG LEE THE DIRECTOR OF LIFE OF PI
WHO WILL SAVE YOU FROM YOURSELF
WILL SMITH
GEMINI MAN
IN THEATRES, REALD 3D AND IMAX
10.11.19

Figure 2. Gemini Man (2019) and a still taken from The Irishman (2019). Source: Petapixel

In other words, films which are reliant on special effects also tend to invite critique of said effects. Herein lies the issue with deepfake technologies in the wild; the use of CGI signposts manipulation, and in the context of blockbuster film productions, suspended disbelief creates skepticism (Bradshaw, 2019). But there is still some disparity between audiences' overall understanding; an audience watching a film with high production values is primed very differently to an audience encountering non-labelled deepfake content in the wild. Context is thus critical for an audience, with the insidiousness of encounters with even labelled deepfakes in the wild providing a clear environmental difference when contrasted with cinematic spaces. While comparisons between encountering labelled and unlabelled deepfake content are subject to an expansive range of factors across a spectrum of human biases (Nickerson, 1998), the field currently lacks a benchmark for establishing how much of a security risk deepfakes could be. And so, this research very specifically homes in on labelled deepfake content hosted on YouTube, given the contentiousness and ethical difficulties of studying unlabelled deepfake content. Concentrating on organic public perceptions towards deepfakes requires us to narrow down the number of variables and focus on a specific type of content – in this instance, content which features a curated selection of well-known public figures. This is a more feasible direction for approaching an underexplored and slippery area which presently lacks robust methods for measurement and benchmarking.

2.1.4 Regulation and Adam's Left Rib: A Conundrum

Regulation of deepfake content is presently problematic, with assessments of deepfakes' technological capabilities framed by the moral judgments of solely policymakers. This often fails to account for the countermeasures currently being taken in the technical field of AI safety, as well as its more epistemological practices and the outright limits of this technology. This is especially the case in the field of detection. On this note, it is useful to briefly return to text models as a precedent for the safety concerns surrounding deepfakes. GPT-2 and GPT-3 have come under scrutiny as progenitors of neural fake news but can also be used inversely to generate methods for detection, in effect countering the same issues created by the models in the first instance (Zeller, 2019). Countermeasures can also be seen in DL models based on recurrent neural networks (RNN) (Güera and Delp, 2018) and convolutional neural networks (CNN) (Shah et al., 2015) which have also been used for deepfake detection, effectively turning the weaknesses of DL architectures into an adversarial foil. This feedback loop is hereafter referred to as the Adam's Left Rib Conundrum.[5] But the trend towards the greater quantity, quality and variety of deepfake technologies (Collins, 2019) means that ground-level dangers which affect civilians at a personal level are fast-becoming issues of national security.

[5] Referring to the Biblical Adam, who in the book of Genesis has a female counterpart shaped from his left rib in the form of Eve, the first woman, who ends up leading him to knowledge but also expulsion from the Garden of Eden. The reference here highlights the apt adversarial parallels which form a continual feedback loop between humans and falling prey to new knowledge generated by the things we create in our own image. Deepfake technologies are a technological issue of our own making, after our own image and the subsequent attempts to banish the issue entirely only serve to generate a new iteration of the problem.

As technologist Aviv Ovadya has noted, disinformation "doesn’t have to be perfect — just good enough" (Warzel, 2019). Ovadya’s comment refers to a manipulated video of Nancy Pelosi slowed down to present the subject as drunk (Calderone, 2021). This video, which was subsequently retweeted by Donald Trump to significant traction, demonstrates the extent to which even less technically sophisticated audiovisual content in real world environments is vulnerable to adversarial perturbation and how volatile this can prove in a political context (Kelly, 2019; Donovan and Paris, 2019) (Figure 3).

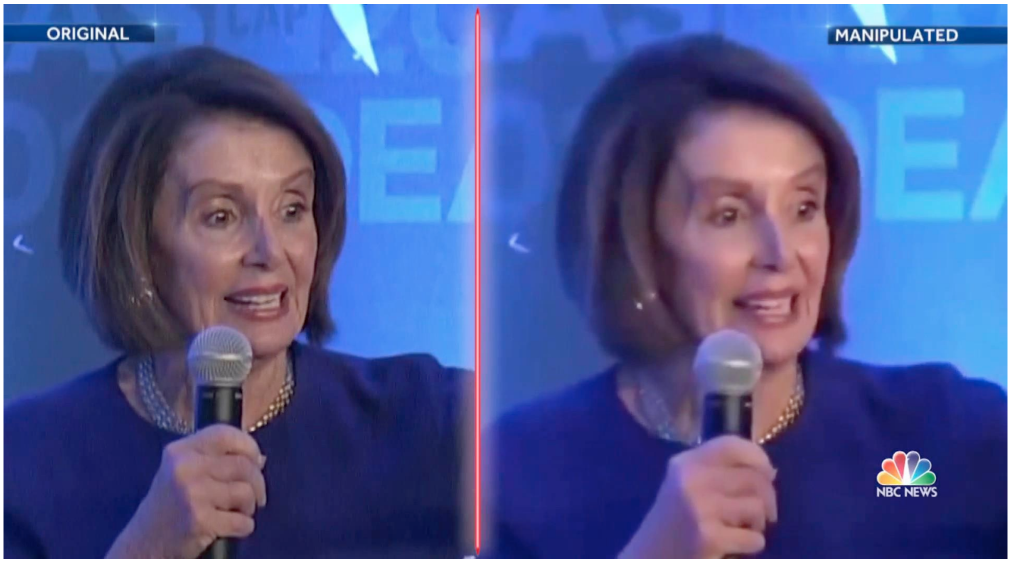


Figure 3. Screenshot of digitally altered clip featuring Nancy Pelosi (2019). Source: NBC News

While a more thorough consideration of political context and deployment is important, there is not yet enough known about the root of deepfake disinformation’s spread and dissemination especially at scale. This spread can be partly attributed to hosting of user-generated deepfake content on social media platforms, with the increasing accessibility of these technologies (Allen and Chan, 2019; Congressional Research Service, 2019) progressively lowering the barrier to entry for frictionless creation and sharing of deepfake content. Notably, the period from 2017 onwards has shown synchronous surges in searches for “deepfake apps” against the main trendline of searches for “deepfake” as a term (Figure 4). This data, obtained from Google Trends, indicates some public appetite across global audiences for engaging with deepfakes through software applications.

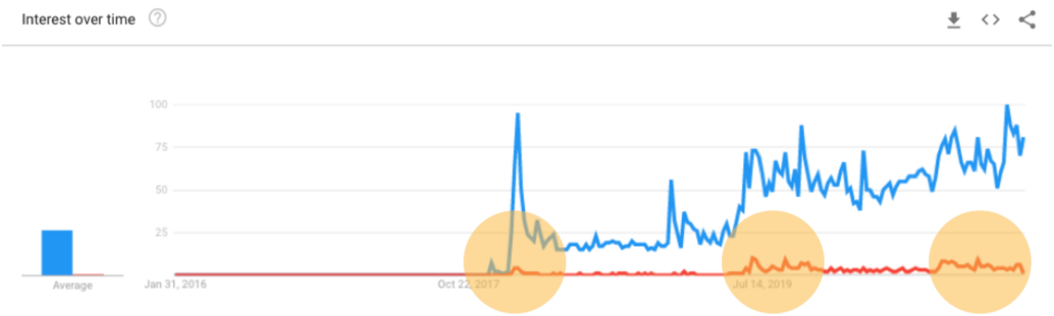


Figure 4. Searches for “deepfake” and “deepfake app” (2016 – 2021). Source: Google Trends

These trends come as little surprise considering that the creation of audiovisual deepfakes using trained models began with hobbyists sharing their creations and engaging with their audience online, with many forums devoted to demoing and sharing the source code and the end products (Cole, 2018; Quilty-Harper, 2021). Deepfake content creators as networked individuals within a community have direct effects on audience belief. This community and its audience are some of the biggest factors in the scaling and democratisation of deepfake technology. The origins of deepfake content emerging out of forum-based communities such as Reddit, provide a formative subculture with its own in-jokes and community figureheads (Larson, Kaleda and Fenstermacher, 2019). These bestow a contextual meaning which is integral to how deepfake technologies should be read in the wild, often being produced with an edge of irony. The role of satire and its ability to be an “alternate configuration” (Taylor, 2020) of political participation has a genealogy which extends as far back more than a decade prior to the popularisation of deepfake content (Reilly, 2012). Subsequently, the “reflexive securitisation” (Taylor, 2020) of deepfake technologies has overlooked key cultural nous, such as the role of sharing discourse and satirical humour in spreading deepfake content and co-constructing understanding of such content. And so, Reddit, and other social media platforms like it, have come to be seen as environments more akin to a “wild west” (Calderone, 2021) with their tendency to feature alternative subcultural norms and the comparatively loose moderation of forums being taken as signalling a threat to democratic rule (Lukito, 2020). The notion of satire as a form of political participation has been somewhat forgotten within current high-level policy-making around deepfake technology, but nonetheless remains a salient part of audiovisual contents’, and now, deepfake contents’, production and dissemination (Ballard, 2016).

But even from this brief exposition, it is evident that there is not only value to be found in examining the content of and context in which deepfakes propagate and spread, but that there is value in scrutinising the public reactions to these. The feedback loop between creators of user-generated content and their audience is worthy of study here given that understanding reception to user-generated deepfake content is needed to fully acknowledge how deepfake content is treated by the group most likely to disseminate it (Chambers and Bichard, 2012; Lee et al., 2021). There is also evidence to suggest that emotive audience feedback is key to understanding malicious deployment of deepfakes, with behavioural factors being a significant arbiter of how quickly disinformation spreads (Kwok and Goh, 2020). With the sentiments for even volatile deepfake content being captured in real-time through commentary under YouTube videos with sensitivity (Boyd, 2014b).

Social media platforms act as participatory sites for exchange and discourse fostered between users and creators; this forms an important part of the approach for delving into the public perceptions of deepfake content. Public commentary is a key component for unpacking the state of play and forthcoming trends in this area. The focus of this research thus homes in tightly on the element of public perception to understand the implications of deepfake disinformation's reach as it scales up. The approach offers a holistic advance on existing mechanisms of detection and education rather than deferring to moralising tendencies which stem from hasty policy-making. Such responses often resort to outright censorship of deepfake content rather than measured and reasonable regulation.

2.2 Deepfake Technologies: The Limits and Potential Considerations of GANs

Having examined the environment which contextualises deepfake deployment, it is also necessary to scrutinise the technical dimensions of audiovisual AI and the foundational issues at the core of the debates on deepfake deployment. The following chapter unpacks these foundational issues to specifically look at the technical aspects of AI technologies in tandem with the theoretical mechanics of disinformation. It is of key importance to highlight that both these technical and theoretical considerations are bound by human biases at their core. These elements prime the analysis in the third chapter which discusses the critical security implications of deepfake deployment in the wild in relation to public perception and discourse.

2.2.1 Out of the Wild and Into the Field: Technical Considerations of GANs, DL and AI/ML Development

This section has surveyed a selection of the technical limitations at the heart of DL technologies and how these lead to real-world security implications. It has introduced themes of safety in greater depth, with detection as a particularly key topic for discussion (Barrett and Baum, 2016).[6] Though deepfake technology's framing and its contexts of deployment have been discussed in relation to wider political discourse, these are rarely reconciled with the more technical aspects of the technology (Brooks, 2021). Arguably a richer assessment of deepfake technology's potency and societally disruptive impacts requires an acknowledgement of the technology's technical vulnerabilities in concert with its audience reception. So, just as the shallowness of post-truth framing is dissected in earlier sections, the shallowness of deepfake technology's technical development must also be explored to fully do justice to this issue in the round and establish an approach to this topic so that it might be breached constructively in the long-term.

AI's major technical limitations stem largely from the epistemological design of ML systems, which have posed fundamental issues for state-of-the-art (SOTA) DL models. One of these fundamental issues is the legacy left by structural anthropomorphism (Špelda and Stritecky, 2021: 89) which sees the latest DL models designed from the same reference points as human cognition. These reference points see world models bear the same foundational fault lines as their originators (LeCun et al., 2015; Schmidhuber 2015; Song et al., 2020).[7] Deepfakes are particularly symptomatic of the structural anthropomorphism which pervades AI systems. This is especially when we consider the extent to

[6] There are several key safety concepts which have been identified as prominent contemporary issues here, but for the scope of this research it has been necessary to focus on a single issue. For a richer discussion of the topic of AI safety, see the three part series by Rudner and Toner (2021ab).

[7] Adjacent structural issues problematise the field of AI, with its lack of rigorous peer review process, varied regulatory status and general expansiveness as a field. These complement existing issues of standardisation, the reproducibility of projects, data and models and in some instances, this sees somewhat arbitrary development of ML tools occur. The field of AI also suffers somewhat from generally over-optimistic outlooks which consider layers of abstractions without resolution. For more on the trending issues in ML scholarship see the discussion by Lipton and Steinhardt (2018).

which their neural networks are modelled on an imperfect set of benchmarks and limitations dictated by instilled researcher bias. The limitations of computing hardware are somewhat synchronous to the major limitations of human cognition; a finding which is well-supported by evidence from neuroscience (Geirhos et al., 2020; Kaur, Kumar and Kumaraguru, 2020; Marshall, 1977). Of these limiting factors, bias is a particularly salient issue in contemporary AI/ML systems and its effects are evident from the outset of training through to the endpoint of deployment. This is a process which is showcased in the technical shallowness of deepfake technologies (Plebe and Grasso, 2019: 516). But given the plethora of GAN architectures which currently exist, the scope here focuses on the widely embedded principles of these models rather than the architectural specificity of particular algorithms.

In essence, the general principle behind GANs sees two neural networks challenging each other for greater representational accuracy by discriminating against certain values until a victor is found (Bengio et al., 2006; LeCun et al., 2015). The zero sum games played out between these two models were originally designed as a method of using competitive coevolution to teach machines to attain more robust results but this has resulted in a number of “pyrrhic wins” (Prabhu and Birhane, 2020) for models which often generate issues which are even more difficult to correct in real-world settings, akin to the Adam’s Left Rib Conundrum. DL models’ corresponding shortcomings mean they are generally difficult to maintain control over outside of lab environments, given that even incremental changes make a significant impact on the behaviour of AI models’ outputs (Gilmer et al. 2018). Further to this, AI is poor at determining complex situations (Hall, 2018: 70) and there is already significant precedent to demonstrate that even in the training phases and iterative stages of deepfake development, neural networks are vulnerable to corruption (Korshunov and Marcel, 2019). The wider real-world implications of these deployments are thus subject to a number of environmental sensitivities. In some instances, this facilitates the work of malicious actors aiming to exploit the vulnerabilities of a complex political system (Tolosana et al., 2020). Some of the direct consequences arising from the release of SOTA DL models in the wild are mentioned in previous sections, but even indirect components of these models, such as the data used to train GANs, can pose a critical vulnerability to real-world security (Afchar et al., 2018; Li et al., 2020; Zi et al., 2020) for being easily exploitable by malicious actors.

As mentioned, GANs share many of the same issues as text models. These become particularly salient in a security context when we account for perturbations which are carried over in the shift from RNN, an older architecture which processes less complex data, such as text, which often comes as single layers, towards CNNs, which process more complex data across “multiple arrays" such as images and audio content (Li, 2018: 10). However, despite the latter’s greater complexity, being made up of both pooling and fully connected layers, CNNs can still be confounded through counterfactual data and are susceptible to the manipulation of training data in the form of data “poisoning” (Koh, Steinhardt and Liang, 2018). These neural networks are easy to exploit from the outset of their life cycles, with wide-ranging issues spanning misclassification and model transferability (Amerini et al., 2019; Zhang and Liu, 2020), to a lack of alignment stemming from largely invariant scenarios used at training stages for CNNs, which ultimately maintain a narrow environmental conception against vulnerabilities (Suratkar et al., 2020). For instance, the latest iteration of DL models in styleGANs, a novel generative adversarial network from Nvidia researchers, have limited schematic variability and cannot replicate backgrounds or artefacts or infer environmental hints (Karras, Laine and Aila, 2019) (Figure 5). Given that adversarial entities and malicious actors seeking to exploit AI’s vulnerable points will usually look to maximise outputs at the same time as minimising costs in real-world environments, these are some of the weaknesses which are most susceptible to exploitation in real-world scenarios. Deepfake technologies, being at a relatively nascent point in their development, have capabilities which are easily perturbed even at a shallow level. The inherent shallowness of deepfake technologies is a limitation which is likely to be amplified in the course of greater democratisation and accessibility of this technology. The role of the public is likely to be a significant one in time, but how they receive and in turn disseminate this is less clear.

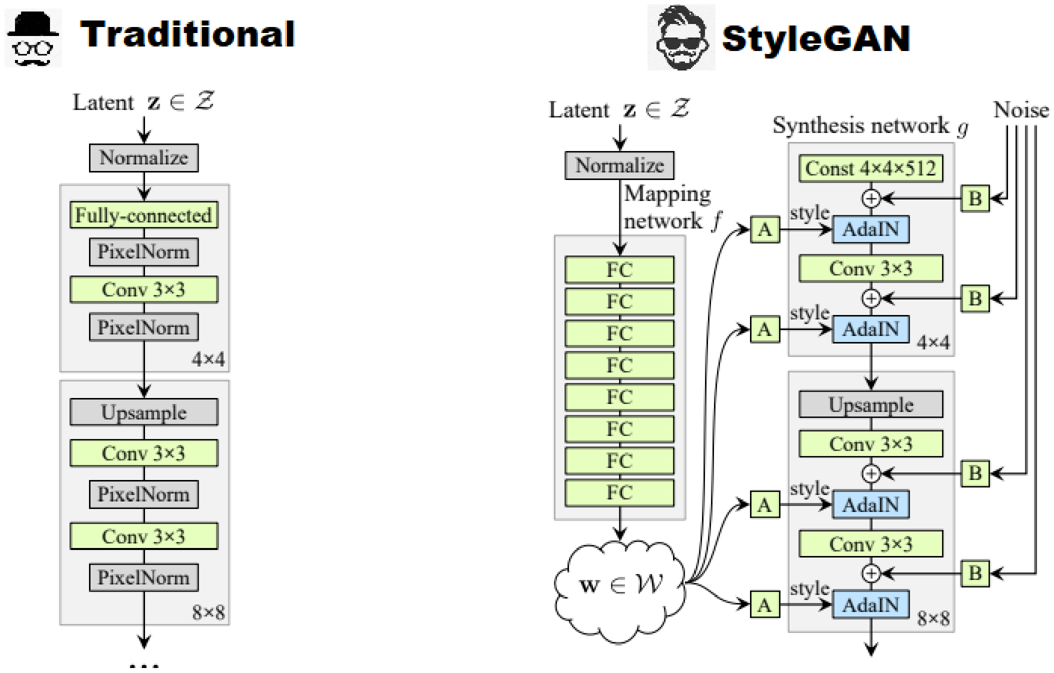


Figure 5. The architecture for styleGAN, a novel generative adversarial network from Nvidia versus the traditional model (2019). Source: Heartbeat.fritz.ai

The interaction between models and human agents is arguably the most potent dynamic to examine here, with public perception playing an understated role here. Advances in the technical sophistication of deepfake technologies are presently seen as the key concern rather than ‘softer’ considerations such as public education and the epistemological benchmarks that come with these. Even belonging to a narrow classification of AI, the political discussions around deepfakes have tended to defer to escalatory narratives and anticipatory ethics more often than not. These are largely biased towards worst-case scenarios, spread by proponents on either extremes of over-optimism about the technical capabilities of AI/ML and SOTA DL technologies (Urbani, 2020) and techno-pessimism concerning actors both bad and incidental (Goodfellow, McDaniel and Papernot, 2018). In both these instances, the tendency to over anthropomorphise AI is a core contention which affects the robustness of DL deployment. The hypothetical nature of these issues is similar to the debates outside of the remit of narrow AI, which illustrates these same issues with democratisation of technology, in particular the scale and speed of its spread. Artificial general and superintelligence (AGI/SI) highlight AGI/SI’s role in spreading misinformation at an exponentially accelerated level, similarly deferring to the worst-case scenario trope (Baum, 2018a: 8). These debates are also directly correlated with the assumptions emerging from many contemporary commentators across journalism and policy-making, but are often anecdotal and lacking an empirical baseline.

The foundational issues which constitute deepfake technology’s technical shallowness in this section are reconciled here with the flimsiness of media trust evidenced in the previous section. It is important to keep in mind that the nature of these aforementioned assessments are for the most part, thought experiments which tend towards in speculation; something which applies to a number of the foundational debates within AI, a field which is somewhat shaky in its core assumptions and sometimes lacking in interpretability (Kasirzadeh and Smart, 2021). The judgements made of deepfake technology lose a number of sensitivities in the rush to foreground deepfake technology as an imminent security issue

(Ignatidou et al., 2019). And while the issue of disinformation at large is a troubling one, the place of deepfake technology within security discourse is fairly siloed, narrowly framed and moreover, lacking a certain balance or proportionality in its assessments (Brooks, 2021). A more balanced argument would suggest that the potency of deepfake content cannot be fully assessed without considering their context for use and by whom. As such, ground-level discourse can help patch this gap.

2.2.2 Theoretical Considerations: Robustness and Detection as Security Issues

Overall, the following section reconciles the aforementioned technical and foundational issues of deepfake technology with the theoretical concerns of the complex real-world security environment. In particular, this section deepens the inquiry by mapping out how these technical issues intersect with “post-truth” media and how this framing sees deepfake technologies implicated as a new tool for disinformation in the public realm. This bridges the gap between AI safety to real-world issues of security.

Having established the major epistemological fallacies within AI’s technical foundations as well as the limitations of taking post-truth framing at face value, it seems evident that close scrutiny of robustness is critical for understanding the nuances of the research questions here, with robustness emerging as a key conceptual driver behind the technical and theoretical shortcomings of deepfake technologies. The shallow end product which results from deepfake technology, as a form of narrow AI, has limited representational capabilities, with many models requiring an appropriately skilled actor to create a widely-convincing impression (Min-Yeong, 2020). This serves to illustrate both the aforementioned issues of structural anthropomorphism in training and in response. These tokenistic appeals to reality serve to highlight the issues of definitional robustness within detection issues in the wild. Issues of definitional robustness run deep, stemming from the epistemological roots of AI and see the detection of deepfake technologies refer largely to the targeting of algorithmically-trained models; that is, deepfakes rather than models needing minimal data, such as shallowfakes (Arik et al., 2018; Simonite, 2020). The proliferation of both deep and shallowfake content in the wild pushes the salience of questions on how to speed up detection at scale, especially when we consider the fast-growing numbers of consumer applications which allow deepfakes to proliferate (Schulz, 2020; Vincent, 2019). But the detection issue becomes a cyclical one, plagued with aptly adversarial dynamics when one considers that deepfake detection tools must be trained with large and diverse data sets to reliably detect deepfakes (Li et al., 2020; Zi et al., 2020). However, the pervasiveness of deep and shallowfake content's proliferation means that current datasets are not sufficient in themselves to withstand the volume of content developed for both innocuous as well as malicious purposes (Tong et al., 2020). But equally, the popularity and growing sophistication of apps for producing such content has also seen a spectrum of less-conventional detection methods continue to adapt accordingly (Gilmer, 2019).

In theory, generative models can be counter-engineered for detection purposes (Cox, Slapakova and Marcellino, 2020). This propensity to pit similar entities against one another in an Adam's Left Rib Conundrum has been a successful approach in previous instances where less complex neural networks such as fake news generators have been used for detection (Botha and Pieterse, 2020). But given that fast-changing developments in technology have outmoded even comprehensive surveys of generative models (Tolosana et al., 2020; Yadav and Salmani, 2019).[8] This "cat and mouse" dynamic of innovation sees the techniques used to

[8] Corporate purchase of deepfake technology from market leaders such as Ancestry confound the system even more, with a sudden moral question of "good" deepfakes for bereavement/psychological support/grievance and tourism. These are also troubling for their novelty and hence, virality which creates a self fulfilling prophecy of the spread of deepfake content.

identify deepfakes often used to pave the way for more sophisticated deepfake techniques (Lyu, 2020). As such, detection tools must be constantly updated with data of increasing sophistication to ensure that they continue to be effective at detecting manipulated media, which is costly and laborious as well as being inefficient (Shu et al., 2020). Alternative methods such as the use of hardware signatures and digital imprinting have been suggested, however these pose their own difficulties being hard to scale in many instances (Katarya and Lal, 2020). This causes issues with a lack of standardisation to resurface. But ultimately deepfake technology is privy to an ever-mercurial set of developments, especially when seen in context of the wider arena of disinformation. And like other emerging technologies in the complex modern security environment which are also affected by minute shifts in user behaviour, this requires progress to be made in a process of continuous innovation rather than developing over-reliance on standardised principles and practices.

Despite the emphasis placed on detection in issues of real-world security, the limited scope of detection efforts so far often overlook epistemic issues at the core of deepfake technologies and their potential usage as disinformation (Agarwal and Varshney, 2019). Traditional forensic methods demonstrate this limited scope by tending towards technical elements of algorithmically trained content, examining surface-level variances of deepfake content, being either signal based (with formats ranging across JPEG, CFA, PRNU) or physics based (based around qualities such as lighting, shadow, reflection) (Li, Chang and Lyu, 2018). However the growing taxonomy of GANs in the wild and the wide-ranging disparities between the source codes these are sired from (Basu, 2019) continues to complicate the issue. The heterogeneity of DL models in combination with the complexity of the environments these models are deployed in means that detecting such content before it spreads is difficult. The origins of deepfake content is not obviously attributable in most instances, especially when the source code is shared across both mainstream platforms, where these models may be used for research as well as less moderated forum spaces, where use veers towards possible criminal intent (Ellis, 2018).

Evidently, detection’s issues are not relegated to technical roadblocks alone. Human agents create additional frictions in the field, being at once a root cause of deepfake disinformation’s spread and a means by which to help resolve real-world security issues. Detection in blind settings is notoriously difficult (Pu et al., 2020); a limitation which researchers have attempted to mitigate through using manual or non-algorithmic detection methods, often mediated by or based entirely on the judgements of human moderators (Korshunov and Marcel, 2020). However, there are obvious limitations to detecting adversarial examples using either human agents or ML-based methods. These are largely fraught with subjectivity; observations which stem from human cognition such as optical bias are transposed to machines. Further, imperceptible manipulations of models such as alterations in pixel intensity, cannot be observed easily by human moderators in the first instance and are consequently still difficult to detect using ML-based methods (Khodabakhsh and Busch, 2020). As we have established, these are weaknesses which are ripe for exploitation by malicious actors. This becomes a cyclical problem especially when we consider existing issues with flawed training data and its interactions with unsupervised learning from the outset of model development, in which data is increasingly defined by the DL models themselves (Karras, Laine and Aila, 2019). This creates a self-fulfilling prophecy in which the model attempts to solve a different task to the one it was originally designed to solve, causing further problems for interpretability. As such, conventional detection methods are presently inadequate and mired in a spectrum of biases rooted in AI’s foundational issues. These conventional detection methods alone may not be the best solution for mitigating the potential effects of AI-generated fake videos in the long-term, and so alternative methods which can be viably used in combination with these should be sought out.

2.3 Dialectics of Public Discourse

The following section further contextualises the avenues of discussion above, looking specifically at the role of the audience and the importance of public discourse in relation to the technical and theoretical perspectives on deepfake technology. This section examines why perception is a meaningful area of research in relation to deepfakes, especially in terms of measuring the risk posed by this technology.

2.3.1 To Err is Human: Bias as a Distinct Factor

Having examined the limitations of the current disinformation environment contextualising this technology, as well as having addressed the lack of definitional robustness in the wider field, it is relevant to discuss the frailties of human agents. Human agents are of critical importance to this discussion, being both a structural reference point for AI as well as an arbiter of deepfake disinformation’s spread.

Human cognition as the basis for SOTA DL applications sees anthropomorphic qualities, such as bias, being built into AI systems at every stage of its development, from training (Pan et al., 2020) to benchmarking (Redondo and Gibert, 2020). Out of this bias comes further factors affecting real-world security contexts including, but not limited to; belief (Pennycook, Cannon and Rand, 2018), moderation (Gillespie, 2020; Veletsianos et al., 2018), the personality traits of audience and users (Eidelman et al., 2012; Wolverton and Stevens, 2020) and the role of personal experiences (Kubin et al., 2021) in what is fundamentally an actor driven phenomenon. As mentioned, the creation and spread of deepfake content forms a feedback loop between itself and its audience. In this, we find our own biases replayed; a factor which plays into the bigger dynamic of witnessing the weaknesses of DL architectures becoming its adversarial foil.

But what is at stake in the worst case ground-level scenario is that these individual biases scale to become societal biases (Erzikova and McLean, 2020) and at a more advanced scale of amplification, this could see the development of what Aviv Ovadya calls “reality apathy” in which the democratic contracts upheld by baseline beliefs are increasingly destabilised and apathy becomes further adopted as a modus operandi in the face of overwhelming helplessness. This deference to low effort thinking and desire to yield to untruth (Warzel, 2018) in the form of alternative facts and conspiratorial thinking is seen by some as the easiest method of coping with the evolving challenges of deepfake technologies (Leiser, Duani and Wagner-Egger, 2017). There is some precedence for these concerns in lay audiences’ tendencies to be biased towards engaging with more emotive content (Kadir, Lokman and Muhammad, 2018b). As such, public perception is central to understanding how deepfake disinformation spreads in the wild. To return briefly to the notion of structural anthropomorphism, these cognitive shortcuts in their most advanced stages are similar to processes which occur during model training. Shortcut learning also sees models demonstrate a bias towards superficially effective means of goal fulfillment (Geirhos et al., 2020), giving some indication as to how cognitive bias could be mobilised to counter malicious uses of deepfake technology.

Given the predisposition towards emotional content, it comes as little surprise that audience reactions should tend towards openly expressing more emotive sentiments. However, this emotional aspect poses vulnerabilities in helping to scale and speed up disinformation’s spread. Alongside well-established correlations between virality and emotional content (Brady, Gantman and Bavel, 2019) there is also evidence to suggest that audiovisual content demonstrates high levels of emotive response (Kadir, Lokman and Muhammad, 2018a), which can lead to the spread of disinformation. Despite the obvious importance of public perception in assessing the actor-driven spread of audiovisual disinformation, this has been a historically under-researched area. Indicators for how important human agents are in spreading disinformation are hinted at in existing studies on disinformation and audience bias. These studies demonstrate that participants are susceptible to falsified videos and fail to distinguish between two different types of fake videos when a biometric reference point for mannerisms or speech patterns is not available (Khodabakhsh and Busch, 2020). This susceptibility to misreading fake content is supplemented by similar studies, showing that audiences tend to exhibit “continued influence effect”; a dissonance between relying on inaccurate information even after a credible correction has been presented. And moreover are subject to “Illusory truth effect” which sees audiences report that more frequently repeated information is truer than novel information on account of its familiarity (Britt et al., 2019). These biases in action are further complicated by additional considerations. One of these considerations is that it is difficult to reconstruct a memory accurately after representations of complex ideas from multiple sources (Larson, Kaleda and Fenstermacher, 2019) and that those who spread disinformation can do so inadvertently by simply misconceiving facts (Starbird, 2019). As such, the case of attribution of malicious actors in disinformation’s spread remains unclear in some instances, making public perception a crucial complement to conventional mitigation methods and an apt baseline for benchmarking threat levels. Evidently, the concerns about deepfake technology are not limited to purely technical or theoretical roadblocks alone and should be considered in light of more sociotechnical explanations. More unified efforts across social and technical specialisms towards the detection of audiovisual disinformation are needed to understand the issue fully and its relation to the bigger picture of disinformation.

2.3.2 Commentary as a Critical Indicator of Perception

Studying the role of public perceptions (Shah et al., 2015) is crucial to filling a research gap which has been largely overlooked in favour of expounding the risks and regulation of deepfakes ahead of ground-level sentiment. Studies on public perception at scale are somewhat sparse in the field of deepfake content, and have been executed largely through formal opinion polls and surveys (Pennycook and Rand, 2020; Tandoc, Lim and Ling, 2020; Wolverton and Stevens, 2020) rather than adopting the immediacy and organic quality of publicly accessible commentary from avenues such as forums which offer a raw but unstructured arena public discourse.

Given how pivotal emotional sentiment has been to the speed and scale of its dissemination, organic public commentary serves as one of the most appropriate avenues to study deepfake technology (Lee, 2019; Martel, Pennycook and Rand, 2019; Mulholland et al., 2016; Vosoughi, Roy and Aral, 2018). And as the most likely deliberate and inadvertent spreaders of deepfake content, laypeople and their perceptions are key to studying disinformation’s spread (Ahmed, 2020; Amodei et al., 2016). In effect, disinformation’s broader scope can be mapped out using ground-level discourse, specifically through user commentary. User commentary allows for organic insights into public perception and an empirical examination of how susceptible the public is to disinformation at scale. This offers further insights into just how justified the concerns about high-level political disruption are. Unpacking the public perceptions of deepfake technology can contribute meaningfully to the discussion as an emerging area of interest and target key blindspots in the current canon of AI safety scholarship.

PART TWO

"Thought achieves more in the world than practice; for, once the realm of imagination has been revolutionized, reality cannot resist."

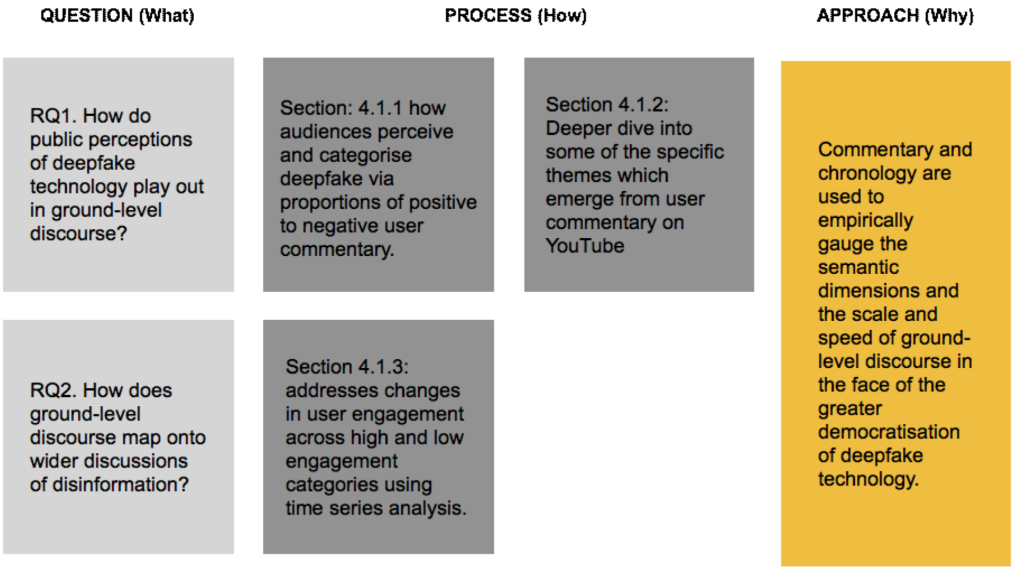
– Georg Wilhelm Friedrich Hegel (1896: 20)

3. Methodology and Research Design

Borrowing from previous approaches used across digital spaces ranging from YouTube (Lee et al., 2021) to online news headlines (Westerlund, 2019), which make use of computationally-assisted semantic analysis, this research model applies and advances the most relevant learnings to examine public commentary deepfake content on YouTube. This study of discourse around audiovisual deepfake content responds to the current scholarship around disinformation, particularly safety research in the AI/ML subfield which has been preoccupied with providing a broad survey of technical safety issues for lab-based environments (Isakov et al., 2020; Mirsky and Lee, 2020; Nguyen and Vu., 2019; Tolosana et al., 2020).

This research takes the form of an inductive inquiry, drawing its samples from user-generated comments on a curated selection of publicly hosted videos on YouTube, the video hosting platform-cum-social network. Given the relative sparseness of empirical research on perceptions of deepfake technology, this research seeks to establish a baseline for how human agents, specifically lay audiences, assess and interact with this emerging technology, which feeds into a broader understanding of the core issues at the heart of disinformation and its spread. Public perception of deepfake content and lay discourse are proportionally underserved areas in the burgeoning canon of deepfake scholarship; this is despite laypeople being the most likely disseminators of disinformation (Ștefăniță, Corbu and Buturoiu, 2018; Tandoc, Lim and Ling, 2020). This research unifies these two factors at a pivotal moment for the greater democratisation of deepfake technology.

Figure 6 below outlines the processes and approach adopted in answering the research questions at stake. The section that follows also details the contextual data collection process, as well as the rationale behind the choice of timeframe, platform and how the resulting sample of labelled political deepfake videos was selected for this research.



QUESTION (What)	PROCESS (How)		APPROACH (Why)
RQ1. How do public perceptions of deepfake technology play out in ground-level discourse?	Section: 4.1.1 how audiences perceive and categorise deepfake via proportions of positive to negative user commentary.	Section 4.1.2: Deeper dive into some of the specific themes which emerge from user commentary on YouTube	Commentary and chronology are used to empirically gauge the semantic dimensions and the scale and speed of ground-level discourse in the face of the greater democratisation of deepfake technology.
RQ2. How does ground-level discourse map onto wider discussions of disinformation?	Section 4.1.3: addresses changes in user engagement across high and low engagement categories using time series analysis.		

Figure 6 Outline of the research model's processes and approach

3.1 Contextual Background

3.1.1 YouTube Comments and Critical Discourse Analysis

Current projections about the effects of deepfake technology are somewhat speculative, being largely transposed from wider debates in AI/ML such as AGI/SI (Baum, 2018b; Everitt, Lea and Hutter, 2018; Castillo, Guarda and Alenda, 2020; Vaccari and Chadwick, 2020;

Yampolskiy and Fox, 2013) or being otherwise subject to broad surveys rather than affording the topic in-depth treatment (Ahmed, 2020; Yadav and Salmani, 2019). This skews policy towards a somewhat elitist and technocratic position which overlooks the actual voices of the people it seeks to protect. As such, established methods of discourse analysis are used here to lend a more structurally sound angle to the wider discussion, unpacking the critical power relations between how audiences relate to deepfakes and how this feeds into disinformation. This taps into the distinct cultural nous around this fast-moving subject and helps reassess the technology in light of the demographic most likely to encounter and disseminate the content created by it (Dupuis and Williams, 2019; Vosoughi, Roy and Aral, 2018).

Discourse analysis' sub-branch, critical discourse analysis (CDA), is methodologically important, being used here to infer trends and patterns across a variety of platforms, networks and databases in a systematic way (Wahl-Jorgensen and Carlson, 2021; Wodak, 2003). In the context of this research, CDA offers a more constructive approach for analysing the unstructured data provided by user commentary which is often messy and fragmentary. CDA's view of discourse as a form of social practice is particularly useful for its preoccupation with how discourse is implicated in power relations (van Dijk, 1993). The scope of CDA also accounts for the social conditions in which discourse is created (Fairclough, 2003: 91). In this way, social interactions and their resulting texts can be seen as a barometer for their respective environments, offering an interpretation of the social system it operates within as well as embodying it (Fairclough, 2003: 91). Discourse as a form of social practice can be subsequently embodied by a wealth of content on the platform it is hosted on, ranging from user-created videos to commentary on this content (Robertson et al., 2013).

CDA offers a critical approach to understanding ideology and how discourse is implicated in power relations, in which the role of text, user commentary in this instance, is an integral part of a sociotechnically entangled feedback loop (Chambers and Bichard, 2012; Dabas et al., 2019); a set of entanglements made up of content, creators and audiences. These sociotechnically entangled components are social processes between audiences and producers and form a co-constructed set of public perceptions around deepfake technology. Social media platforms such as YouTube offer a breeding ground for opinions, being co-constructed by a complex web of users' choices and actions (Boyd, 2014a; Dubovi and Tabak, 2020). There are similar patterns which occur across several platforms, in which these co-constructed sociotechnical entanglements map onto more formal disinformation campaigns (Boyd, 2014b). These networks see misinformation spread wittingly or otherwise by the same networks of users (Mericle, 2020; Pantserev, 2020).

This dynamic is essential for understanding disinformation, which is itself a feedback loop that is similarly co-constructed, continuously entangled and full of interdependent processes in this "realisation" of deepfake deployment in the wild (Heydari et al., 2019; Hussain et al., 2018). Commentary as a unit of measurement here is socially shaped as well as socially shaping. This returns to the themes explored in the first part of this research, on public sensemaking around post-truth, user bias and the Adam's Left Rib Conundrum. And while commentary is a powerful factor in reinforcing belief (Heydari et al., 2019), lay voices and the opinions exhibited in discourse cannot be easily broken down into discrete units of measurement. This is particularly given how relations of participants and their roles in producing discourse are not always equal (Momeni and Sageder, 2013). In this context commentary found on YouTube videos may lead to some disparity of opinion.

Social networks, YouTube in particular, as platforms which host as well as construct discourse, exemplifies this set of inequalities well. The networked audience exhibits a range of comprehension in response to content, with resulting exchanges spanning a wealth of cultural responses (Fairclough, 2003). For instance, the disparity between groups more familiar with the subcultural nous of video hosting platforms and comment boards are more readily able to read and access deepfake content (Ahmed, 2020). As such, networked discourse can provide “patterns of experience” (Halliday, 1985: 106) which in turn can map directly onto patterns of ideologies. and help us better understand how sensemaking is co-constructed.

To make sense of these complex feedback loops and answer the research questions at stake, the approach here adopts the analytical structure of describing, interpreting and explaining following Fairclough's (1995) model for CDA. This is appropriate for closely examining the entangled, sociotechnically co-constructed patterns of both ideologies and experience generated by user commentary. Further to this, the application of CDA is a method which allows the distinct nuances of the social interactions which constitute the wider subculture of the network to be drawn out (Mejova and Srinivasan, 2012). As a distinct difference, CDA’s approach takes the hallmarks of discourse analysis and navigates the gaps between discrete data to pit value judgments against discrete information, through coding, which ensures the validity and replicability, and sampling, through selecting specific clips within a set time frame (Bell, 2005: 138-37). But even as a non-invasive method which is automated by computer-aided text and visual analysis here, CDA is not totally without fault. The application of CDA in this study of user comments may also result in systemic bias throughout the process of collecting and analysing information. But while these entanglements are challenging to explore, they also yield compelling insights into alternative behavioural patterns which may have not been considered in the realm of security.

Though there are various ways of exploring public perception, user commentary on audiovisual content is an appropriate behavioural barometer for studying deepfake technology, especially as a nascent phenomenon. The use of commentary for inferring trends and patterns across data has well-established precedents in the social sciences. This is especially in relation to studies on YouTube and its subcultural nous as a platform (Amarasekara and Grant, 2019; Erzikova and McLean, 2020). Comments can be used to reveal matters of perception via particular pieces of video content and so, rather than merely being an end user "gimmick" they are heavyweight enough for semantic analysis (Schultes, Dorner and Lehner, 2013: 661). When treated as a form of critical discourse, YouTube commentary reveals insights into how users perceive labelled fake content. User commentary has also been the basis of previous studies to infer political sentiments (Mulholland et al., 2016) and trustworthiness in political figures (Heydari et al., 2019).

Commentary has also been implemented in broader studies on forms of political communication; these amply demonstrated how YouTube can be an apt basis for comprehending political commentary (Chambers and Bichard, 2012). This is of particular interest given the shift in the platform's patterns of usership from "passive consumption" to generating the "objects of social exchange" (Schultes, Dorner and Lehner, 2013: 659); that is, rather than merely observing user-generated content, users increasingly initiate discursive reactions. Underrepresented areas of public discourse, such as those found informally and organically on social media (Bicquelet, 2017) can feed into a greater contextual understanding of disinformation. And YouTube has provided a fertile site for this, offering insights into patterns of online behaviour, particularly around political messaging and usership (Kolotaev and Kollnig, 2020).

3.1.2 YouTube: Social Network as Wild Environment

YouTube, as a mainstream social network, is a popular site of cultural and political exchange; making it an apt if not ideal space to study perceptions of deepfake content in the wild. This aptness comes partly from hosting a wide variety of audiovisual content (Gillespie, 2020). A factor which draws engagement from broad audiences in substantial quantities. This also facilitates the collection of user commentary at scale, out of everyday, organic public encounters. Further, the platform's technical traits, such as the ease of accessing video comments through YouTube's Data API v2.0, as well as its reasonable terms of service (ToS) for researchers and the sheer volume of information hosted on the platform make it the most compelling candidate for studying the phenomenon of labelled deepfakes in the wild. This narrows down the search for videos by matching key criteria and metadata to make the most of publicly available resources. For instance, patterns of user behaviours can be inferred from the data rich pool that YouTube offers across large volumes of commentary and allows correlations between content and discourse to be unpacked as they relate to a sprawling phenomenon such as disinformation. YouTube's status as a popular video sharing platform facilitates real-time discussion between producers and audiences and in doing so, provides an unparalleled public forum for opinion-rich discussions (Möller et al., 2019). This mine of commentary across a broad spectrum of content is key to helping establish a firmer baseline for disinformation and its leakage onto other platforms (Halpern and Gibbs, 2013). As such, user comments as the key unit of measurement in this research allow for the analysis of both direct and less direct expressions of sentiments about deepfake technology.

However, deepfake content is moderated under still-changing laws – this swathe of issues adds to the slipperiness of researching the underbelly of deepfake technologies with any certainty or replicability (Villasenor, 2019a). And though there are other sites which could be examined in this instance, YouTube is by far the richest platform to ground our understanding of deepfake proliferation, providing a trove of data which unveils the socially-shaped structures underlying networks of users. Being an open forum for discussion, YouTube offers insights across the spectrum of proliferation, discussion and dissemination of content (Chambers and Bichard, 2012; Halpern and Gibbs, 2013; Heydari et al., 2019). This is especially in regards to the platform's status as a popular hub for hobbyists of deepfake videos, with YouTube hosting the largest amount of deepfake content on the surface web besides closed discussion boards and forum-based platforms, such as Reddit (Lewis and Nelson, 2019). However, these closed forums are more difficult to access, it is also more difficult to interpret at a subcultural level, as well as being more ephemeral (Hagen et al., 2020; Papadamou et al., 2020).[9] YouTube has been chosen over other alternatives as a platform which allows for sufficient data quality at scale, exhibiting high user engagement as well as a high density of deepfake content.[10]

In a notable recent example of how YouTube has been implemented in research about the novel area of deepfake technologies, YoungAh Lee et al.'s (2021) study has pointed to the importance of YouTube's meta-frames in influencing audience perceptions through a study of audience comments (n= 2689) on the platform's top ten labelled deepfake videos.

[9] But besides deepfake content favouring closed forums and non-surface web spaces which are less easy to penetrate, concerns around ethics and a lack of resources pose a practical barrier to the research here, which focuses strictly on open sources from end to end.

[10] It should be additionally noted that some of the most popular Reddit videos are disseminated on YouTube, indicating some porosity across platforms (Ajder et al., 2019a). And so, given that deepfake content is hosted across other video platforms such as Vimeo and social networks like Instagram and TikTok, it is harder to fully capture direct, causal links between content and commentary to glean an accurate representation of the phenomenon. And though Twitter is often seen as the dominant tool for drawing out how users and their commentary are networked, owing partly to similar aforementioned technical traits, YouTube has been aptly utilised in previous research on the perception of deepfake technology.

But as well as being a site for studying nascent and emerging trends, YouTube also showcases the contours of new political behaviours. The platform's burgeoning status as a space for expressing political speech and opinion has advanced alongside a trend towards microtargeting of commercial deepfake content (Dobber et al., 2021) and personalised content (Borgesius et al., 2018; Rymes, 2012) more generally. The latter of these has been marked out as having particularly potent ramifications for disinformation. Assessing YouTube’s dynamics as a social platform can shed light on how particular localised tenets of disinformation are spread. However, it should be acknowledged that YouTube’s content only provides a partial examination of the whole environment. The particular cultural and semantic nous which make up these dynamic online communities makes the whole difficult to interpret and assess with substantial applicability to every context (Hall, 2018: 73), but ultimately the design of this research aims to provide a real-time snapshot of the most prominent sites of organic discourse about deepfake content.

3.2 Sampling and Data Collection Process

Current safety efforts are directed largely towards the study of algorithmically trained audiovisual deepfake content rather than audio deepfakes, cheap or shallowfakes, which can be easily generated without technical expertise (Paris and Donovan, 2019). As such, understanding the public reaction to deepfake content is critical for supporting these safety efforts and gauging a baseline response to audiovisual disinformation. This is particularly important as access to the software and other tools to create deepfake content becomes increasingly democratised (Simonite, 2020). Deepfake content is often faced with content takedowns and platform bans which are factors echoed in YouTube’s approach to moderation (Bechmann, 2020; Diakopoulos and Johnson, 2019; Gillespie, 2020). This makes the status of deepfake content’s longevity on mainstream platforms precarious and fraught with ephemerality. The videos selected for this sample have been retained by YouTube without the threat of being similarly disrupted. This provides a stable way to feasibly observe the mid-term effects of an otherwise slippery subject matter. This retention by the platform is partly due to their status as labelled deepfakes, with labelled content often signalling a disclaimer to moderators (Nassetta and Gross, 2020: 5). These often bypass filters for deceptive content within moderation standards (Liu and Wu, 2020). The labelled videos selected here therefore offer a more fixed point to study public perception and its implications for disinformation. The labelled content chosen here features an explicit mention of the term “deepfake" either as part of the video name or descriptor.

The initial sample for this research was derived from a composite of methods used to identify videos in recent research on audiovisual deepfake content (Choi and Segev, 2016; Schultes, Dorner and Lehner, 2013). This involved creating ranked searches for the keyword “deepfake" and variants thereof to find the most appropriate candidates for study. For overall feasibility, the initial searches in this sample were narrowly focused on outputs featuring only safe for work (SFW) content of well-known public figures from English-language videos. Though the curation and manual sampling of the labelled videos chosen here introduces an element of selection bias, the process of manual sampling is a preferable tradeoff when compared to alternative methods which are more sensitive to a constant flux of user engagements on the platform. For instance, manual sampling avoids pitfalls such as relying on YouTube’s hierarchical meta-frames whose ranking categories include likes, dislikes and “most viewed” or “most played” categories, which are highly changeable and dependent on the other material uploaded to the platform (Schultes, Dorner and Lehner, 2013).

This study of user comments draws from a selection of labelled deepfake videos hosted on YouTube, featuring well-known public figures. The videos selected for this study are published between 2018 and 2021; a range which accounts for the term's first use in public and a mass-scale release of deepfake content into the public domain. This sample was chosen partly for its resonance with themes expressed in wider policy-making, featuring high-profile public figures who might reasonably become targeted as part of disinformation campaigns in the wild (Hsu, 2018; Hwang, 2018), with a cross-section of videos curated to examine public perception across high and low engagement content, with data collection finalised in early April 2021. To keep the scope of research focussed, this research's key concerns around the policy and security surrounding disinformation campaigns means that the well-known public figures chosen here err towards figures of national importance, for instance, heads of states rather than celebrities or figures from the entertainment industry. In order of published date on the platform these sources are; former President of the United States, Barack Obama by digital media, news and entertainment company, Buzzfeed from 2018; Russian President, Vladimir Putin by non-profit organisation, RepresentUs from 2019; Supreme Leader of North Korea, Kim Jong-un, from the same organisation in the same year; and the Queen of the United Kingdom by British public-service television network, Channel 4 from 2020.

This research uses indicators similar to studies which stratify audiovisual content beyond thematic classifications and metadata (Choi and Segev, 2016), making use of engagement metrics as a criteria. For ease of analysis, this sample is grouped by absolute volumes of comments into higher and lower engagement criteria under labels of “high” and “low” engagement respectively. The former of these sets the benchmark for engagement at a minimum bound of n= ≥1000 comments and a maximum bound at n= ≤10000 comments. As such, the deepfakes of Obama and the Queen qualify as content with high engagement levels with n= 8793 and n= 3576 comments respectively, whilst the criteria for low engagement content sees a minimum bound of n= ≥100 comments and a maximum bound set at n= ≤1000 comments. In the low engagement set, the videos of Putin and Kim feature n= 270 and n= 377 comments respectively. This forms a total of n= 13016 comments in the sample.

Within the high category of engagement, the Obama video is often cited as a progenitor of the discussion of political deepfakes in the wild (Figure 7). This viral BuzzFeed video has been referenced as a seminal piece of deepfake content, having attracted a substantial amount of engagement since its publication. This video has spurred a wealth of discussions around deepfake content; both on the video itself as well as the implications it has for deepfake technologies more broadly (Ajder, 2019b). Channel 4’s deepfake of the Queen follows in a similar vein (Figure 8), also being a piece of viral content which has witnessed high levels of engagement following its release on Christmas day in 2020. Within the low engagement group, the “dictator” series by RepresentUs, sees these videos of Putin and Kim (Figures 9 and 10) garner a less dramatic number of views and comments in terms of absolute volumes of comments. However, these absolute numbers bely a relatively high comment-to-view ratio when compared with high engagement videos in the sample. Across the resulting sample, the content of these videos traipses the boundary between satire and public education. In these, the educational components are enacted directly through unveiling the means of production behind the deepfake content in question, or indirectly through other means such as the script, organisational branding and the video descriptor. Table 1 below gives an overview of the chosen data sampled in this research.

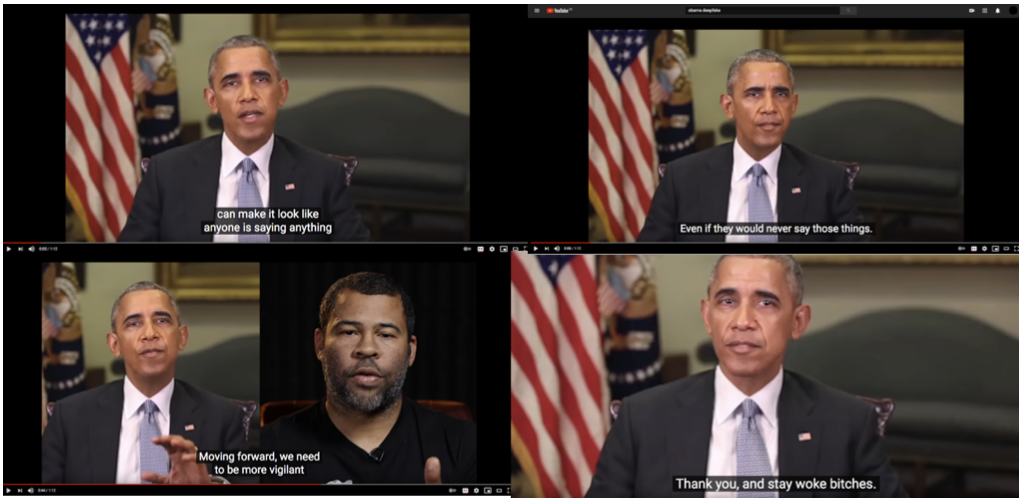

can make it look like
anyone is saying anything
Even if they would never say those things.
Moving forward, we need
to be more vigilant
Thank you, and stay woke bitches.

Figure 7. Screenshots from “You Won’t Believe What Obama Says In This Video! 😉” (2018). Source: YouTube

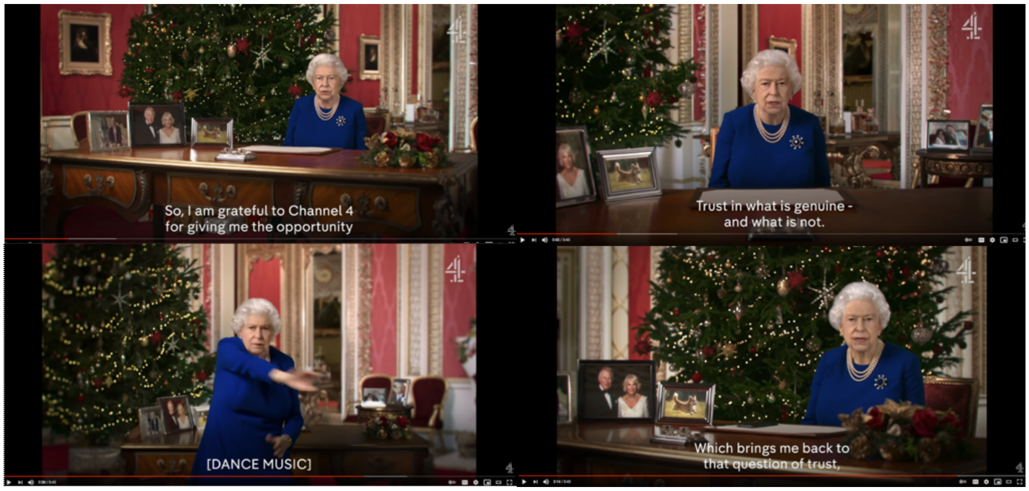
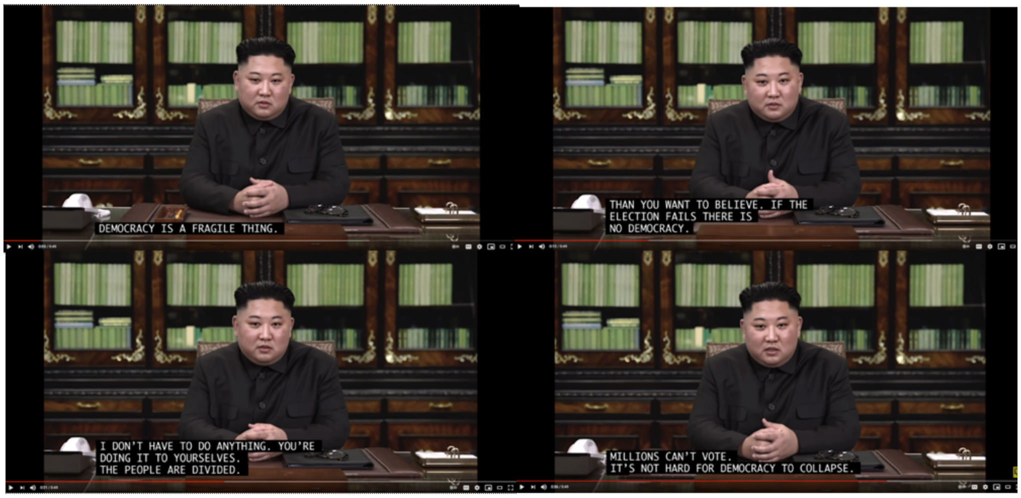

So, I am grateful to Channel 4
for giving me the opportunity
Trust in what is genuine -
and what is not.
[DANCE MUSIC]
Which brings me back to
that question of trust,

Figure 8. Screenshots from “Deepfake Queen: 2020 Alternative Christmas Message” (2020). Source: YouTube



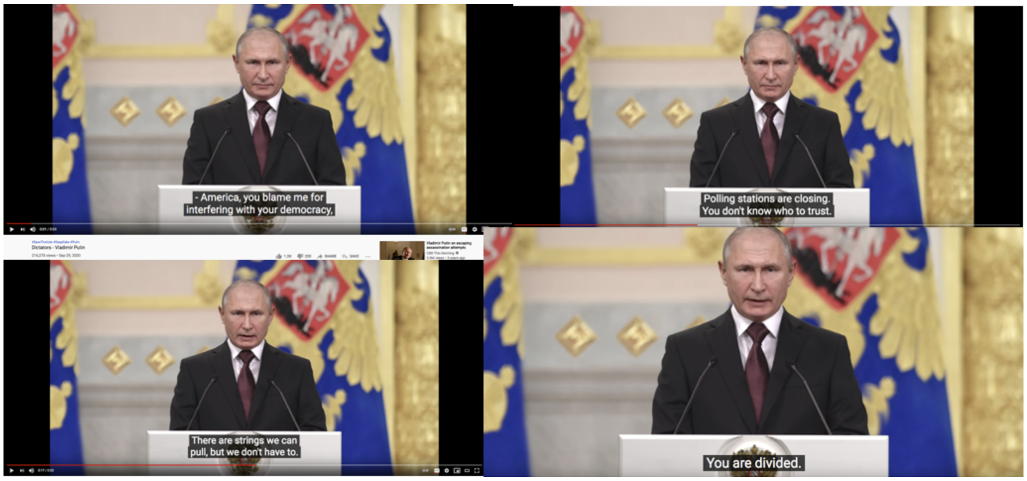
DEMOCRACY IS A FRAGILE THING.

THAN YOU WANT TO BELIEVE. IF THE ELECTION FAILS THERE IS NO DEMOCRACY.

I DON'T HAVE TO DO ANYTHING. YOU'RE DOING IT TO YOURSELVES. THE PEOPLE ARE DIVIDED.

MILLIONS CAN'T VOTE. IT'S NOT HARD FOR DEMOCRACY TO COLLAPSE.

Figure 9. Screenshots from "Dictators - Kim Jong-Un" (2020). Source: YouTube



- America, you blame me for interfering with your democracy,

Polling stations are closing. You don't know who to trust.

There are strings we can pull, but we don't have to.

You are divided.

Figure 10. Screenshots from "Dictators - Vladimir Putin" (2020). Source: YouTube

Table 1. Known Public Figures: Key Data of Individual YouTube Videos from Selected Sample[11]

Title	Published Date	Length (Mins)	View Count	Comment Count

[11] All data is correct as of 13th April 2021. To watch the full list of videos, see: https://www.YouTube.com/playlist?list=PLruuWSlwgyWlc6ZztCp2j9mzhIsJ4B4IJ

High Engagement				
You Won’t Believe What Obama Says In This Video! 😉	Apr 17, 2018	1:12	8,453,010	8793
Deepfake Queen: 2020 Alternative Christmas Message	Dec 25, 2020	3:45	2,235,477	3576
Low Engagement				
Dictators - Kim Jong-Un	Sep 29, 2020	0:49	446,645	377
Dictators - Vladimir Putin	Sep 29, 2020	0:33	227,953	270

Other sources which have also been considered for also being of potential relevance here are two now-infamous videos featuring Facebook CEO, Mark Zuckerberg and Belgian Prime Minister, Shophie Wilmès, with the former created by the artists Bill Posters and Daniel Howe, and the latter by environmental movement, Extinction Rebellion. These videos have been removed several times from YouTube’s platform despite being labelled as deepfake content and so videos have not been included in this selection, being comparatively less stable than the existing sample. Besides these alternatives, videos of former US President Donald Trump have also been omitted here despite being an appropriate candidate for study. Despite being of theoretical interest and relevance, the extent of Trump’s popularity as a subject of satirical deepfake content means that content of the former US President could be a standalone area of research in itself, given the volume of such content on YouTube alone.

Ultimately, the sample here balances a pertinent public collection of figures with a spectrum of engagement metrics. This is crucial for responding to some of the key contentions in the field around safety and policy-making, as well as answering the research questions at hand.

3.3 Data Processing and Cleaning

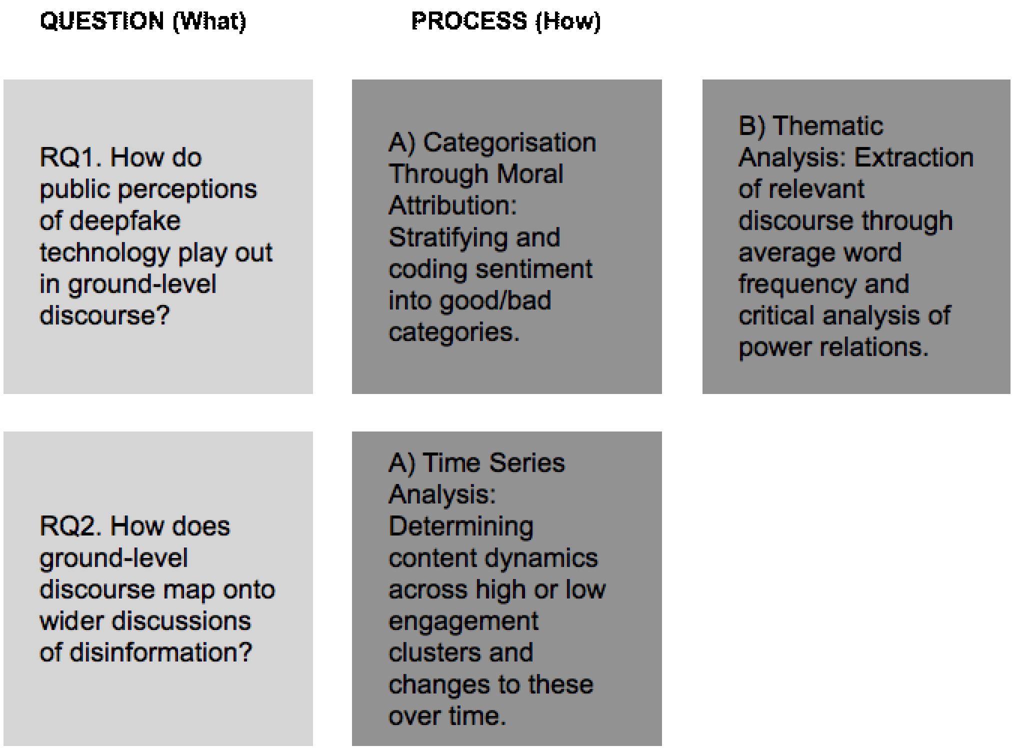
The large, unstructured datasets extracted from these videos consist of users sharing their opinions, debating points and disseminating content. To account for a lack of similar existing datasets, the initial extraction process used two open-source browser-based tools (Netlytic, Coberry) to triangulate the raw data to ensure quality data was obtained. The use of two tools here is intended to mitigate the disparities and data gaps created by computationally-aided extraction in some instances, with different classifiers counting some comments and not others (Burnham et al., 2008: 51; Gruzd, 2016). Further to these comments about data purity, open source tools are selected here for maximum replicability and data generalisability.

Additionally, the volume of comments extracted for this study means that the dataset has been subject to rigorous pre-processing to create a more convenient format for analysis. Following best practice for greater standardisation and replicability of this data in other programs, the removal of duplicates, spam, special characters and an excessive use of emojis, has been carried out where necessary. This process produced discrete data for each video which were then unified into a single dataset for cross-comparison of overall trends once robustness checks on the raw data were completed.

4. Research Findings for Analysis and Discussion

This section reports on the findings and interrogates these results in context of the research questions. To briefly restate the questions at stake in this project, these are as follows:

- RQ1. How do public perceptions of deepfake technology play out in ground-level discourse?
- RQ2. How does ground-level discourse map onto wider discussions of disinformation?



QUESTION (What)	PROCESS (How)	
RQ1. How do public perceptions of deepfake technology play out in ground-level discourse?	A) Categorisation Through Moral Attribution: Stratifying and coding sentiment into good/bad categories.	B) Thematic Analysis: Extraction of relevant discourse through average word frequency and critical analysis of power relations.
RQ2. How does ground-level discourse map onto wider discussions of disinformation?	A) Time Series Analysis: Determining content dynamics across high or low engagement clusters and changes to these over time.	

Figure 11. Research question and methods by section

To build on the processes and approaches labelled in Figure 6, Figure 11 here unpacks these ideas more concretely in regards to specific processes used. The former of the research questions here assesses the semantic dimensions of ground-level discourse. Section 4.1.1 presents insights into how audiences perceive and categorise deepfake content by examining the proportions of positive to negative user commentary across the sample. This is followed by a deeper dive into some of the specific themes which emerge from user commentary on YouTube in Section 4.1.2. These iterative interrogations of user commentary provide pertinent areas for discussion in response to the first question concerning what is said at the level of public discourse.

The second question builds on this semantic exploration of public perception and looks at the burgeoning trend towards the greater democratisation of deepfake technology. Section 4.1.3 addresses changes in user engagement across high and low engagement categories using time series analysis. Analysing a chronology of deepfake content allows us to take a more empirical look at the gauging the scale and speed of this technology and how it maps onto disinformation.

4.1 Research Findings

4.1.1 Categorisation Through Moral Attribution

In accounting for one of the most prominent sites of discourse on deepfake content, this in-depth study of a sample of YouTube-hosted deepfake videos studies civilian users' behaviour to empirically understand how deepfake technologies might implicate the wider security environment. Categorisation through moral attribution distinguishes whether the fear-mongering claims made by policymakers, lawmakers and regulators are equally matched by lay perceptions of labelled deepfake content. Investigating how users categorise labelled deepfake content clarifies some of the more opaque sociotechnical entanglements which emerge in this topic. Moreover, this narrowed focus on moral attribution as a facet of public perception and the subsequent use of clearly distinguished categories classified as either “good” or “bad” are based on similar research about media frames and audience frames (Lee et al., 2021: 4) which use similar categories to suggest whether commentary is good (spurring interest and humour) or bad (inciting negativity and fear). The respective sets of terms used for both good and bad moral attribution here have been standardised using Netlytic's default “feelings" category. This allows us to garner a clearer impression of audience perception at scale and scopes out whether the technology is perceived as an inherent threat by the general public, even when it is labelled as knowingly faked.

Table 2 displays the sample's fifteen most frequently-used terms. Capping the highest ranked terms here facilitates a deeper and more concentrated analysis of the commentary and is also practical for studying this highly slippery area, given that highly ranked words are less volatile to relative changes in comment frequency. For a full list of terms used to code moral attribution through good and bad sentiments, see Appendix A. The terms shown in the table below constitute the bulk of word frequencies in the sample, being 92.5% of the overall volume of commentary. This breaks down further to be 94.9% and 87.2% of the total number of word frequencies coded good and bad respectively. Being a significant proportion of the total number of comments, the terms in Table 2 are roughly representative of the sample as a whole.

Table 2. Moral Attribution: Overall Frequency of Good/Bad Sentiments

“Good” Sentiment	**Frequency**	**“Bad” Sentiment**	**Frequency**
'good'	419	'bad'	189
'funny'	273	'scary'	83
'great'	176	'dangerous.'	57
'nice'	98	'creepy'	34
'hilarious'	67	'terrible'	32
'kind'	63	'evil'	21

'perfect'	45	'awful'	20
'happy'	34	'scary'	13
'fine'	19	'ashamed'	13
'calm'	17	'angry'	10
'fair'	13	'worried'	10
'fantastic'	12	'hurt'	8
'proud'	11	'ill'	8
'lovely'	7	'upset'	8
'proud'	6	'lazy'	6

Taken as a representative sample, the absolute volume of commentary per category in Table 2 sees the list of sentiments coded as good being overwhelmingly greater than those coded as bad. The volume of comments coded as good is twice the size of corresponding bad sentiments within the same sample. This is of a total of n= 1328 and n= 459 sentiments respectively coded good and bad. The overwhelming skew towards comments coded as good contradicts some of the concerns inferred by policymakers and the moral panics of wider policy-making.

In terms of a semantic breakdown of the commentary in Table 2, the most frequently occurring terms are concentrated in the use of “good” and “bad” themselves but overall, the number of occurrences following these initial terms drops significantly. The second most-frequently used terms are “funny” and “scary” respectively; this corresponds with trends identified earlier in this research about the nexus between humour and fear. Variations of these themes also occur further down the ranking, with “great”, “nice” and “hilarious” juxtaposed with “dangerous,” “creepy” and “terrible” in turn. These terms lose their direct relevance to the video content itself as the usage frequency decreases, but the commentary towards the bottom of this list generally demonstrates similar trends to the above, forming relatively direct contrasts (“perfect” versus “awful”, “calm” versus “angry”) to one another when compared with the terms earlier in the rankings. These less frequently-occurring terms in Table 2 exhibit a narrowing gap towards the bottom of the rankings, as another indicator of dwindling relevance. Overall, public perception is seen to consistently err towards positive moral attribution, especially where commentary is most concentrated.

These trends are largely consistent with the results obtained from the individual videos seen in Table 3. This table shows that the videos in the sample skew overall towards positive moral attribution. But a closer examination reveals that the margin of difference between good to bad sentiments is narrower in the lower engagement groups, with a dramatic difference being more notable in the high engagement category. The results see around two thirds responding positively in the former, high engagement group (between 72% and 79%), and slightly fewer responding positively in the latter, lower engagement group (between 53% and 64%).

Table 3. Moral Attribution: Frequency of Good/Bad Sentiment of Individual Videos

Title	Type	“Good” Sentiment	“Bad” Sentiment
You Won’t Believe What Obama Says In This Video! 😉	High Engagement	689	186
Deepfake Queen: 2020 Alternative Christmas Message	High Engagement	549	210

Dictators - Kim Jong-Un	Low Engagement	47	26
Dictators - Vladimir Putin	Low Engagement	43	37

These high concentrations of positive sentiment could be taken as an indicator of how lay commentary co-constructs commonsense knowledge, with moral sentiment being more confidently ascribed as positive and more pronounced amongst a high engagement category which features many user engagements, and is inversely less certain when content features lower engagement. The aforementioned issues around bias could also fit into this pattern of co-constructed sensemaking, with greater volumes of positive sentiments begetting more of the same. So just as the effects of continued influence and Illusory truth can be reinforced in a feedback loop by other users (Britt et al., 2019; Larson, Kaleda and Fenstermacher, 2019), user commentary on labelled deepfake content could prove to be useful in revealing deepfake content to others in the network, erasing the uncertainty which would otherwise prevail in the lower engagement categories.

4.1.2 Thematic Analysis

Using word frequencies navigates some of the standard challenges facing the semantic processing of audiovisual content and allows common themes and trends to be identified with relative ease (Choi and Segev, 2016). Exploring word frequencies in this sample showcases the most defined areas of public opinion. This serves to highlight pertinent areas for discussion within this large and unstructured dataset of user commentary. Table 4 shows the four highest ranked terms for each video across the sample to capture a cross-section of key themes. Within the selected sample, these are double processed through open source tools and cross-checked for additional robustness. These terms are subsequently used as a baseline for the thematic extraction of comments seen in this section in Table 5.

Table 4. Overview of Average Number of Top 4 Frequently Occurring Terms

You Won’t Believe What Obama Says In This Video! 😉	Deepfake Queen: 2020 Alternative Christmas Message	Dictators - Kim Jong-Un	Dictators - Vladimir Putin
News = 1503	Queen= 338	People= 44	Putin= 41
Obama= 970	Fake= 301	Democracy= 44	Fake= 30
Fake= 854	Real= 276	Fake= 40	People= 25
Video= 610	People= 269	Real= 39	Russia= 25

A tier-system has been used to stratify unstructured user commentary more meaningfully into primary, secondary and tertiary tiers – labelled here as Tiers 1 through to 3. These derive from the most frequently occurring terms averaged across the cross-section seen in Table 4 to reconcile these and make sense of what is said by the public. As a topline, we can observe a consistently high volume of commentary across the sample about the figures at the centre of the video themselves, who are directly referenced by name across the majority of the sample, with “Dictators - Kim Jong-Un” being the exception here. Otherwise, secondary themes of the discussion see mentions of “fake”, “real” or both these terms in some instances, dominate as themes across high and low engagement groups. This is followed by mentions of “people” and “democracy,” which form other, tertiary areas of discussion.

From this a cross-section of commentary featuring high engagement metrics from each video, being primarily ranked by the number of likes and then by number of replies, can be used to establish a more nuanced discussion of this dataset and unpack the critical power relations within discourse on this network.

In a cross-comparison of commentary across the high and low engagement groups in this sample, Table 5 shows that the most popular strains of discourse to emerge from Tier 1 seem to be knowing jokes about the subjects depicted in the video, mostly about the quality of mimicry and the character traits of the subjects, with the former of these noting voice as a key trait (“Gotta find a better voice for Putin, this one's too deep”; “Wow, Barack Obama does a mean Jordan Peele impersonation”; “I thought Kim Jong Un couldn’t speak any English...apparently he is fluent now”). This is typical in many ways, with other studies of user commentary demonstrating a precedent for users addressing personal attributes of the video subject as a thematic constant (Veletsianos et al., 2018). The overview of the commentary gathered from this tier puts positive, satirical sentiments at the forefront of this group. The popularity of wry comments about the public figures seen in these videos seems to confirm the findings from the previous sections; the figure at the centre of these samples drive commentary which is more funny than scary and in turn, more good than bad.

Table 5. Selection of High Engagement Commentary on Tier 1 (Video Subjects)

Comment	Like Count	Reply Count
Wow, Barack Obama does a mean Jordan Peele impersonation	10123	44

Wow Obama does a great Jorden [sic] Peele	2539	17
Sound [sic] like Obama mixed with one of the muppets!	1251	8
Plot twist: Queen died some years ago, but is deepfaking everytime when shes showing yourself at TV	434	20
Imagine the Queen watching this right now 😂 and being like this is the truth	354	9
Alt-Queen: “You hear that, Andrew? That's the sound of inevitability, that's the sound of your death, goodbye, Andrew.” Andrew: “Jeffrey understood me, mother.”	339	4
Gotta find a better voice for Putin, this one's too deep	39	2
God bless Vladimir Putin	21	0
putin	7	0
I thought Kim Jong Un couldn’t speak any English...apparently he is fluent now	3	1
🙁 For second I thought Kim was being thoughtful and concerned about us poor Americans.	3	4
The person who did this has a kindergarden [sic] understanding of the DPRK and Kim Jong-Un	3	3

Moving from a less localised set of trends which result from Tier 1, on the subjects of the videos themselves, commentary which looks more broadly beyond the particular public figures here, can be found in the commentary around fake and real. An overview of the discourse in Tier 2 produces results which are largely not dissimilar to the previous tier, with the results from Table 6 maintaining that parodic responses are potent when it comes to garnering engagement. However, a key difference between the first and second tiers of commentary is that the latter generally veers towards a more serious appraisal of deepfake technology, with a notable admixture of commentary relating to deepfake content in a specific as well as more general context. Some users offer ambiguous assessments on the potential of this technology in relation to its ability to fool audiences ("I didnt [sic] realize this was computer generated. I just thought it was recorded on a flip phone...") and otherwise, this tier showcases the general viewer's demonstrable level of context awareness with a nous for the wider implications of how the audience reception of deepfakes plays out in a broader context ("Wasn't perfect, but does show how far this type of technology has come. There will certainly be a time where [it] is almost impossible to tell the difference"'; "these DeepFakes are really dangerous"). This resonates with the key findings of Lee et al's (2021) study which found a significant number of comments within deepfake content appraising believability.

Table 6. Selection of High Engagement Commentary on Tier 2 (Fake/Real)

Comment	Like Count	Reply Count
Plot twist: the Jordan peele face was a deep fake made by Obama	1767	12
This is more real than the real video	1647	24
Wasn't perfect, but does show how far this type of technology has come. There will certainly be a time where is almost impossible to tell the difference	1416	57

Plot Twist : The Queen's been a deepfake for 50 years.	1401	35
who here knew from the start this was fudged and was fake.	1222	26
I didnt [sic] realize this was computer generated. I just thought it was recorded on a flip phone...	645	4
I know I'm supposed to be afraid of electoral fraud, but I'm more afraid of this deepfake.	423	12
Thats [sic] a good deep fake.	171	1
these DeepFakes are really dangerous	153	12
Plot twist, this is not deepfake.	137	4
His real voice in English is a lot deeper than this. Doesn't sound like him at all.	34	9
"It's okay to fake videos if it's for OUR side" Okay then, just don't get mad when the other side does it for their own purposes	34	7

The trends which surface in Tiers 1 and 2 still feature prominently in this third tier about people and democracy. However, there is an incremental but notable shift here towards more serious commentary in the form of longer form opinions. While the number of satirical quips dwindles in this tier, the number of developed opinions rises, with users sharing thoughts about political factions and democratic principles (“A republic is a form of democracy. We do not live in a PURE democracy.”), as well as expanding on concerns about deepfake content exhibited in the previous tier (“People need to be more aware of what they see on the internet especially on Facebook, be more cautious about what they read and watch”). Table 7 details some of these comments in greater specificity which documents commentary around domestic and geopolitical tensions, such as opinions about the US, Korea and Russia (“The fact that millions of people point to fabricated Russia nonsense while saying literally 0 about Israel is at best extremely laughable”).

Table 7. Selection of High Engagement Commentary on Tier 3 (People/Democracy)

Comment	Like Count	Reply Count
People need to be more aware of what they see on the internet especially on Facebook, be more cautious about what they read and watch.	1146	7
This is coming from the people who write things like what your favorite pancake says about your future gay lover.	893	9
What people thought would happen to deepfakes. 2020- BAKA MITA	157	0
I think they could have made this more realistic... But pulled back to stop it REALLY freaking people out.	63	1
Just shows you how images can easily be imitated, even if just parody Wake up people	45	3

trying to warn you The Queen and most of senior royals been dealt wih [sic]		
How is this deeply disrespectful? It’s just a good laugh. Some people are so boring	41	1
The US must own its problems. It seems that the system is reaching its limit of sustainability. The solution is simple but painful: Let go of militarism, reinvest in your people and accept that your GDP might reflect the size of your population.	32	4
A republic is a form of democracy. We do not live in a PURE democracy.	8	0
You think this garbage will get people to take you seriously?	7	2
It's times like this I'm glad I'm a separatist I don't care for either side people always say vote for the Lesser evil.. I rather vote for neither.. the green party or independent and two I can join a separatist party.. but no matter what you do you can't replace government with another government they'll always become corrupt We the People doesn't work in America anymore.. it's all	7	0

about self-interest and Mimi me next to capitalism of course..		
While I support and promote the electoral reforms to protect and advance democratic ideals put forward by Represent.US, I do take exception to the use of DeepFakes and the image of Kim Jong-Un for fear-mongering purposes. *Propaganda and fear are threats to democracy no less significant than electoral malfeasance.* It is further worth considering that the dictatorial state presiding over North Korea is a direct consequence of Western intervention in the social, economic and political affairs on the Korean Peninsula. Too many people forget that the war with North Korea has yet to come to an end. The economic war of attrition has persisted since military hostilities ended in 1953 after 3 years of conflict that reduced 80% of North Korea to rubble with millions killed.	6	8
The fact that millions of people point to fabricated Russia nonsense while saying literally 0 about Israel is at best extremely laughable	6	1

The spectrum of commentary in the overall sample ranges between amusement and concern, and within the reactions to each video, a variety of aptitudes for interpreting labelled deepfake content is showcased. This commentary demonstrates a set of disparities in literacy towards deepfake content and otherwise signals an imbalance of power within the network. These power relations play out as users collectively unveil and unpack the implications of deepfake content for each other on the network. This is ultimately a form of co-constructed sensemaking with the complex sociotechnical entanglements of deepfake content being further complicated by a set of contradictory audience reactions in satirical and cautionary commentary. This blended use of humour and caution as mechanisms of discourse give rise to a decentralised but distinct, approach to lay-education in commentary on YouTube. What is indicated in these findings so far is that comments are more critically reflexive than initially assumed, with the ground-level discourse attempting to draw attention to the issue of deepfake content and stimulate greater thinking on the issue.

4.1.3 Mapping Change Through Time Series Analysis

Overall, the data range which spans the period 2018 – 21, suggests that the novelty of individual videos featuring deepfake content is short-lived, receiving less engagement over time and no dramatic resurgence in comments after an initial peak in engagement (see Figures 12 to 15). Taking levels of all-time engagement for each video since their individual publication, these results prove consistent across both high and low engagement groups in this sample, which feature a sharp spike, reaching an average peak of n= 1584 comments across high engagement videos and a peak of n= 64.5 comments across low engagement videos. This figure is sustained for the first few weeks up to the period of a month, followed by a steady decline in comments thereafter. This drop is more dramatic in the instance of high engagement videos.

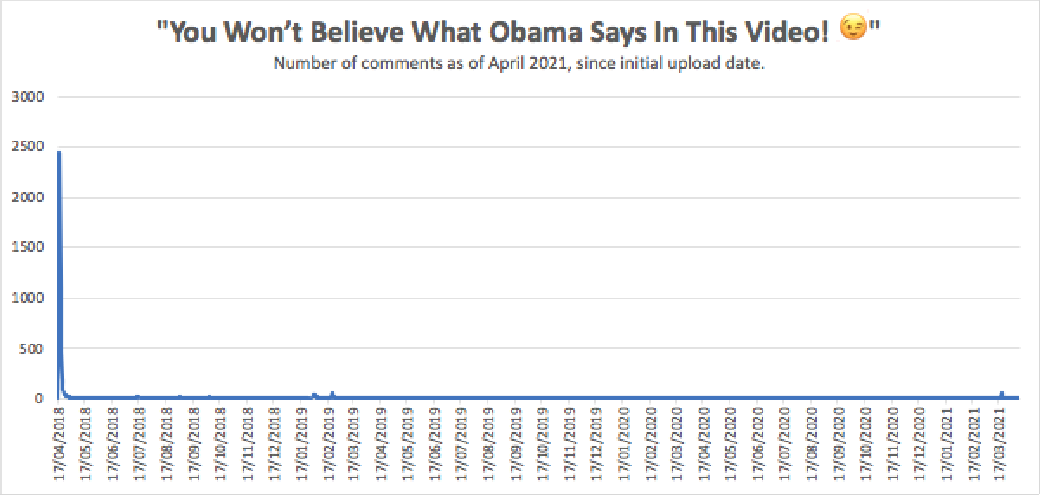


Figure 12. Time Series Chart for “You Won’t Believe What Obama Says In This Video! 😉” – High Engagement Group (2018 – 2021). Source: YouTube

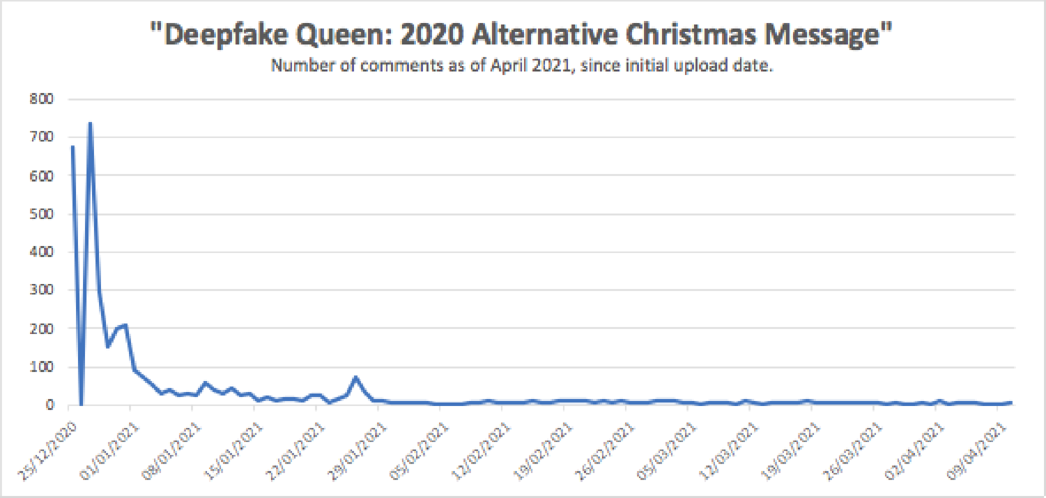


Figure 13. Time Series Chart for “Deepfake Queen: 2020 Alternative Christmas Message” – High Engagement Group (2020 – 2021). Source: YouTube

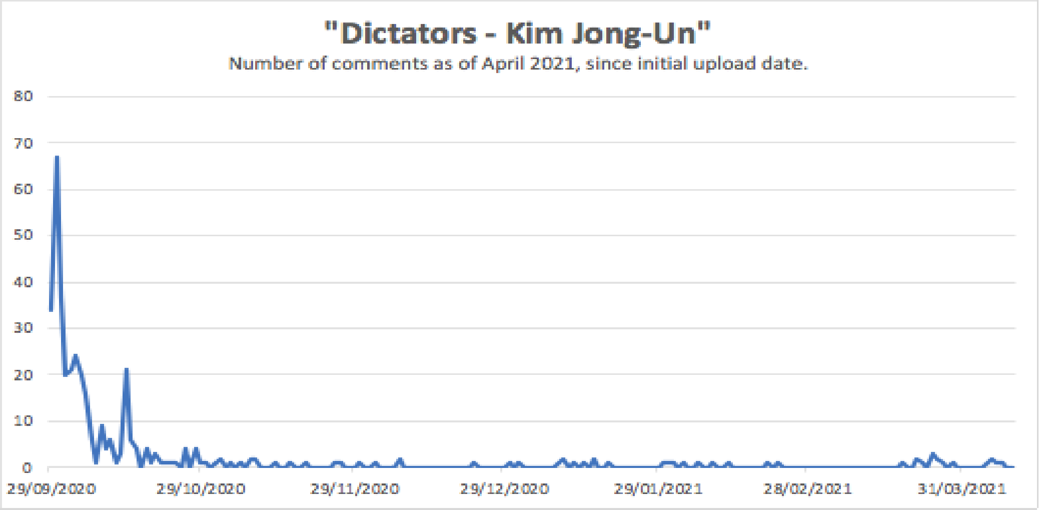


Figure 14. Time Series Chart for “Dictators - Kim Jong-Un” – Low Engagement Group (2020 – 2021). Source: YouTube

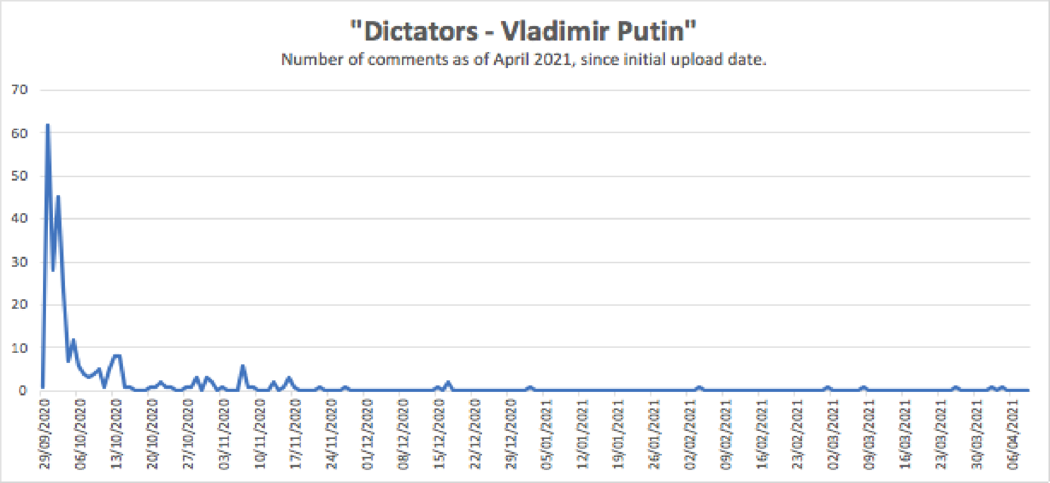


Figure 15. Time Series Charts for “Dictators - Vladimir Putin” – Low Engagement Group (2020 – 2021). Source: YouTube

The results here are largely congruent with the patterns followed by other popular content on YouTube. The patterns seen in these results are not atypical of content on YouTube, which often sees a similar attrition of engagement and a diffusion of content over time (Susarla, Oh and Tan, 2012). This is true of both high and low engagement categories in this sample. Deepfake content experiences artificially inflated engagement from the outset of their publication to the platform as a novel piece of content. However, evidence from the figures above suggest the initial traction gained is unlikely to be sustained or replicated in a comparable way over time. This is an indicator that as the phenomenon becomes less novel, engagement with the content tends to fall as soon as it reaches its peak, and that other, newer clips are likely to receive attention instead. However, this could be inferred to be a localised trend on the platform, given that while the number of overall searches for deepfake content on the web has generally increased over time, it can be noted that the comments on each individual video in the sample declines exponentially over time. These findings support the view that producers of ground-level discourse perceive deepfake content differently to those in positions of power who hold policymaking positions. When paired with findings from earlier in this chapter, ground-level discourse reveals a picture of labelled deepfake content which shows it to be within the conventional bounds of novelty for lay audiences, who in dialogue with other users, form a well-balanced set of assessments and approaches to interpretation.

5. Discussion

Having opened with a discussion of the broader social context of deepfake content to explore the sociotechnical dimensions of the technology and the disinformation sphere at large, this research explores public perceptions of deepfake content hosted on YouTube. The key findings from this research have seen that the greater number of sentiments coded as good far outweighs those coded as bad. Throughout the sample of user commentary studied there is a tendency for discourse to meet at the nexus of humour and caution, with satire emerging as a popular mechanism for political expression. This research also finds that a decentralised but distinct approach to lay-education is present amongst YouTube's audience. It also finds that deepfake content loses engagement on YouTube over time and experiences attrition after an initial period of novelty. These two interrelated questions are as follows:

- RQ1. How do public perceptions of deepfake technology play out in ground-level discourse?
- RQ2. How does ground-level discourse map onto wider discussions of disinformation?

This section concludes the analysis component of the research, unifying the findings in the previous section to answer the research questions at stake.

5.1 RQ1: Labelling as a Progenitor of the Funny-Bad Nexus

This research unpacks the complex sociotechnical entanglements of deepfake technology to provide a baseline for how labelled deepfakes are perceived in the wild. And though the comment classification system here could be perceived to be a "lightweight alternative to complex image processing" (Schultes, Dorner and Lehner, 2013: 661), it is an effective way of extracting meaning and interpreting ground-level discourse around deepfake content. This discourse offers organic insight into how the general public actually perceive deepfakes as a phenomenon and is important considering that laypeople are the demographic most likely to spread disinformation. Analysis of user commentary reveals that discourse around deepfake technologies is still fraught by a series of power relations, marked out for instance in Sections 4.1.1 and 4.1.2 by audience disparities in digital literacy. These disparities are expressed through varied levels of emotional response towards deepfake content, which are sometimes contradictory.

Examining YouTube's subcultural tenets as a civic forum reflects some of the sensationalism propounded by policymakers and journalists, revealing similar findings to adjacent studies on public commentary which sees audiences exercise "participatory potential" when faced with contrary narratives by policymakers (Pinto-Coelho, Carvalho and Seixas, 2019), with some users reflecting on the wider political context that surrounds this content “The fact that millions of people point to fabricated Russia nonsense while saying literally 0 about Israel is at best extremely laughable.” The results here demonstrate forms of alternative sensemaking in action, with the public perceptions of deepfake content acting as an innocuous form of education or public “inoculation” (Compton and Pfau, 2009). Humour is particularly important to fulfilling this participatory potential and inoculating audiences. Humour is a potent aspect of YouTube’s subcultural nous and is used in several capacities to educate other users. As seen in Tables 5 and 6, this is something which is especially evident in comments exercising humorous tropes to convey this; “Plot twist: the Jordan peele face was a deep fake made by Obama”; “Plot Twist : The Queen's been a deepfake for 50 years”; “Plot twist: Queen died some years ago, but is deepfaking everytime when shes showing yourself at TV.” The role that satire plays in sensemaking is one which current scholarship around disinformation tends to overlook, despite satire being a powerful form of political expression (Reilly, 2012).

Ultimately, the ground-level results here pivot away from the moral panic narrative propounded by policymakers. The consistent skew of positive to negative sentiments demonstrates that the moral panic exuded from a policy-making perspective fails to account for alternatively expressed, co-constructed means of sensemaking which arises from ground-level discourse. Further, the analysis here confirms that these views from policymakers only capture a fragment of the full scope of public responses, and that there are also major disparities between legislation, policy recommendations and the security community’s attempt to understand the foundational issues at the heart of AI. The attempt to enforce hasty countermeasures may do more inchoate harm than good, considering that English-speaking audiences on YouTube presently have a grasp of the technology's potential and are also evolving a reflective set of viewpoints on the issue, with any disparities in knowledge patched by more informed users.

And to return to earlier sections on similar adversaries being pitted against each other in real-world security environments, as in the case of fake news, the outcomes of this research suggest that similarly, the problems created and exacerbated by deepfake content can also be resolved or even prevented by the same means. This is parallel to developments in the technical field of AI safety, which has already experienced some success with fighting deepfakes using other deepfakes (Li et al., 2020). An issue which is fundamentally human at its core cannot be solved by technical patches alone, and so, AI safety and real-world security concepts need to be reconciled (Aliman, Kester and Yampolskiy, 2020).

5.2 RQ2: Democratisation and the Speed and Scale of Bias

In regards to answering the second of these questions on how commentary about deepfake content maps onto the bigger picture of disinformation, we see that commonsense thinking ultimately prevails despite concerns around the current speed and scale of this technology's proliferation. Trends inferred from the time series analysis carried out in Section 4.1.3 builds on the previous sections to better understand civilian behaviour over time and finds that deepfakes attract high levels of engagement for their novelty.

And so, returning to the ideas explored in the first part of this research about truth being co-constituted by public perception (Dubovi and Tabak, 2020), the treatment of deepfake technology as an "epistemic" threat to truth itself (Fallis, 2020) seems heavy handed if we consider how people come to treat falsified content over time and the attrition of engagement videos typically face. The traditionally speculative and anecdotal approach taken by more technocratic policy-making expresses concerns about the public's uncritical absorption of content, but when paired with the findings from Sections 4.1.1 and 4.1.2, we note that greater engagement also means greater opportunities for the public to enter into discourse and interrogate the content in question.

These interactions with deepfake content are not solely opportunities for disruption and there is precedence for social media acting as a “catalyst” (Halpern and Gibbs, 2013) for user deliberation, with comments pointing out the more malign attributes of such technology; “People need to be more aware of what they see on the internet especially on Facebook, be more cautious about what they read and watch.” This study has shown that the public perception of deepfake content is more critical than widely thought, with users appraising this content and educating other users on the platform that this content occurs on. And while this research does not contend the potential dangers of deepfakes it does critique the extreme binaries that have been assumed of this technology. It also counters fearmongering assessments which do not go far enough into interrogating the root concepts at the core of deepfake technologies. As such, deepfake content should be considered part of a gradual knowledge acquisition process rather than being something to censor outright, given that greater exposure to labelled deepfakes could incite critical discourse if the results evidenced here are to be considered.

Despite greater access to tools which conflate the quality of amateur and professional content (Simonite, 2020), the findings here suggest that users are likely to find deepfake content less sensational over time when presented with labelled content. There is sparse long-term evidence to suggest that discrete pieces of deepfake content are of sustained interest to the public beyond their novelty value. This trend is not dissimilar to the previous hype experienced by software tools such as Adobe Photoshop which was touted as having significant potential for being a tool of disinformation (Lewis and Nelson, 2019). This greater desensitisation to deepfake content could be potential interest for current AI safety efforts, especially those directed towards detection.

5.3 Ethical Considerations

The data here focuses on user-generated content and commentary which is shared publicly. This dataset of comments is processed in line with the University of Glasgow's Code of Ethics as well as drawing best practice from the Association of Internet Researchers (AOIR) Ethics Committee's Ethical Decision-Making and Internet Research Recommendations (2016). The combination of these guidelines formulate an ethical approach to the research undertaken here and ensures that reasonable duty of care is exercised in this treatment of internet-based subjects who are only ever referred to indirectly in the research data. Significant effort has been made to balance the rights of subjects with an assessment of the contextual harms, even those which are not immediately apparent.

To mitigate any more localised ethical issues, this dataset is compliant with local data protection laws and GDPR regulations, with subject identifiers stripped where necessary to ensure that specific authors remain anonymous (Burnham et al., 2008: 283-302). YouTube as a content host states that users may freely publish and re-share collected data, and that the responsibility for video content on the platform belongs to those who upload it to the platform (YouTube, 2021). And as such, this research operates within the bounds of YouTube's ToS and does not knowingly compromise between impact and ethics, nor does this research and its results carry any significant risk or discomfort to human subjects.

5.4 Limitations

While this research sheds light on the complex interactions between deepfake content and their audiences, there are two key limitations which emerge, pertaining largely to the social and technical elements of the research.

As one of the first studies to explore labelled deepfakes in the wild and their sociotechnical entanglements, the data here accounts solely for YouTube as a wild environment. Further, the period accounted for here captures only a snapshot of a fast moving phenomenon. As such, the dataset of data here has limited generalisability when scaled to other environments and platforms, with each wild environment bearing its own distinctive subcultural traits and networked behaviours which dictate how discourse and disinformation spreads. While the resulting commentary is highly context specific, it is for the most part, a representative sample and otherwise provides a focussed baseline for scoping future research.

The second of these limitations concerns ethical barriers and the difficulties this poses for understanding the causal inferences behind this sample; that is, how groups of users within the sample might be stratified by demographic data. This would provide more in-depth, causal reasoning behind user commentary; user data on gender, class, ethnicity and so on would highlight the less visible dimensions of how audiences perceive deepfake content, and bring to salient issues around disparities in perception by more marginalised groups. However, this is highly sensitive data and difficult to access, so this research compensates for these limitations by adopting more rigorous methods to find contextual confirmation; for instance, adopting robustness checks to glean a fuller scope of data without resorting to the invasive disaggregation of each user’s metadata.

This study is intended to act as a concrete basis for future analysis, with the use of open source data and tools intended to build-in replicability. And so, despite the outlined limitations, the analysis and data here provide critical insights for understanding the role of the public interactions in the wild and how these play out in constructing a security narrative. It is hoped that the set of approaches adopted here can advocate for greater sensitivity around future research into deepfake technology's sociotechnical aspects.

5.5 Further Directions for Research

As a topline, this research has found that public perception of labelled deepfake content hosted on YouTube contradicts many of the existing concerns and worst-case assumptions made by policymakers. This research hopes to lay the foundations for more sustained research on deepfake content in the long term. To achieve this goal, this research could be expanded in two directions. The first concerns how malicious deepfake content plays out on other platforms in multi-channel efforts towards disinformation, with a second related research direction suggesting that closer attention be paid to how the spread of deepfake content maps onto non-Western contexts.

To briefly outline the first of these proposed research directions, exploring how opinions on deepfake content play out on other platforms could draw from aggregate-level data collected from one or several of the aforementioned content platforms on both the surface web and the dark web. Unifying localised conversations on these platforms could prove complementary to the research here, providing greater robustness for discourse-based policy work, especially given the fast-encroaching concerns with the democratisation of this technology.

This research has chosen to privilege the primary language of YouTube users, focusing on English language discourse only. The wider approach to user commentary in this research could be transposed to non-English content, with deepfake content in non-Western contexts becoming a pertinent area of further study. Treating deepfakes as a phenomenon which constitutes a singlehanded shift in the disinformation environment is a perspective largely fraught with Western cultural biases, with the concerns around deepfake technology arising largely from US policy-making interests (Botha and Pieterse, 2020; Karanicolas, 2020). And so, despite deepfake technologies and disinformation being touted as endemic in the West, the democratisation of these technologies means that the issues at stake are fast-becoming global in scale.

To continue confining research to solely democratic contexts or political threats to the Global North risks repeating the same failings of current policy-making which affords an artificially narrow lens to the issue. There needs to be an attempt to broaden the scope of focus towards the protection of marginalised groups, particularly the gendered and racial fissures which are already so pervasive within malicious deployment of this technology (Bae, 2019; Ellis, 2018; Karanicolas, 2020; Stolk, 2020). An increasingly localised focus will allow a broad spectrum of security professionals and citizens alike to mitigate the main concerns which come with these technologies.

CONCLUSION

"Only education is capable of saving our societies from possible collapse, whether violent, or gradual."

– Jean Piaget (1934: 31)

This research has explored the security dimensions of safety and the public perception of emerging deepfake technologies as they are deployed at scale and speed in complex real-world environments. This study has taken commentary from deepfake content on YouTube to gain a baseline for understanding public reactions to audiovisual GANs in the wild and how this feeds into disinformation. This research aims to provide a corrective to the aforementioned issues of robustness and reconcile these with the technical limitations of deepfake technologies to build a robust basis for future studies on the real-world deployment of deepfake technology.

The approach here hopes to prove a useful resource for Security Studies scholars and practitioners, especially those seeking to understand how the wider security environment will be affected by technologically enabled disinformation campaigns. This section closes with the theoretical implications carried by this research and how this feeds into a richer understanding of AI safety, as well as how empirically-informed policy could provide a practical approach to resolving the legal and regulatory challenges ahead.

6.1 Theoretical Implications

To return to the research questions which concern discourse and disinformation, this research reconciles two still-disparate halves of the conversation, bridging the gap between high-level policy and technical discussions to understand the wider sociotechnical dimensions of the issue. This is particularly salient given the growing scale and speed of deepfake technologies and their democratisation.

This research attempts to counter some of the assessments made of deepfake technologies as facilitators of disinformation, which has often led to shortsighted solutions and have been seen to hinder progress in wider DL applications. As is true of any technology, deepfakes are not an inherent threat, they are reflective instead of a combination of social and technical issues which more often than not reflect our own worst tendencies. Deepfakes are expected to disrupt the conditions which make reaching socio-political consensus possible, but the research results here prove that the public is less naive than policymakers assume, expressing critical thinking and exercising agency when faced with labelled deepfake content.

However, deepfake content also redefines the quintessential question of ‘security for whom?’ for a generation of users with abundant, frictionless access to this technology. As such, we cannot afford to be complacent about our existing infrastructures for risk despite their seeming sophistication in countering these emerging threats across a multiplicity of source codes. Building civic perspectives from the ground-up is crucial for understanding evolving trends within the sprawling issue of disinformation. It is hoped that this exploration of ground-level discourse can be of long-term benefit to research and policy-making communities.

6.2 Policy Conclusion

The discussion around deepfake technology as a tool of disinformation requires input from ground-level discourse to account for what the public engages with and finds compelling. But despite the undeniable importance of ground-level discourse in assessing public perceptions and its wider application within policy, there are ultimately no panaceas for the worst effects of disinformation given how multifaceted the issue of disinformation is. Even within scholarly issues of AI safety, a broad range of actors are needed to collaborate on the issues outlined here, extending beyond the preoccupations of technical and academic fields, to civil society and regulatory solutions.

Being propagated through ground-level discourse, through its creators and audience, deepfake content needs to be carefully considered in light of its use within disinformation. Such content sees an equally sprawling set of challenges arise for the key legal and regulatory issues within policy, particularly those who are already disadvantaged. But the legal initiatives taken in anticipation of deepfake disinformation and its inchoate harms will likely have resounding implications on a societal level, and so these initiatives will need to tread the line between protection of citizens and complying with constitutional free speech requirements.

Of these proposed solutions, preventative measures taken to inoculate the public against the malicious use of deepfake technologies begins with understanding a baseline; in this instance, this comes from gauging the perceptions of labelled examples of such content. Unifying discrete areas of research for use in policy-making is just the beginning of a collective effort towards countering disparities between different levels of digital literacy. This is especially given how AI and its development have traditionally been something of a black box, privileging particular groups with knowledge over others. And so, to echo the sentiments of many of the scholars cited in this work, there is a need for policymakers to engage with citizen protection in the context of those at the margins who are already disproportionately affected along the fault lines of gender, race and socioeconomic status.

References

Afchar, D., Nozick, V. and Yamagishi, J. (2018) MesoNet: A Compact Facial Video Forgery Detection Network. Research Gate. Available at: https://www.researchgate.net/publication/327435226_MesoNet_a_.

Agarwal, S. and Varshney, L. R. (2019) 'Limits of Deepfake Detection: A Robust Estimation Viewpoint', ArXiv.

Ahmed, S. (2020) 'Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size', Telematics and Informatics. doi: 10.1016/j.tele.2020.101508.

Ajder, H. (2019a) The ethics of deepfakes aren't always black and white, Podium | The Next Web. Available at: https://thenextweb.com/podium/2019/06/16/the-ethics-of-deepfakes-arent-always-black-and-white/ (Accessed: 1 March 2021).

Ajder, H. (2019b) The State of Deepfakes: Landscape, Threats, and Impact. Amsterdam, Netherlands: Deeptrace.

Aliman, N.-M., Kester, L. and Yampolskiy, R. V. (2020) 'Transdisciplinary AI Observatory - Retrospective Analyses and Future-Oriented Contradistinctions', ArXiv.

Allen, G. and Chan, T. (2020) Artificial Intelligence and National Security, Belfer Center for Science and International Affairs. Available at: https://www.belfercenter.org/publication/artificial-intelligence-and-national-security (Accessed: 21 March 2021).

Amarasekara, I. and Grant, W. (2019) 'Exploring the YouTube science communication gender gap: A sentiment analysis', Public understanding of science. doi: 10.1177/0963662518786654.

Amerini, I. et al. (2019) 'Deepfake Video Detection through Optical Flow based CNN', in 2019 Ieee/Cvf International Conference on Computer Vision Workshops (iccvw), pp. 1205–1207. doi: 10.1109/ICCVW.2019.00152.

Amodei, D. et al. (2016) ‘Concrete Problems in AI Safety’, arXiv:1606.06565 [cs]. Available at: http://arxiv.org/abs/1606.06565 (Accessed: 28 February 2021).

Antinori, A. (2019) Terrorism and DeepFake: from Hybrid Warfare to Post-Truth Warfare in a Hybrid World, Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics (eciair 2019). Edited by P. Griffiths and M. N. Kabir, pp. 23–30. doi: 10.34190/ECIAIR.19.053.

AOIR, (2016) Ethical Decision-Making and Internet Research:

Recommendations from the AoIR Ethics Working Committee (Version 2.0), https://aoir.org/reports/

Arik, S. O. et al. (2018) ‘Neural Voice Cloning with a Few Samples’, arXiv:1802.06006 [cs, eess]. Available at: http://arxiv.org/abs/1802.06006 (Accessed: 1 March 2021).

Ascott, T. (2020) ‘Microfake: How small-scale deepfakes can undermine society’, Journal of Digital Media and Policy, 11(2), pp. 215–222. doi: 10.1386/jdmp_00018_1.

Bae (2019) ‘A Study on Criminal Regulations against Deepfake porn’, HUFS Law Review, 43(3), pp. 169–187.

Ballard, T. (2016) ‘YouTube Video Parodies and the Video Ideograph’, Rocky Mountain Review, 70(1), pp. 10–22. Available at: http://www.jstor.org/stable/24898564 (Accessed: 26 February 2021).

Barrett, A. and Baum, S., D., (2016). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. Journal of Experimental and Theoretical Artificial Intelligence 29(2).

Basu, K. (2019) ‘Identification of the Source(s) of Misinformation Propagation Utilizing Identifying Codes’, in Companion Proceedings of The 2019 World Wide Web Conference. New York, NY, USA: Association for Computing Machinery (WWW ’19), pp. 7–11. doi: 10.1145/3308560.3314200.

Bateman, J. (2020) Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios, Carnegie Endowment for International Peace. Available at: https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237 (Accessed: 1 March 2021).

Baum, S. D. (2018a) ‘Countering Superintelligence Misinformation’, Information, 9(10), p. 244. doi: 10.3390/info9100244.

Baum, S. D. (2018b) ‘Superintelligence Skepticism as a Political Tool’, Information, 9(9), p. 209. doi: 10.3390/info9090209.

Bechmann, A. (2020) ‘Tackling Disinformation and Infodemics Demands Media Policy Changes’. doi: 10.1080/21670811.2020.1773887.

Bell, J., 2005. Doing Your Research Project. 4th ed. Maidenhead: Open University Press.

Bengio, Y. et al. (2006) ‘Greedy layer-wise training of deep networks’, in Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press (NIPS’06), pp. 153–160.

Bicquelet, A. (2017) ‘Using online mining techniques to inform formative evaluations: An analysis of YouTube video comments about chronic pain’. doi: 10.1177/1356389017715719.

Borgesius, F. J. Z. et al. (2018) ‘Online Political Microtargeting: Promises and Threats for Democracy’, Utrecht Law Review, 14(1), pp. 82–96. doi: 10.18352/ulr.420.

Botha, J. and Pieterse, H. (2020) ‘Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security’, in Payne, B. K. and Wu, H. (eds) Proceedings of the 15th International Conference on Cyber Warfare and Security (iccws 2020), pp. 57–66. doi: 10.34190/ICCWS.20.085.

Boyd, M. S. (2014a) '(New) participatory framework on YouTube? Commenter interaction in US political speeches'. doi: 10.1016/J.PRAGMA.2014.03.002.

Boyd, M. S. (2014b) 'Participation and recontextualisation in New Media: Political Discourse Analysis and YouTube', in. doi: 10.1075/DAPSAC.55.12BOY.

Braddock, K. (2020) Weaponized Words: The Strategic Role of Persuasion in Violent Radicalization and Counter-Radicalization. Cambridge University Press.

Bradshaw, T. (2019) Deepfakes: Hollywood's quest to create the perfect digital human. Available at: https://www.ft.com/content/9df280dc-e9dd-11e9-a240-3b065ef5fc55 (Accessed: 28 February 2021).

Brady, W., Gantman, A. P. and Bavel, J. V. V. (2019) 'Attentional capture helps explain why moral and emotional content go viral.', Journal of experimental psychology. General. doi: 10.1037/xge0000673.

Britt, M. A. et al. (2019) 'A Reasoned Approach to Dealing With Fake News', Policy Insights from the Behavioral and Brain Sciences, 6(1), pp. 94–101. doi: 10.1177/2372732218814855.

Brooks, C. F. (2021) 'Popular Discourse Around Deepfakes and the Interdisciplinary Challenge of Fake Video Distribution.', Cyberpsychology, behavior and social networking. doi: 10.1089/cyber.2020.0183.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., et al., (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. https://doi.org/10.17863/CAM.22520

Burnham, P., Lutz, K.G., Grant, W.G., and Layton-Henry, Z., 2008. Research Methods in Politics (2nd

edition) London: Palgrave MacMillan.

Cadwalladr, C. (2017) The great British Brexit robbery: how our democracy was hijacked, the Guardian. Available at: http://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy (Accessed: 20 March 2021).

Calderone, M. (2019) 'It's kind of the Wild West': Media gears up for onslaught of deepfakes, POLITICO. Available at: https://politi.co/2YcRKzt (Accessed: 1 March 2021).

Cerdan Martinez, V. and Padilla Castillo, G. (2019) 'Audio-visual fake history: deepfake and the woman in a fake and perverse imaginary', Historia Y Comunicacion Social, 24(2), pp. 505–520. doi: 10.5209/hics.66293.

Chadwick, A. and Vaccari, C. (2019) 'News sharing on UK social media: misinformation, disinformation, and correction', in.

Chakhoyan, A. (2018) Deep fakes could destroy democracy. Can they be stopped?, World Economic Forum. Available at: https://www.weforum.org/agenda/2018/11/deep-fakes-may-destroy-democracy-can-they-be-stopped/ (Accessed: 7 March 2021).

Chambers, B. and Bichard, S. L. (2012) 'Public Opinion on YouTube: A Functional Theory Analysis of the Frames Employed in User Comments Following Sarah Palin's 2008 Acceptance Speech', Int. J. E Politics. doi: 10.4018/jep.2012040101.

Charlie Warzel (2018) Believable: The Terrifying Future Of Fake News, BuzzFeed News. Available at: https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news (Accessed: 8 March 2021).

Chintha, A. et al. (2020) 'Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection', IEEE Journal on Selected Topics in Signal Processing, 14(5), pp. 1024–1037. doi: 10.1109/JSTSP.2020.2999185.

Choi, S., Oh, S. U. and Lee, S.E. (2019) 'Deepfake Image Manipulation: Crisis of Factuality and Creation of Punctum by Deep Automation', Media, Gender and Culture, 34(3), pp. 339–380. doi: 10.38196/mgc.2019.09.34.3.339.

Citron, D. and Chesney, R. (2019) ‘Deepfakes and the New Disinformation War’, 29 June. Available at: https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war (Accessed: 28 February 2021).

Cole, S. (2018) We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. Available at: https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley (Accessed: 1 March 2021).

Collins, A. (2019) Forged authenticity: governing deepfake risks. Available at: https://www.epfl.ch/research/domains/irgc/specific-risk-domains/projects-cybersecurity/forging-authenticity-governing-deepfake-risks/ (Accessed: 1 March 2021).

Compton, J. and Pfau, M. (2009) ‘Spreading Inoculation: Inoculation, Resistance to Influence, and Word-of-Mouth Communication’, Communication Theory, 19(1), pp. 9–28. doi: 10.1111/j.1468-2885.2008.01330.x.

Congressional Research Service, (2019). Artificial Intelligence and National Security. Federation of American Scientists. Available at: https://fas.org/sgp/crs/natsec/R45178.pdf

Cox, K., Slapakova, L. and Marcellino, W. (2020) A Machine Learning Approach Could Help Counter Disinformation. Available at: https://www.rand.org/blog/2020/06/a-machine-learning-approach-could-help-counter-disinformation.html (Accessed: 28 February 2021).

Dabas, C. et al. (2019) ‘Analysis of Comments on Youtube Videos using Hadoop’, 2019 Fifth International Conference on Image Information Processing (ICIIP). doi: 10.1109/ICIIP47207.2019.8985907.

Diakopoulos, N. and Johnson, D. (2019) Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections. SSRN

Scholarly Paper ID 3474183. Rochester, NY: Social Science Research Network. doi: 10.2139/ssrn.3474183.

Dobber, T. et al. (2021) ‘Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?’, International Journal of Press-Politics, 26(1), pp. 69–91. doi: 10.1177/1940161220944364.

Dowdeswell, T. and Goltz, N. (2020) ‘The clash of empires: regulating technological threats to civil society’. doi: 10.1080/13600834.2020.1735060.

Dubois, E. and Blank, G. (2018) ‘The echo chamber is overstated: the moderating effect of political interest and diverse media’, Information, Communication and Society, 21(5), pp. 729–745. doi: 10.1080/1369118X.2018.1428656.

Dubovi, I. and Tabak, I. (2020) ‘An empirical analysis of knowledge co-construction in YouTube comments’, Comput. Educ. doi: 10.1016/j.compedu.2020.103939.

Dupuis, M. and Williams, A. (2019) ‘The Spread of Disinformation on the Web: An Examination of Memes on Social Networking’, 2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00256.

Eidelman, S. et al. (2012) ‘Low-Effort Thought Promotes Political Conservatism’, Personality and Social Psychology Bulletin. doi: 10.1177/0146167212439213.

Ellis, E. G. (2018) ‘Yes, People Can Put Your Face on Porn. No, the Law Can’t Help You’, Wired. Available at: https://www.wired.com/story/face-swap-porn-legal-limbo/ (Accessed: 1 March 2021).

Erzikova, E. and McLean, C. (2020) Drowning Out the Message Together: Analysis of Social Media Comments on a Political Sex Scandal, undefined. Available at: /paper/Drowning-Out-the-Message-Together%3A-Analysis-of-on-a-Erzikova-McLean/d4228ee7c6880ed8e387c3733bc86aa7adec7568 (Accessed: 28 February 2021).

Everitt, T., Lea, G. and Hutter, M. (2018) ‘AGI Safety Literature Review’, arXiv:1805.01109 [cs]. Available at: http://arxiv.org/abs/1805.01109 (Accessed: 11 March 2021).

Evtimov, I., Cui, W., Kamar, E., Kiciman, E., Kohno, T. and Li, J., (2020). Security and Machine Learning in the Real World. Available at: https://arxiv.org/abs/2007.07205.

Fairclough, N. (2003) Analysing Discourse: Textual Analysis for Social Research, undefined. Available at: /paper/Analysing-Discourse%3A-Textual-Analysis-for-Social-Fairclough/ca42c0b46569fac64177b7b95d20e414e8895801 (Accessed: 28 February 2021).

Fallis, D. (2020) ‘The Epistemic Threat of Deepfakes’, Philosophy and Technology. doi: 10.1007/s13347-020-00419-2.

Fazio, L. (2020a) ‘Pausing to consider why a headline is true or false can help reduce the sharing of false news’, Harvard Kennedy School Misinformation Review, 1(2). doi: 10.37016/mr-2020-009.

Fazio, L., (2020b). “Out-of-Context Photos Are a Powerful Low-Tech Form of Misinformation”. The Conversation, February 15, 2020. Available at: https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959

Foer, F. (2018) The Era of Fake Video Begins, The Atlantic. Available at: https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/ (Accessed: 1 March 2021).

Geirhos, R. et al. (2020) ‘Shortcut Learning in Deep Neural Networks’, arXiv:2004.07780 [cs, q-bio]. Available at: http://arxiv.org/abs/2004.07780 (Accessed: 9 March 2021).

Gillespie, T. (2020) ‘Content moderation, AI, and the question of scale’, Big Data and Society, 7(2), p. 2053951720943234. doi: 10.1177/2053951720943234.

Gilmer, M. (2019) As concern over deepfakes shifts to politics, detection software tries to keep up, Mashable. Available at: https://mashable.com/article/deepfakes-political-threat-detection/ (Accessed: 20 March 2021).

Goodfellow, I., McDaniel, P. and Papernot, N. (2018) Making Machine Learning Robust Against Adversarial Inputs | July 2018 | Communications of the ACM. Available at: https://cacm.acm.org/magazines/2018/7/229030-making-machine-learning-robust-against-adversarial-inputs/fulltext (Accessed: 1 April 2021).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., (2014). Generative Adversarial Networks. Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680.

Gosse, C. and Burkell, J. (2020) ‘Politics and porn: how news media characterizes problems presented by deepfakes’, Critical Studies in Media Communication, 37(5), pp. 497–511. doi: 10.1080/15295036.2020.1832697.

Grave, J. (2019) ‘The Politics of Pictures: Approaching a Difficult Concept’, Social Epistemology, 33(5), pp. 442–451. doi: 10.1080/02691728.2019.1652862.

Grothaus, M., (2021). Trust No One: Inside the World of Deepfakes. Hodder Studio: London.

Gruzd, A. (2016) Netlytic: Software for Automated Text and Social Network Analysis. Available at http://Netlytic.org

Guera, D. and Delp, E. J. (2019) ‘Deepfake Video Detection Using Recurrent Neural Networks’, in Proceedings of AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance. doi: 10.1109/AVSS.2018.8639163.

Hagen, S. et al. (2020) ‘Cross-platform mentions of the QAnon conspiracy theory’, Zenodo. doi: http://dx.doi.org/10.5281/ZENODO.3758479.

Hale, W. (2019) First Federal Legislation on Deepfakes Signed Into Law, JD Supra. Available at: https://www.jdsupra.com/legalnews/first-federal-legislation-on-deepfakes-42346/ (Accessed: 12 March 2021).

Hall, H. K. (2018) ‘Deepfake Videos: When Seeing Isn’t Believing’, Catholic University Journal of Law and Technology, 27(1), pp. 51–76. Available at: https://scholarship.law.edu/jlt/vol27/iss1/4.

Halliday, M.A.K. 1985, *An introduction to functional grammar,* Edward Arnold, London.

Halpern, D. and Gibbs, J. (2013) ‘Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression’, Comput. Hum. Behav. doi: 10.1016/j.chb.2012.10.008.

Hao, K. (2020) The owner of WeChat thinks deepfakes could actually be good, MIT Technology Review. Available at: https://www.technologyreview.com/2020/07/28/1005692/china-tencent-wechat-ai-plan-says-deepfakes-good/ (Accessed: 1 March 2021).

Hartmann, K. and Giles, K. (2020) ‘The Next Generation of Cyber-Enabled Information Warfare’, in International Conference on Cyber Conflict, CYCON, pp. 233–250. doi: 10.23919/CyCon49761.2020.9131716.

Harwell, D. (2019) ‘Top AI researchers race to detect “deepfake” videos: “We are outgunned”’, Washington Post. Available at:

https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/ (Accessed: 1 March 2021).

Hegel, G.W.F. 1896, *Hegel's Philosophy of right,* Bell, London.

Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Hoven, J. V. D., Zicari, R.V. and Zwitter, A., (2017). “Will Democracy Survive Big Data and Artificial Intelligence?” Scientific American. Available at: https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence

Heydari, A. et al. (2019) ‘YouTube Chatter: Understanding Online Comments Discourse on Misinformative and Political YouTube Videos’, ArXiv.

Hussain, M. N. et al. (2018) ‘Analyzing disinformation and crowd manipulation tactics on youtube’, in Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, pp. 1092–1095. doi: 10.1109/ASONAM.2018.8508766.

Hussain, S. et al. (2020) ‘Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples’, arXiv:2002.12749 [cs]. Available at: http://arxiv.org/abs/2002.12749 (Accessed: 6 March 2021).

Hwang, T. (2018) Deepfakes won’t wreck politics this year even if politicians might, New Scientist. Available at: https://www.newscientist.com/article/2174590-deepfakes-wont-wreck-politics-this-year-even-if-politicians-might/ (Accessed: 1 March 2021).

Hwang, T. (2020) ‘Deconstructing the Disinformation War’, in. doi: 10.35650/md.2053.d.2020.

Ignatidou, S. et al. (2019) Deepfakes, shallowfakes and speech synthesis: tackling audiovisual manipulation, European Science-Media Hub.

Available at: https://sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/ (Accessed: 1 March 2021).

Isakov M., Gadepally V., Gettings, K. M. and Kinsy M. A., (2020). Survey of Attacks and Defenses on Edge-Deployed Neural Networks. Available at: https://arxiv.org/abs/1911.11932.

Jasanoff, S. and Simmet, H.R., (2017). "No funeral bells: Public reason in a ‘post-truth’ age", Social studies of science, vol. 47, no. 5, pp. 751-770.

Jurgenson, N. (2019) The Social Photo: On Photography and Social Media. Verso.

Kadir, S. A., Lokman, A. M. and Muhammad, M. (2018a) ‘Analysis of Laddering Downwards for Classification of Item and Category Based on Emotional Values in Political Video on YouTube’, in. doi: 10.1007/978-981-10-8612-0_67.

Kadir, S. A., Lokman, A. M. and Muhammad, M. (2018b) ‘Identification of positive and negative emotion towards political agenda videos posted on YouTube’, in. doi: 10.1007/978-981-10-8612-0_79.

Kalpokas, I. (2020) ‘Problematising reality: the promises and perils of synthetic media’, SN Social Sciences. doi: 10.1007/s43545-020-00010-8.

Karanicolas, M. (2020) The Countries Where Democracy Is Most Fragile Are Test Subjects for Platforms’ Content Moderation Policies, Slate Magazine. Available at: https://slate.com/technology/2020/11/global-south-facebook-misinformation-content-moderation-policies.html (Accessed: 28 February 2021).

Karras, T., Laine, S. and Aila, T. (2019) ‘A Style-Based Generator Architecture for Generative Adversarial Networks’, arXiv:1812.04948 [cs, stat]. Available at: http://arxiv.org/abs/1812.04948 (Accessed: 30 March 2021).

Kasirzadeh, A. and Smart, A. (2021) ‘The Use and Misuse of Counterfactuals in Ethical Machine Learning’, in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery (FAccT ’21), pp. 228–236. doi: 10.1145/3442188.3445886.

Katarya, R. and Lal, A. (2020) ‘A Study on Combating Emerging Threat of Deepfake Weaponization’, in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 485–490. doi: 10.1109/I-SMAC49090.2020.9243588.

Kaur, S., Kumar, P. and Kumaraguru, P. (2020) ‘Deepfakes: Temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory’, Journal of Electronic Imaging, 29(3). doi: 10.1117/1.JEI.29.3.033013.

Kawahata, Y. (2019) ‘Examination of Analysis Method of Opinion Distribution in News Media Transferred on Web’, in IES. doi: 10.1007/978-3-030-37442-6_14.

Khodabakhsh, A., Ramachandra, R. and Busch, C. (2019) ‘Subjective evaluation of media consumer vulnerability to fake audiovisual content’, in 2019 11th International Conference on Quality of Multimedia Experience, QoMEX 2019. doi: 10.1109/QoMEX.2019.8743316.

Kietzmann, J., Mills, A. J. and Plangger, K. (2020) ‘Deepfakes: perspectives on the future “reality” of advertising and branding’, International Journal of Advertising. doi: 10.1080/02650487.2020.1834211.

Knight, W. (2020) ‘Deepfakes Aren’t Very Good. Nor Are the Tools to Detect Them’, Wired. Available at: https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/ (Accessed: 1 March 2021).

Koh, P. W., Steinhardt, J. and Liang, P. (2018) ‘Stronger Data Poisoning Attacks Break Data Sanitization Defenses’, arXiv:1811.00741 [cs, stat]. Available at: http://arxiv.org/abs/1811.00741 (Accessed: 1 April 2021).

Kolotaev, Y. and Kollnig, K. (2020) ‘Perceptions of YouTube’s political influence’, ArXiv.

Korshunov, P. and Marcel, S. (2020) ‘Deepfake detection: humans vs. machines’, ArXiv.

Korshunov, P. and S. Marcel (2019) ‘Vulnerability assessment and detection of Deepfake videos’, in 2019 International Conference on Biometrics (ICB). 2019 International Conference on Biometrics (ICB), pp. 1–6. doi: 10.1109/ICB45273.2019.8987375.

Krafft, P. M. and Donovan, J. (2020) ‘Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign’. doi: 10.1080/10584609.2019.1686094.

Kubin, E. et al. (2021) ‘Personal experiences bridge moral and political divides better than facts’, Proceedings of the National Academy of Sciences of the United States of America, 118(6). doi: 10.1073/pnas.2008389118.

Kwok, A. O. J. and Koh, S. G. M. (2020) ‘Deepfake: a social construction of technology perspective’, Current Issues in Tourism. doi: 10.1080/13683500.2020.1738357.

Kwok, A.OJ., and Koh, S.G.M., (2020). Deepfake: a social construction of technology perspective, Current Issues in Tourism, DOI: 10.1080/13683500.2020.1738357

Landon-Murray, M., Mujkic, E. and Nussbaum, B. (2019) ‘Disinformation in Contemporary U.S. Foreign Policy: Impacts and Ethics in an Era of

Fake News, Social Media, and Artificial Intelligence’. doi: 10.1080/10999922.2019.1613832.

Larson, K., Kaleda, K. and Fenstermacher, L. (2019) ‘Applying cognitive psychology principles to the (dis)information environment: an examination of discourse comprehension, memory, and fusion of news articles’, in Defense + Commercial Sensing. doi: 10.1117/12.2522278.

LeCun, Y., Bengio, Y. and Hinton, G., (2015). "Deep learning", Nature (London), vol. 521, no. 7553, pp. 436-444.

Lee, Y. et al. (2021) ‘To Believe or Not to Believe: Framing Analysis of Content and Audience Response of Top 10 Deepfake Videos on YouTube’, Cyberpsychology, Behavior, and Social Networking, 24(3), pp. 153–158. doi: 10.1089/cyber.2020.0176.

Leiser, D., Duani, N. and Wagner-Egger, P. (2017) ‘The conspiratorial style in lay economic thinking’, PLOS ONE, 12(3), p. e0171238. doi: 10.1371/journal.pone.0171238.

Lewis, J. A. and Nelson, A. (2019) Trust Your Eyes? Deepfakes Policy Brief. Available at: https://www.csis.org/analysis/trust-your-eyes-deepfakes-policy-brief (Accessed: 1 March 2021).

Li, L. et al. (2020) ‘Advancing high fidelity identity swapping for forgery detection’, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5073–5082. doi: 10.1109/CVPR42600.2020.00512.

Li, X. et al. (2020) ‘Fighting Against Deepfake: Patchand Pair Convolutional Neural Networks (PPCNN)’, in The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020, pp. 88–89. doi: 10.1145/3366424.3382711.

Li, Y. (2018) ‘Deep Reinforcement Learning’, arXiv:1810.06339 [cs, stat]. Available at: http://arxiv.org/abs/1810.06339 (Accessed: 26 March 2021).

Li, Y. et al. (2020) ‘Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics’, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3204–3213. doi: 10.1109/CVPR42600.2020.00327.

Li, Y., Chang, M.C. and Lyu, S. (2018) ‘In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking’, in 2018 10th Ieee International Workshop on Information Forensics and Security (wifs).

Liu, Y. and Wu, Y. F. B. (2020) ‘FNED: A Deep Network for Fake News Early Detection on Social Media’, ACM Transactions on Information Systems, 38(3), p. 25:1-25:33. doi: 10.1145/3386253.

Lukito, J. (2020) ‘Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017’. doi: 10.1080/10584609.2019.1661889.

Lyu, S. (2020) ‘Deepfake detection: Current challenges and next steps’, in 2020 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2020. doi: 10.1109/ICMEW46912.2020.9105991.

Maddocks, S., (2020). "‘A Deepfake Porn Plot Intended to Silence Me’: exploring continuities between pornographic and ‘political’ deep fakes", Porn studies (Abingdon, UK), pp. 1-9.

Manor, I. (2019) ‘The Specter of Echo Chambers—Public Diplomacy in the Age of Disinformation’, in. doi: 10.1007/978-3-030-04405-3_5.

Marcus, G. (2018) The deepest problem with deep learning, Medium. Available at: https://medium.com/@GaryMarcus/the-deepest-problem-with-deep-learning-91c5991f5695 (Accessed: 28 February 2021).

McBeth, K. (2018) The Infocalypse is Coming, International Policy Digest. Available at: https://intpolicydigest.org/the-infocalypse-is-coming/ (Accessed: 8 March 2021).

Mejova, Y. and Srinivasan, P. (2012) ‘Political speech in social media streams: YouTube comments and Twitter posts’, in WebSci ’12. doi: 10.1145/2380718.2380744.

Mericle, M. E. (2020) ‘Citizen Science in the Digital Age: Rhetoric, Science, and Public Engagement’. doi: 10.1080/00335630.2020.1744819.

Metzger, M. J. and Flanagin, A. J. (2013) ‘Credibility and trust of information in online environments: The use of cognitive heuristics’, Journal of Pragmatics, 59, pp. 210–220. doi: 10.1016/j.pragma.2013.07.012.

Min-Yeong, L. E. E. (2020) ‘Deepfake as powerized algorithm and public opinion regulated by post-truth’, Study on The American Constitution, 31(1), pp. 199–241.

Mirsky, Y. and Lee, W. (2021) ‘The Creation and Detection of Deepfakes: A Survey’, ACM Computing Surveys, 54(1), p. 7:1-7:41. doi: 10.1145/3425780.

Möller, A. et al. (2019) ‘Exploring User Responses to Entertainment and Political Videos: An Automated Content Analysis of YouTube’. doi: 10.1177/0894439318779336.

Momeni, E. and Sageder, G. (2013) ‘An empirical analysis of characteristics of useful comments in social media’, in WebSci. doi: 10.1145/2464464.2464490.

Mulholland, E. et al. (2016) ‘Analysing Emotional Sentiment in People’s YouTube Channel Comments’, in ArtsIT/DLI. doi: 10.1007/978-3-319-55834-9_21.

Munari, A. Jean Piaget. *Prospects* 24**,** 311–327 (1994). https://doi.org/10.1007/BF02199023

Murphy, H. (2019) Cyber security companies race to combat ‘deepfake’ technology. Available at: https://www.ft.com/content/63cd4010-bfce-11e9-b350-db00d509634e (Accessed: 1 March 2021).

Nabokov, V.V. Strong Opinions. New York: Vintage Books, 1990.

Nassetta, J. and Gross, K. (2020) ‘State media warning labels can counteract the effects of foreign misinformation’, Harvard Kennedy School Misinformation Review. doi: 10.37016/mr-2020-45.

Nguyen, A. and Vu, H. (2019) ‘Testing popular news discourse on the “echo chamber” effect: Does political polarisation occur among those relying on social media as their primary politics news source?’, First Monday. doi: 10.5210/FM.V24I6.9632.

Nguyen, T. T., Nguyen, C.M., Nguyen, D.T., Nguyen, D.T. and Nahavandi, S., (2020). Deep Learning for Deepfakes Creation and Detection: A Survey. Available at: arXiv:1909.11573

Nickerson, R. (1998) ‘Confirmation Bias: A Ubiquitous Phenomenon in Many Guises’, Review of General Psychology, 2, pp. 175–220. doi: 10.1037/1089-2680.2.2.175.

Nimmo, B. et al. (2019) ‘#OperationFFS: Fake Face Swarm | Graphika’, MediaWell. Available at: https://mediawell.ssrc.org/2019/12/20/operationffs-fake-face-swarm-graphika/ (Accessed: 19 March 2021).

Nonnecke, B., (2019) ‘Anti-Deepfake Law in California Is Far Too Feeble’, Wired. Available at: https://www.wired.com/story/opinion-californias-anti-deepfake-law-is-far-too-feeble/ (Accessed: 1 March 2021).

O’Sullivan, D. (2018) Inside the Pentagon’s race against deepfake videos. Available at: https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/ (Accessed: 28 February 2021).

Padilla Castillo, G., Garcia Guarda, M. L. and Cerdan Alenda, V. (2020) ‘Digital moral literacy for the detection of deepfakes and audiovisual fakes’, Cic-Cuadernos De Informacion Y Comunicacion, 25, pp. 165–181. doi: 10.5209/ciyc.68762.

Pan, D. et al. (2020) ‘Deepfake Detection through Deep Learning’, in 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT). 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), pp. 134–143. doi: 10.1109/BDCAT50828.2020.00001.

Pantserev, K. A. (2020) ‘The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability’, in Jahankhani, H. et al. (eds) Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity. Cham: Springer International Publishing (Advanced Sciences and Technologies for Security Applications), pp. 37–55. doi: 10.1007/978-3-030-35746-7_3.

Papadamou, K. et al. (2020) ‘Understanding the Incel Community on YouTube’, ArXiv.

Paris, B and Donovan, J., 2019. "Deepfakes and Cheap Fakes." Data and Society, September 18, 2019. Available at: https://datasociety.net/library/deepfakes-and-cheap-fakes/

Paris, J. D., Britt (2019) Deepfakes Are Troubling. But So Are the “Cheapfakes” That Are Already Here., Slate Magazine. Available at: https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html (Accessed: 16 February 2021).

Pennycook, G., Bear, A., et al. (2020) ‘The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings’, Management Science, 66(11), pp. 4944–4957. doi: 10.1287/mnsc.2019.3478.

Pennycook, G., Cannon, T. D. and Rand, D. G. (2018) ‘Prior Exposure Increases Perceived Accuracy of Fake News’, Journal of experimental psychology. General. doi: 10.1037/xge0000465.

Pennycook, G., McPhetres, J., et al. (2020) ‘Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention’. PsyArXiv. doi: 10.31234/osf.io/uhbk9.

Pinto-Coelho, Z., Carvalho, A. and Seixas, E. C. (2019) ‘News discourse and readers’ comments: Expanding the range of citizenship positions?’ doi: 10.1177/1464884917707595.

Plebe, A. and Grasso, G., (2019). ‘The Unbearable Shallow Understanding of Deep Learning’ Minds and Machines 29(4): pp. 515-553.

Polyakova, A. and Fried, D. (2019) ‘Europe is starting to tackle disinformation. The US is lagging.’, Brookings, 20 June. Available at: https://www.brookings.edu/blog/order-from-chaos/2019/06/20/europe-is-starting-to-tackle-disinformation-the-us-is-lagging/ (Accessed: 28 February 2021).

Porup, J. M. (2021) Deepfake videos: How and why they work — and what is at risk, CSO Online. Available at: https://www.csoonline.com/article/3293002/deepfake-videos-how-and-why-they-work.html (Accessed: 4 June 2021).

Prabhu, V. U. and Birhane, A. (2020) ‘Large image datasets: A pyrrhic win for computer vision?’. Available at: http://arxiv.org/abs/2006.16923 (Accessed: 26 March 2021).

Pu, J. et al. (2020) ‘NoiseScope: Detecting Deepfake Images in a Blind Setting’, in Annual Computer Security Applications Conference. New York, NY, USA: Association for Computing Machinery (ACSAC ’20), pp. 913–927. doi: 10.1145/3427228.3427285.

Quilty-Harper, C. (2019) Anyone can now play with sophisticated AIs thanks to a desktop app, New Scientist. Available at: https://www.newscientist.com/article/2205794-anyone-can-now-pla y-with-sophisticated-ais-thanks-to-a-desktop-app/ (Accessed: 1 March 2021).

Redondo, R. and Gibert, J. (2020) ‘Extended Labeled Faces in-the-Wild (ELFW): Augmenting Classes for Face Segmentation’, arXiv:2006.13980 [cs]. Available at: http://arxiv.org/abs/2006.13980 (Accessed: 28 February 2021).

Reilly, I. (2012) ‘Satirical Fake News and/as American Political Discourse’. doi: 10.1111/J.1542-734X.2012.00812.X.

Robertson, S. et al. (2013) ‘Political discourse on social networking sites: Sentiment, in-group/out-group orientation and rationality’, Inf. Polity. doi: 10.3233/IP-130303.

Rubin, V. L. (2019) ‘Disinformation and misinformation triangle’, J. Documentation. doi: 10.1108/jd-12-2018-0209.

Rudner, T. G. J. and Toner, H. (2021a) ‘Key Concepts in AI Safety: An Overview’, Center for Security and Emerging Technology. Available at: https://cset.georgetown.edu/research/key-concepts-in-ai-safety-an-overview/ (Accessed: 21 March 2021).

Rudner, T. G. J. and Toner, H. (2021b) ‘Key Concepts in AI Safety: Robustness and Adversarial Examples’, Center for Security and Emerging Technology. Available at: https://cset.georgetown.edu/research/key-concepts-in-ai-safety-robustness-and-adversarial-examples/ (Accessed: 21 March 2021).

Ryan, A. and Hii, A. (2019) Disinformation takes on a new face: ‘Deepfakes’ and the current legal landscape | Lexology. Available at: https://www.lexology.com/library/detail.aspx?g=757d4839-4281-4d84-b7be-016daaf8c378 (Accessed: 28 February 2021).

Rymes, B. (2012) ‘Recontextualizing YouTube: From Macro—Micro to Mass-Mediated Communicative Repertoires’, Anthropology and Education Quarterly, 43(2), pp. 214–227. Available at:

http://www.jstor.org/stable/23249786 (Accessed: 26 February 2021).

Schapiro, Z. (2020) ‘DEEP FAKES Accountability Act: Overbroad and Ineffective – Intellectual Property and Technology Forum’. Available at: https://bciptf.org/2020/04/deepfakes-accountability-act/ (Accessed: 4 March 2021).

Schick, N. (2020a) Cheapfakes did more political damage in 2020 than deepfakes, MIT Technology Review. Available at: https://www.technologyreview.com/2020/12/22/1015442/cheapfakes-more-political-damage-2020-election-than-deepfakes/ (Accessed: 15 March 2021).

Schick, N., (2020b). Deep Fakes and the Infocalypse: What You Urgently Need To Know. Monoray: London.

Schmidhuber, J., 2015. "Deep learning in neural networks: An overview", Neural networks, vol. 61, pp. 85-117.

Schultes, P., Dorner, V. and Lehner, F. (2013) ‘Leave a Comment! An In-Depth Analysis of User Comments on YouTube’, in Wirtschaftsinformatik.

Schulz, J. (2020) The Deepfake iPhone Apps Are Here, Lawfare. Available at: https://www.lawfareblog.com/deepfake-iphone-apps-are-here (Accessed: 15 March 2021).

Shah, D.V., Culver, K.B., Hanna, A., Macafee, T. and Yang, J., (2015). "Computational Approaches to Online Political Expression: Rediscovering a Science of the Social" in Coleman, S. and Freelon, D., Handbook of digital politics. Edward Elgar Publishing: Cheltenham.

Shelley, P.B. & Cook, A.S. 1890, *A defense of poetry,* Ginn, Boston, Mass.

Shu, K., Bhattacharjee, A., et al. (2020) ‘Combating disinformation in a social media age’, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. doi: 10.1002/widm.1385.

Shu, K., Wang, S., et al. (2020) ‘Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements’, ArXiv. doi: 10.1007/978-3-030-42699-6_1.

Simonite, T. (2020) ‘Cheap, Easy Deepfakes Are Getting Closer to the Real Thing’, Wired. Available at: https://www.wired.com/story/cheap-easy-deepfakes-closer-real-thing (Accessed: 1 March 2021).

Song L., Sehwag V., Bhagoji A. N. and Mittal P., (2020). A Critical Evaluation of Open-World Machine Learning. In: ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Available at: https://arxiv.org/abs/2007.04391.

Spelda, P. and Stritecky, V. (2021) ‘What Can Artificial Intelligence Do for Scientific Realism?’, Axiomathes, 31(1), pp. 85–104. doi: 10.1007/s10516-020-09480-0.

Spicer, R. N. (2018) ‘Lies, Damn Lies, Alternative Facts, Fake News, Propaganda, Pinocchios, Pants on Fire, Disinformation, Misinformation, Post-Truth, Data, and Statistics’, in. doi: 10.1007/978-3-319-69820-5_1.

Starbird, K. (2019) ‘Disinformation’s spread: bots, trolls and all of us’, Nature. doi: 10.1038/d41586-019-02235-x.

Ştefăniţă, O., Corbu, N. and Buturoiu, R. (2018) ‘Fake News and the Third-Person Effect: They Are More Influenced Than Me and You’, in. doi: 10.24193/JMR.32.1.

Stolk, L. (2020) ‘If deepfakes are a threat, this is it: A feminist perspective on the impact of deepfake pornography.’, The Hmm, 5 June. Available at: https://thehmm.nl/if-deepfakes-are-a-threat-this-is-it/ (Accessed: 1 March 2021).

Suratkar, S. et al. (2020) ‘Exposing DeepFakes Using Convolutional Neural Networks and Transfer Learning Approaches’, in 2020 IEEE 17th India Council International Conference (INDICON). 2020 IEEE

17th India Council International Conference (INDICON), pp. 1–8. doi: 10.1109/INDICON49873.2020.9342252.

Susarla, A., Oh, J. and Tan, Y. (2012) ‘Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube’, Inf. Syst. Res. doi: 10.1287/isre.1100.0339.

Tammekänd, J., Thomas, J. and Peterson, K., (2020). Deepfakes 2020: The Tipping Point. Sentinel. Available at: https://thesentinel.ai/report

Tandoc, E. C., Lim, D. and Ling, R. (2020) ‘Diffusion of disinformation: How social media users respond to fake news and why’. doi: 10.1177/1464884919868325.

Taylor, B. C. (2020) ‘Defending the state from digital Deceit: the reflexive securitization of deepfake’, Critical Studies in Media Communication. doi: 10.1080/15295036.2020.1833058.

Thies, J. et al. (2018) ‘HeadOn: Real-time Reenactment of Human Portrait Videos’, ACM Transactions on Graphics 2018 (TOG).

Thies, J. et al. (2020) ‘Face2Face: Real-time Face Capture and Reenactment of RGB Videos’, arXiv:2007.14808 [cs]. Available at: http://arxiv.org/abs/2007.14808 (Accessed: 1 March 2021).

Thornhill, J. (2019) New tools are evolving in the fight against deepfakes. Available at: https://www.ft.com/content/4183b400-f960-11e9-98fd-4d6c20050229 (Accessed: 1 March 2021).

Thornhill, J. (2020) The astonishingly good but predictably bad AI program. Available at: https://www.ft.com/content/51f1bb71-ce93-4529-9486-fec96ab3dc4d (Accessed: 1 March 2021).

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J., 2020. "Deepfakes and beyond: A Survey of face manipulation and fake detection", Information fusion, vol. 64, pp. 131-148.

Toner, H. (2020) ‘GPT-2 Kickstarted the Conversation About Publication Norms in the AI Research Community’, Center for Security and Emerging Technology, 1 May. Available at: https://cset.georgetown.edu/article/gpt-2-kickstarted-the-conversation-about-publication-norms-in-the-ai-research-community/ (Accessed: 1 March 2021).

Tong, X. et al. (2020) ‘An Overview of Deepfake: The Sword of Damocles in AI’, 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL). doi: 10.1109/CVIDL51233.2020.00-88.

Tredinnick, L. and Laybats, C. (2019) ‘Reality filters: Disinformation and fake news’. doi: 10.1177/0266382119874267.

Urbani, S. (2020) AI won’t solve the problem of moderating audiovisual media, First Draft. Available at: https://firstdraftnews.org:443/latest/ai-moderating-media/ (Accessed: 28 February 2021).

Vaccari, C. and Chadwick, A. (2020) ‘Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News’. doi: 10.1177/2056305120903408.

Van Dijk, T.A., 1993. Principles of critical discourse analysis. *Discourse & society*, *4*(2), pp.249-283.

Veletsianos, G. et al. (2018) ‘Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading, and moderation on YouTube TED talk comments’, PloS one. doi: 10.1371/journal.pone.0197331.

Venkataramakrishnan, S. (2019) Can you believe your eyes? How deepfakes are coming for politics. Available at: https://www.ft.com/content/4bf4277c-f527-11e9-a79c-bc9acae3b654 (Accessed: 1 March 2021).

Venkataramakrishnan, S. (2020) After deepfakes, a new frontier of AI trickery: fake faces. Available at:

https://www.ft.com/content/b50d22ec-db98-4891-86da-af34f06d1cb1 (Accessed: 1 March 2021).

Verdoliva, L. (2020) 'Media Forensics and DeepFakes: An Overview', IEEE Journal of Selected Topics in Signal Processing, 14(5), pp. 910–932. doi: 10.1109/JSTSP.2020.3002101.

Villasenor, J. (2019a) 'Artificial intelligence, deepfakes, and the uncertain future of truth', Brookings, 14 February. Available at: https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/ (Accessed: 1 March 2021).

Villasenor, J. (2019b) 'Deepfakes, social media, and the 2020 election', Brookings, 3 June. Available at: https://www.brookings.edu/blog/techtank/2019/06/03/deepfakes-social-media-and-the-2020-election/ (Accessed: 1 March 2021).

Vincent, J. (2019) This website uses AI to turn your selfies into haunted classical portraits, The Verge. Available at: https://www.theverge.com/tldr/2019/7/22/20703810/ai-classical-portrait-apps-selfie-web-transformation (Accessed: 1 March 2021).

Vosoughi, S., Roy, D. and Aral, S. (2018) 'The spread of true and false news online', Science. doi: 10.1126/science.aap9559.

Wahl-Jorgensen, K. and Carlson, M. (2021) 'Conjecturing fearful futures: Journalistic discourses on deepfakes', Journalism Practice. Available at: https://orca.cf.ac.uk/139988/ (Accessed: 7 April 2021).

Wang, Y. et al. (2020) 'Understanding the Use of Fauxtography on Social Media', ArXiv.

Weiss, G. (2020) The new struggle for truth in the era of deepfakes, The Strategist. Available at: https://www.aspistrategist.org.au/the-new-struggle-for-truth-in-the-era-of-deepfakes/ (Accessed: 28 February 2021).

Westerlund, M. (2019) 'The Emergence of Deepfake Technology: A Review', in.

Whyte, C., (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge, Journal of Cyber Policy, 5:2, 199-217, DOI: 10.1080/23738871.2020.1797135

Wodak, R. (2003) ‘Critical Discourse Analysis’. doi: 10.1002/9781118584194.CH22.

Wolverton, C. and Stevens, D. (2020) ‘The impact of personality in recognizing disinformation’, Online Inf. Rev. doi: 10.1108/oir-04-2019-0115.

Woolley, S., 2020. The Reality Game: A gripping investigation into deepfake videos, the next wave of fake news and what it means for democracy. Endeavour: London.

Yadlin-Segal, A. and 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), pp. 36–51. doi: 10.1177/1354856520923963.

Yampolskiy, R. and Fox, J. (2013) ‘Safety Engineering for Artificial General Intelligence’, Topoi, 32(2), pp. 217–226. doi: 10.1007/s11245-012-9128-9.

Zeller, R. (2019) Why We Released Grover, The Gradient. Available at: https://thegradient.pub/why-we-released-grover/ (Accessed: 20 March 2021).

Zhang, Z. and Liu, Q. (2020) Detect Video Forgery by Performing Transfer Learning on Deep Neural Network.

Zi, B. et al. (2020) ‘WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection’, ACM Multimedia. doi: 10.1145/3394171.3413769.

APPENDIX A: List of Terms for Moral Attribution

Good 'Keyword'	'# of Posts'	Bad 'Keyword'	'# of Posts'
'good'	419	'bad'	189
'funny'	273	'scary'	83
'great'	176	'dangerous.'	57
'nice'	98	'creepy'	34
'hilarious'	67	'terrible'	32
'kind'	63	'evil'	21
'perfect'	45	'awful'	20
'happy'	34	'scary'	13
'fine'	19	'ashamed'	13
'calm'	17	'angry'	10
'fair'	13	'worried'	10
'fantastic'	12	'hurt'	8
'proud'	11	'ill'	8
'lovely'	7	'upset'	8
'proud'	6	'lazy'	6
'successful'	5	'nasty'	6
'silly'	5	'annoyed'	5
'lucky'	5	'cruel'	5
'friendly'	5	'tired'	5
'brave'	5	'foolish'	3
'amused'	5	'obnoxious'	3

'determined'	4	'tense'	3
'thoughtful'	3	'arrogant'	2
'lucky'	3	'bored'	2
'jolly'	3	'dull'	2
'healthy'	3	'jealous.'	2
'obedient'	2	'clumsy'	1
'faithful'	2	'defeated'	1
'excited.'	2	'depressed'	1
'eager'	2	'disgusted'	1
'wonderful'	1	'disturbed'	1
'witty'	1	'envious'	1
'splendid'	1	'frightened'	1
'smiling'	1	'hungry'	1
'relieved'	1	'weary'	1
'lively'	1	-	-
'gentle'	1	-	-
'energetic'	1	-	-
'encouraged'	1	-	-
'comfortable'	1	-	-

APPENDIX B: Frequently Occurring Terms (Top 100)

You Won’t Believe What Obama Says In This Video! 😉	Deepfake Queen: 2020 Alternative Christmas Message	Dictators - Kim Jong-Un	Dictators - Vladimir Putin
957;"news"	279;"fake"	42;"people"	34;"Putin"
848;"Obama"	276;"real"	39;"fake"	28;"fake"
787;"fake"	256;"people"	39;"real"	25;"people"

598;"video"	230;"Queen"	38;"democracy"	24;"Russia"
411;"people"	153;"voice"	32;"Kim"	18;"democracy"
395;"sources"	147;"funny"	25;"deepfake"	15;"video"
313;"trusted"	146;"video"	20;"China"	14;"real"
309;"woke"	129;"Channel"	20;"video"	13;"America"
307;"buzzfeed"	120;"good"	19;"country"	11;"countries"
305;"know"	104;"better"	19;"North"	11;"dictator"
260;"real"	101;"technology"	19;"right"	11;"different"
250;"voice"	101;"time"	19;"system"	11;"message"
243;"believe"	98;"years"	18;"point"	11;"propaganda"
235;"good"	92;"deepfake"	17;"Korea"	10;"country"
232;"Trump"	92;"now"	17;"really"	10;"good"
229;"trust"	84;"channel"	17;"voting"	10;"Russian"
223;"tell"	80;"deep"	16;"speech"	9;"China"
219;"bitches"	73;"year"	15;"deep"	9;"deepfake"
217;"look"	72;"believe"	15;"democratic"	9;"English"
201;"source"	71;"point"	14;"free"	9;"point"
194;"Stay"	70;"never"	14;"way"	9;"Trump"
189;"president"	70;"right"	14;"world"	8;"end"
181;"lol"	69;"media"	13;"government"	8;"great"
175;"Jordan"	69;"world"	13;"nothing"	8;"love"
162;"obama"	68;"way"	13;"republic"	8;"maybe"
160;"now"	67;"actually"	13;"scary"	8;"mean"
153;"time"	64;"message"	13;"shit"	8;"president"
146;"media"	64;"news"	12;"every"	8;"republic"
145;"mouth"	64;"speech"	12;"Korean"	8;"right"
134;"something"	61;"Christmas"	12;"political"	8;"speak"
132;"Peele"	60;"old"	12;"power"	8;"take"
130;"BuzzFeed"	59;"things"	12;"thing"	8;"using"
128;"things"	56;"need"	12;"thought"	8;"voice"
127;"technology"	54;"joke"	12;"want"	8;"way"
126;"thing"	52;"God"	11;"Democracy"	7;"bad"
123;"anything"	50;"bit"	11;"end"	7;"care"

121;"bad"	50;"British"	11;"ever"	7;"funny"
121;"Fake"	50;"disrespectful"	11;"good"	7;"man"
120;"CNN"	50;"watch"	11;"life"	7;"never"
119;"better"	48;"bad"	11;"now"	7;"news"
119;"point"	48;"everything"	11;"rule"	7;"next"
118;"never"	46;"hope"	11;"war"	7;"now"
117;"News"	46;"van"	10;"anyone"	6;"agree"
116;"deep"	44;"far"	10;"freedom"	6;"anyone"
114;"want"	44;"love"	10;"keep"	6;"course"
113;"face"	44;"truth"	10;"party"	6;"deep"
110;"videos"	44;"want"	10;"trying"	6;"elections"
109;"thought"	43;"always"	10;"well"	6;"fall"
103;"sound"	43;"person"	9;"actually"	6;"first"
101;"man"	42;"life"	9;"Americans"	6;"Good"
100;"internet"	42;"show"	9;"footage"	6;"law"
99;"truth"	42;"Trump"	9;"look"	6;"political"
98;"made"	41;"anything"	9;"never"	6;"time"
96;"trump"	41;"deepfakes"	9;"Nothing"	6;"vote"
94;"someone"	41;"lol"	9;"propaganda"	6;"war"
94;"years"	41;"nothing"	9;"support"	6;"without"
93;"world"	41;"probably"	9;"technology"	6;"years"
92;"Fox"	40;"shit"	9;"things"	5;"already"
92;"great"	39;"country"	9;"work"	5;"Americans"
91;"watch"	39;"family"	9;"wrong"	5;"another"
89;"President"	38;"BBC"	8;"agree"	5;"attention"
88;"funny"	38;"money"	8;"back"	5;"become"
87;"anyone"	37;"enough"	8;"dictatorship"	5;"better"
87;"guy"	37;"watching"	8;"election"	5;"Biden"
87;"mean"	36;"great"	8;"first"	5;"dangerous"
87;"talking"	36;"mean"	8;"law"	5;"everyone"
86;"new"	36;"since"	8;"literally"	5;"evil"
83;"love"	36;"trust"	8;"live"	5;"face"
82;"use"	35;"CGI"	8;"man"	5;"find"

80;"everything"	35;"comedy"	8;"president"	5;"foreign"
80;"true"	35;"dead"	8;"social"	5;"God"
79;"ever"	35;"face"	8;"something"	5;"going"
79;"many"	35;"find"	8;"stop"	5;"government"
79;"trying"	34;"new"	8;"style"	5;"hope"
78;"back"	34;"tech"	8;"take"	5;"idea"
78;"used"	33;"comment"	8;"true"	5;"leader"
75;"Wow"	33;"fact"	8;"U.S"	5;"less"
75;"wrong"	33;"laugh"	7;"America"	5;"lol"
71;"pretty"	33;"someone"	7;"communists"	5;"lot"
69;"words"	33;"true"	7;"countries"	5;"matter"
68;"done"	32;"Deep"	7;"Deepfake"	5;"may"
67;"sounds"	32;"end"	7;"legislation"	5;"nice"
66;"joke"	32;"Fake"	7;"little"	5;"others"
65;"America"	32;"last"	7;"lol"	5;"problem"
64;"public"	32;"little"	7;"matter"	5;"RepresentUs"
63;"nothing"	32;"making"	7;"message"	5;"show"
62;"left"	31;"around"	7;"military"	5;"stop"
61;"information"	31;"best"	7;"public"	5;"technology"
61;"person"	31;"first"	7;"RepresentUs"	5;"threat"
61;"Yeah"	31;"fun"	7;"threat"	5;"two"
59;"lmao"	31;"TV"	7;"time"	5;"understand"
59;"propaganda"	30;"behind"	7;"Western"	5;"use"
58;"away"	30;"part"	6;"American"	5;"used"
58;"best"	30;"trying"	6;"another"	5;"Vote"
57;"CGI"	30;"videos"	6;"anything"	5;"whole"
57;"Hillary"	29;"actual"	6;"better"	5;"word"
56;"long"	29;"ago"	6;"Chinese"	4;"actually"
55;"lie"	29;"Andrew"	6;"class"	4;"answer"
55;"money"	29;"day"	6;"dangerous"	4;"anything"
54;"black"	28;"dance"	6;"data"	4;"believe"