Fordham Intellectual Property, Media and Entertainment Law Journal

Volume 31 *XXXI*
Number 4

Article 3

2021

Ocularcentrism and Deepfakes: Should Seeing Be Believing?

Katrina G. Geddes
New York University, kgg283@nyu.edu

Ocularcentrism and Deepfakes: Should Seeing Be Believing?

Cover Page Footnote

J.S.D. Candidate, New York University, School of Law; Fellow, Engelberg Center on Innovation Law & Policy, NYU Law; Fellow, Information Law Institute, NYU Law.

For helpful comments and insights on this piece, I want to thank the ILI fellows, Professor Barton Beebe, Stefano Martiniani, and Felix Barber.

Ocularcentrism and Deepfakes: Should Seeing Be Believing?

Katrina Geddes*

The pernicious effects of misinformation were starkly exposed on January 6, 2021, when a violent mob of protestors stormed the nation's capital, fueled by false claims of election fraud. As policymakers wrestle with various proposals to curb misinformation online, this Article highlights one of the root causes of our vulnerability to misinformation, specifically, the epistemological prioritization of sight above all other senses ("ocularcentrism"). The increasing ubiquity of so-called "deepfakes"—hyperrealistic, digitally altered videos of events that never occurred—has further exposed the vulnerabilities of an ocularcentric society, in which technology-mediated sight is synonymous with knowledge. This Article traces the evolution of visual manipulation technologies that have exploited ocularcentrism and evaluates different means of addressing the issues raised by deepfakes, including the use of copyright law.

* J.S.D. Candidate, New York University, School of Law; Fellow, Engelberg Center on Innovation Law & Policy, NYU Law; Fellow, Information Law Institute, NYU Law. For helpful comments and insights on this piece, I want to thank the ILI fellows, Professor Barton Beebe, Stefano Martiniani, and Felix Barber.

INTRODUCTION

Since its unholy beginnings in pornography, deepfake technology has, understandably, been the subject of widespread criticism and outrage. Broadly speaking, a "deepfake" is a hyperrealistic video that has been digitally altered to depict an event or events that never occurred.[1] At the individual level, pornographic and other harmful kinds of deepfakes can cause significant psychological and reputational harm.[2] At the collective level, the dissemination of deepfakes affects our ability to differentiate authentic from inauthentic content, rendering us more vulnerable to misinformation.[3] This effect, however, is not limited to deepfakes; photographs and videos have long been vulnerable to manipulation. The problem, then, is not deepfakes *per se*, but our *uncritical* and *disproportionate* reliance on technology-mediated sight, and our insistence that *seeing is believing*. The initial purpose of this Article is to understand the historical persistence of "ocularcentrism," or the epistemological prioritization of sight above other human senses,[4] and, secondly, to situate deepfakes within this social history—do deepfakes represent the *limit* of our tolerance for visual manipulation, and if so, why? Do they truly threaten visual truth in a way that earlier

---

1 Mika Westerlund, *The Emergence of Deepfake Technology: A Review*, 9 TECH. INNOVATION MGMT. 39, 40 (2019).

2 *See, e.g.*, Anne Pechenik Gieseke, *"The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography*, 73 VAND. L. REV. 1479, 1479 (2020).

3 *See, e.g.*, Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J. L. TECH. 1, 2 (2020); Robert Chesney & Danielle Keats Citron, *21st Century-Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security*, 78 MD. L. REV. 882, 883–84 (2019); Holly Kathleen Hall, *Deepfake Videos: When Seeing Isn't Believing*, 27 CATH. U. J. L. & TECH. 51, 52 (2018).

4 *See, e.g.*, Jenni Lauwrens, *Can You See What I Mean? An Exploration of the Limits of Vision in Anti-Ocularcentric Contemporary Art*, 47 DE ARTE 26, 28 (2012); Martin Jay, *Scopic Regimes of Modernity, in* VISION AND VISUALITY 3, 3 (Hal Foster ed., 1988), MARTIN JAY, DOWNCAST EYES: THE DENIGRATION OF VISION IN TWENTIETH-CENTURY FRENCH THOUGHT 3 (1993).

technologies have not? If so, should we *abandon* ocularcentrism—or cling to the credibility of visual evidence?

To date, existing scholarship on deepfakes has failed to differentiate between, and tailor solutions for, the *individual* and *collective* harms associated with their dissemination. Such tailoring is needed to preserve the substantial utility that deepfakes offer. Deepfake audio recreated the speech that John F. Kennedy intended to deliver shortly before his assassination, using recordings of 831 speeches he delivered in his lifetime, and offering hope to patients who have lost their voices to illness.[5] Researchers have used deepfake technology to create animated, photorealistic avatars of deceased persons and portrait subjects.[6] Museum visitors can interact with life-size deepfakes of long-dead artists, constructed from archival footage.[7] Deepfake technology can be used to anonymize vulnerable sources,[8] generate multilingual voice petitions,[9] produce synthetic MRI images that protect patient privacy,[10] synthesize news

---

[5] *JFK Unsilenced*, CEREPROC, https://www.cereproc.com/en/jfkunsilenced [https://perma.cc/K7BB-3B4U].

[6] Egor Zakharov et al., *Few-Shot Adversarial Learning of Realistic Neural Talking Head Models, Computer Vision and Pattern Recognition*, 2019 IEEE/CVF INT'L CONFERENCE ON COMP. VISION 9459, 9459 (2019); Westerlund, *supra* note 1, at 41–43.

[7] Dami Lee, *Deepfake Salvador Dalí Takes Selfies with Museum Visitors*, THE VERGE (May 10, 2019, 8:50 AM), https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum [https://perma.cc/9M8P-T975].

[8] Rebecca Heilweil, "*How Deepfakes Could Actually Do Some Good*," VOX (June 29, 2020, 11:10 AM), https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-chechnya [https://perma.cc/4THG-ULMQ].

[9] Guy Davies, *David Beckham 'Speaks' 9 Languages for New Campaign to End Malaria*, ABC NEWS (Apr. 9, 2019, 12:51 PM), https://abcnews.go.com/International/david-beckham-speaks-languages-campaign-end-malaria/story?id=62270227 [https://perma.cc/XQ7Q-C3DQ]; *see also* Kim Lyons, *An Indian Politician Used AI to Translate His Speech into Other Languages to Reach More Voters*, THE VERGE (Feb. 18, 2020, 5:35 PM), https://www.theverge.com/2020/2/18/21142782/india-politician-deepfakes-ai-elections [https://perma.cc/Y8DM-N7PZ].

[10] Hoo-Chang Shin et al., Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks 1 (Sept. 13, 2018) (unpublished manuscript) (available at https://arxiv.org/abs/1807.10225.

reports,[11] improve video-game graphics,[12] reverse the aging process,[13] re-animate old photos,[14] and elevate fanfiction.[15] If, like most forms of technology, deepfakes are capable of both beneficial and harmful use, how should the technology be regulated to maximize its utility and minimize its harm?

---

[11] Simon Chandler, *Reuters Uses AI to Prototype First Ever Automated Video Reports*, FORBES (Feb. 7, 2020, 8:30 AM), https://www.forbes.com/sites/simonchandler/2020/02/07/reuters-uses-ai-to-prototype-first-ever-automated-video-reports/?sh=35d9aa87a2a7 [https://perma.cc/PG2G-KX2S].

[12] James Vincent, *Nvidia Has Created the First Video Game Demo Using AI-Generated Graphics*, THE VERGE (Dec. 3, 2018, 8:00 AM), https://www.theverge.com/2018/12/3/18121198/ai-generated-video-game-graphics-nvidia-driving-demo-neurips [https://perma.cc/F864-AHVN].

[13] Jacob Kastrenakes, *When Diplo and The Strokes Need a Deepfake, They Go to This Guy*, THE VERGE (Mar. 4, 2020, 12:30 PM), https://www.theverge.com/2020/3/4/21164607/the-fakening-deepfakes-strokes-diplo-memes-music-industry-elon-musk-jeff-bezos-star-trek [https://perma.cc/ZJ9Q-YLGX]. *See also* The Strokes, *The Strokes – Bad Decisions (Official Video)*, YOUTUBE (Feb. 18, 2020), https://www.youtube.com/watch?v=5fbZTnZDvPA&t=9s [https://perma.cc/9BVU-B8FY].

[14] Alex Hern, *Deep Nostalgia: 'Creepy' New Service Uses AI to Animate Old Family Photos,* THE GUARDIAN (Mar. 1, 2021), https://www.theguardian.com/technology/2021/mar/01/deep-nostalgia-creepy-new-service-ai-animate-old-family-photos [https://perma.cc/BG5L-5UG3]

[15] *See, e.g.*, Jay Peters, *This Disturbingly Realistic Deepfake Puts Jeff Bezos and Elon Musk in a Star Trek Episode*, THE VERGE (Feb. 20, 2020, 3:35 PM), https://www.theverge.com/tldr/2020/2/20/21145826/deepfake-jeff-bezos-elon-musk-alien-star-trek-the-cage-amazon-tesla [https://perma.cc/6ZXE-NG43]; Chaim Gartenberg, *Deepfake Edits Have Put Harrison Ford into Solo: A Star Wars Story, for Better or for Worse*, THE VERGE (Oct. 17, 2018, 3:37 PM), https://www.theverge.com/2018/10/17/17990162/deepfake-edits-harrison-ford-han-solo-a-star-wars-story-alden-ehrenreich [https://perma.cc/3WK6-8W6B]; KC Ifeanyi, *According to this Deepfake, Neo Taking the Blue Pill in 'The Matrix' Would've Been 'Office Space'*, FAST COMPANY (Feb. 19, 2020), https://www.fastcompany.com/90465563/according-to-this-deep-fake-neo-taking-the-blue-pill-in-the-matrix-wouldve-been-office-space [https://perma.cc/2VEN-2GVY]; Lee Moran, *Jon Snow Says Sorry for 'Game of Thrones' Finale in Convincing Deepfake*, HUFFINGTON POST (June 14, 2019, 9:18 AM), https://www.huffpost.com/entry/game-of-thrones-deepfake-jon-snow_n_5d038623e4b0985c419bded2 [https://perma.cc/4F59-XKHQ]; Zack Sharf, *The Shining Deepfake Goes Viral with Jim Carrey Starring as Jack Torrance*, INDIEWIRE (July 10, 2019, 12:14 PM), https://www.indiewire.com/2019/07/the-shining-jim-carrey-deepfake-video-viral-1202156857/ [https://perma.cc/JT8B-XZND]; Sven Charleer, *Family Fun with Deepfakes. Or How I Got My Wife onto The Tonight Show*, MEDIUM (Feb. 2, 2018), https://towardsdatascience.com/family-fun-with-deepfakes-or-how-i-got-my-wife-onto-the-tonight-show-a4454775c011 [https://perma.cc/X9YP-VGJL].

This Article will explore this question through the lens of copyright law and policy. The creation of deepfakes depends heavily on access to, and manipulation of, audiovisual content—much of which is protected by copyright law. Accordingly, copyright represents a natural lens through which to evaluate the unique social issues raised by the creation and dissemination of deepfakes. Part I will explain the technical process by which deepfakes are created and evaluate whether a deepfake video would constitute transformative fair use.[16] Part II will discuss both the *individual* and *collective* harms generated by the dissemination of deepfakes, including the erosion of our ability to differentiate authentic from inauthentic content. It will interrogate the historical basis for the normative claim that *seeing is believing* and problematize the role of ocularcentrism in promoting both surveillance and misinformation. Part III will evaluate a variety of legal and regulatory measures that have been proposed to address the harms caused by deepfakes. The Conclusion will summarize the discussion contained within the Article and provide final thoughts.

I. ARE DEEPFAKES PROTECTED BY FAIR USE?

For now, this question remains theoretical; no judicial proceeding has yet determined whether fair use protects the creators of deepfakes from copyright infringement liability. So, the question becomes conditional: *should* deepfakes be protected by fair use? Would such protection be consistent with the evolution of fair use jurisprudence and the overarching policy objectives of the copyright regime? These are the questions that will be explored in this section.

A. Background

First, it is important to understand the technical process by which deepfakes are created. The term *deepfake*—a combination of "deep learning" and "fake"—generally refers to synthetic content

[16] The copyrightability of deepfakes as transformative fair uses would be consistent both with the long jurisprudential history of fair use, as well as copyright law's ostensible content neutrality (i.e., the availability of copyright protection should not depend on whether the work is a photograph of candy or an Impressionist painting).

created by an artificial neural network,[17] but the term has colloquially been used to describe a broad spectrum of hyperrealistic content.[18] At the sophisticated end of the spectrum, a recurrent neural network ("RNN") can generate synthetic video footage of an individual from an audio recording.[19] The process of mapping from a one-dimensional (audio) signal to a three-dimensional time-varying image is technically challenging, but bears substantial utility.[20] For example, an individual who is hearing-impaired could lip-read a synthetic video generated from over-the-phone audio.[21] Researchers from the University of Washington trained a RNN on seventeen hours of video footage of President Obama delivering 300 weekly addresses.[22] From this corpus of video footage, they extracted 1.9 million video frames.[23] For every output video frame, the RNN detects mouth landmarks (18 points along the outer and inner contours of the lip) to generate a sparse mouth shape.[24] The mouth shape and lower region of the face are given texture before the synthesized mouth region is blended into the target video.[25] The target video is then re-timed to ensure that the natural head motion matches the input audio.[26] Essentially, the RNN maps mouth shapes from raw audio to create synthetic footage that can be composited into the mouth region of a target video for photorealistic results.[27]

Another sophisticated technique for generating deepfakes is a generative adversarial network ("GAN"), which pairs a discriminative algorithm (which predicts a label, given certain features) and a

---

17 Yisroel Mirsky & Wenke Lee, *The Creation and Detection of Deepfakes: A Survey*, ACM COMPUTING SURVS., Jan. 2020, at 1.

18 BRITT PARIS & JOAN DONOVAN, DEEPFAKES AND CHEAPFAKES: THE MANIPULATION OF AUDIO AND VISUAL EVIDENCE 10–11 (2019).

19 Supasorn Suwajanakorn et al., *Synthesizing Obama: Learning Lip Sync from Audio*, ACM TRANSACTIONS ON GRAPHICS, July 2017, at 2.

20 *Id.* at 1.

21 *Id.*

22 *Id.* at 8.

23 *Id.*

24 *Id.* at 4.

25 *Id.* at 3.

26 *Id.*

27 The photorealism generated by the audio-to-shape (mouth shape) neural network can be observed by using a pixel difference map to compare the groundtruth video of the input audio and the input audio composited into the target video.

generative algorithm (which predicts features, given a certain label).[28] For example, a discriminative algorithm would try to predict whether a particular email should be classified as "spam" given its contents, whereas a generative algorithm would try to predict the features of an email that had already been classified as spam. Deepfakes are created by the interaction of these algorithms: the "generator" generates artificial images that *resemble* the images in the training set, and the "discriminator" evaluates these images for authenticity—whether they came from the training set or not.[29] As these algorithms interact, the generator learns to create sufficiently realistic images to fool the discriminator.[30]

A similar deep learning technique, known as Video Dialogue Replacement ("VDR"), was used to create a deepfake of Mark Zuckerberg discussing the profitability of personal data.[31] Artists Barnaby Francis and Daniel Howe created the deepfake using the proprietary algorithm of an Israeli technology start-up known as Canny AI.[32] Canny engineers trained their deep learning algorithm on a twenty-one second clip from the target video as well as video footage of the voice actor speaking, then reconstructed the frames in the target video to match the facial movements of the voice actor.[33] No audio recordings of Zuckerberg were used.[34]

At the other end of the deepfake spectrum, less sophisticated actors can create "cheap fakes," or lower-quality deepfakes, using

---

[28] Russell Spivak, *Deepfakes: The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 342–43 (2019).

[29] *Id.* at 343.

[30] HAI X. PHAM ET AL., GENERATIVE ADVERSARIAL TALKING HEAD: BRINGING PORTRAITS TO LIFE WITH A WEAKLY SUPERVISED NEURAL NETWORK 4 (2018).

[31] Bill Posters, *'Imagine This…'*, VIMEO (June 12, 2019, 9:30 AM), https://vimeo.com/341794473 [https://perma.cc/6SGD-UWKC].

[32] Samantha Cole, *This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies*, VICE (June 11, 2019, 2:25 PM), https://www.vice.com/en/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy [https://perma.cc/8MKA-7TD9].

[33] *Id.*

[34] A comparison of the deepfake with the original (altered) video is available here: Multimedia LIVE, *Artists Create Zuckerberg 'Deepfake' Video*, YOUTUBE (June 13, 2019), https://www.youtube.com/watch?v=cnUd0TpuoXI [https://perma.cc/283M-93DM].

consumer-grade software or simple video-editing techniques.[35] For example, Adobe After Effects and FakeApp were used to create a deepfake of President Obama delivering impersonated audio by Jordan Peele.[36] And recent cheap fakes that attracted significant attention were created using simple video editing techniques. A video of House Speaker Nancy Pelosi was slowed down to create the impression of slurred speech,[37] and a video of CNN correspondent Jim Acosta interacting with a White House intern was sped up to suggest physical assault.[38] In each of these cases, the cheap fake was widely circulated on social media platforms before its falsity was recognized.[39] The purpose of this technical summary is simply to highlight the breadth of digitally manipulated media that falls under the umbrella term "deepfake." For the purposes of evaluating the copyright issues raised by deepfakes, however, this Article will focus on two videos created using sophisticated deep learning techniques.

B. Kim Kardashian Deepfake

In 2019, the same artists who created the Mark Zuckerberg deepfake, Barnaby Francis and Daniel Howe, posted on YouTube a deepfake of Kim Kardashian describing the profitability of data

---

35 PARIS & DONOVAN, *supra* note 18, at 2; *see, e.g.*, Nic and Pancho, *Why Chihuahuas Don't Run on the Snow?*, YOUTUBE (Jan. 7, 2014), https://www.youtube.com/watch?v=JDaLg7G8rH0 [https://perma.cc/8S3W-559D]; Nic and Pancho, *Is Pancho Alive? Why Chihuahuas Don't Run in the Snow – The Making of*, YOUTUBE (Jan. 19, 2016), https://www.youtube.com/watch?v=nrWO2CHgBCU [https://perma.cc/KQZ2-PVh5-].

36 James Vincent, *Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA About Fake News*, THE VERGE (Apr. 17, 2018, 1:14 PM), https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed [https://perma.cc/89F4-7HGJ].

37 Sarah Mervosh, *Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump*, N.Y. TIMES (May 24, 2019), https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html [https://perma.cc/M2X3-CNJ7].

38 Bijan Stephen, *The White House Used a Doctored Video to Tell a Lie*, THE VERGE (Nov. 8, 2018, 6:49 PM), https://www.theverge.com/2018/11/8/18076532/fake-doctored-video-cnn-cspan-infowars-sarah-huckabee-sanders-jim-acosta [https://perma.cc/U698-9EXY].

39 Drew Harwell, *Faked Pelosi Videos, Slowed to Make Her Appear Drunk, Spread Across Social Media*, THE WASH. POST (May 24, 2019), https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [https://perma.cc/Z3MY-W9AW].

extraction.[40] Condé Nast, the copyright owner of the original video that had been modified to generate the deepfake,[41] indicated that they wished to block it and YouTube removed it from all territories.[42] To date, the deepfake has not been reinstated on YouTube, although it is still available on Instagram[43] and Vimeo.[44] It also appears in an exhibition at the Annka Kultys Gallery in London, titled "Dissimulation," alongside deepfakes of other public figures including Morgan Freeman and Donald Trump.[45]

If Francis and Howe had challenged YouTube's takedown of their Kardashian deepfake, and Condé Nast had sued for copyright infringement, Condé Nast would have needed to show not only that Francis and Howe had copied from its work, but that the copying rose to the level of improper appropriation.[46] This is not a simple case of comprehensive copying: the original audio has been replaced with synthetic audio, and although much of the original video footage has been reproduced, Kim's facial expressions (particularly in the mouth region) would have been altered to match the new audio input. It is more likely a case of fragmented literal similarity,

---

[40] Samantha Cole, *The Kim Kardashian Deepfake Shows Copyright Claims Are Not the Answer*, VICE (June 19, 2019), https://www.vice.com/en/article/j5wngd/kim-kardashian-deepfake-mark-zuckerberg-facebook-youtube [https://perma.cc/L2BV-BEM7].

[41] The original video ("73 Q's with Kim Kardashian West") can be viewed here: *Keeping Up with the Wests: Kim, Kanye (and Their Kids!) Answer 73 Questions*, VOGUE (Apr. 11, 2019), https://www.vogue.com/article/73-questions-with-kim-kardashian-west [https://perma.cc/M9PA-RTAC].

[42] Cole, *supra* note 40. It is still unclear whether Content ID automatically flagged the deepfake, and Condé Nast's default response to Content ID claims is to block them, or whether Condé Nast itself filed a DMCA takedown notice.

[43] *See* Bill Posters (@bill_posters_uk), INSTAGRAM (June 1, 2019), https://www.instagram.com/p/ByKg-uKlP4C/ [https://perma.cc/7FX5-NPVU].

[44] *See* Bill Posters, VIMEO (July 1, 2012), https://vimeo.com/user12695491 [https://perma.cc/Q84H-PXZM].

[45] *See 'Dissimulation,' Solo Show @ Annka Kultys Gallery, Opening Today*, BILL POSTERS (Oct. 7, 2020), http://billposters.ch/dissimulation-solo-show-annka-kultys-gallery-opening-today/ [https://perma.cc/6J5R-B9AX].

[46] Laureyssens v. Idea Grp., Inc., 964 F.2d 131, 140 (2d Cir. 1992) ("A plaintiff must first show that his or her work was actually copied. Copying may be established either by direct evidence of copying or by indirect evidence, including access to the copyrighted work, similarities that are probative of copying between the works, and expert testimony. If actual copying is established, a plaintiff must then show that the copying amounts to an improper appropriation by demonstrating that substantial similarity to protected material exists between the two works.").

where only parts of the original have been reproduced. In this case, Condé Nast would need to show that the part(s) taken include copyrightable expression, and are qualitatively and quantitatively substantial.[47] The original video footage contains expressive elements that may meet the requirements of independent creation and originality[48]—for example, the camera angles, lighting, choreography, costume design, pacing, and the single continuous take.[49] However, the defendants could argue that these compositional

---

[47] Bridgeport Music, Inc. v. Dimension Films, 410 F.3d 792, 797–98 (6th Cir. 2005) (recognizing the "fragmented literal similarity" standard but declining to apply it in cases of digital sampling).

[48] Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 345 (1991) ("Original, as the term is used in copyright, means only that the work was independently created by the author (as opposed to copied from other works), and that it possesses at least some minimal degree of creativity.").

[49] *See, e.g.*, Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53, 58 (1884):

> "A photograph is the mere mechanical reproduction of the physical features or outlines of some object, animate or inanimate, and involves no originality of thought or any novelty in the intellectual operation connected with its visible reproduction in shape of a picture….[T]he process is merely mechanical, with no place for novelty, invention, or originality. It is simply the manual operation, by the use of these instruments and preparations, of transferring to the plate the visible representation of some existing object, the accuracy of this representation being its highest merit…[I]n regard to the photograph in question…'[P]laintiff made the same[] entirely from his own original mental conception, to which he gave visible form by posing the said Oscar Wilde in front of the camera, selecting and arranging the costume, draperies, and other various accessories in said photograph, arranging the subject so as to present graceful outlines, arranging and disposing the light and shade, suggesting and evoking the desired expression, and from such disposition, arrangement, or representation, made entirely by plaintiff, he produced the picture in suit.' These findings, we think, show this photograph to be an original work of art, the product of plaintiff's intellectual invention.";

Gentieu v. Tony Stone Images/Chicago, Inc., 255 F. Supp. 2d 838 (N.D. Ill. 2003) ("[F]or photographs a copyright does not extend to the subject matter of the image itself, but instead protects the expression of the subject as contained in such elements of the author's composition as the selection of lighting, shading, camera angle, background and perspective.").

elements are standard video-interview conventions—*scènes à faire*—and thus not copyrightable.[50]

If Condé Nast can prove that these features of the video are sufficiently original to be copyrightable, the next question is whether the footage taken was qualitatively and quantitatively substantial. At this point, we need to differentiate between the footage that was directly reproduced in the deepfake, and the footage that was used to *train* the algorithm to synthesize Kim's voice and facial movements. With respect to the former, the deepfake appropriates roughly twenty-two seconds of the eleven-minute *Vogue* interview with substituted, synthetic audio and synthetic mouth footage to match the new audio input. Quantitatively, this segment seems insubstantial: it represents roughly three percent of the original video. Qualitatively, this segment is not more important than other parts of the original video: Kim is answering the same kinds of trivial questions that appear in the rest of the video. A court may not find that this rises to the level of improper appropriation.

The second aspect that must be considered here is the footage that was used to *train* the deep learning algorithm. Without access to the corpus of training footage that was used, we can only infer that some, or all, of the *Vogue* interview was used to generate synthetic audio of Kim speaking and synthetic mouth footage to match. Reproduction of this footage within a training dataset is unlikely to be sufficiently transitory to fall outside the scope of copyright law.[51] However, training the algorithm on factual elements of the copyrighted footage—e.g., the physical features of Kim's face—does not

---

[50] *See, e.g.*, Bill Diodata Photography, LLC v. Kate Spade, LLC, 388 F. Supp. 2d 382, 392 (S.D.N.Y. 2005) ("[A]spects of the BDP Photograph that necessarily flow from its idea are not protectible. Under the doctrine of scènes à faire, elements of an image that flow naturally and necessarily from the choice of a given concept cannot be claimed as original."); *Gentieu*, 255 F. Supp. 2d at 848 ("Although in some cases the contrived positioning of a subject has been protected the poses are not copyrightable elements where they follow necessarily from the choice of the subject matter or are otherwise unoriginal.").

[51] Only reproductions of a copyrighted work that are "copies" may constitute infringement, and "copies" are "fixed" in material form such that they are sufficiently permanent or stable to permit them to be perceived, reproduced or otherwise communicated for a period of more than transitory duration. 17 U.S.C. § 101; *see also* Cartoon Network LP, v. CSC Holdings, Inc., 536 F.3d 121, 129 (2d Cir. 2008) (finding that the copyrighted works were not "fixed" in the buffers for a period of more than transitory duration because they resided there for no more than 1.2 seconds before being automatically overwritten).

implicate the video's protectable aspects,[52] and such non-expressive use may fall within the text and data mining exception to copyright infringement.[53]

If Condé Nast can establish improper appropriation, it then falls on Francis and Howe to argue that, nevertheless, they are shielded from copyright infringement liability by the doctrine of fair use. In the United States, fair use is codified as a four-factor analysis: (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole;

---

[52] *See, e.g.*, Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018); Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J. L. & ARTS 45, 51 (2017); Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 U.C. DAVIS L. REV. 893 (2019); Benjamin Sobel, *A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning, in* A.I. & INTELL. PROP. (Reto Hilty, Jyh-An Lee, Kung-Chung Liu, eds., Oxford Univ. Press) (forthcoming 2020).

[53] *See, e.g.*, Kelly v. Arriba Soft Corp., 336 F.3d 811, 819 (9th Cir. 2003) (explaining that the use of copyrighted images for thumbnail images in a visual search engine is transformative because it serves an entirely different function than the owner's original images, it does not supplant the need for the originals, and the use benefits the public by enhancing internet information gathering techniques); Field v. Google, Inc., 412 F. Supp. 2d 1106, 1119 (D. Nev. 2006) (ruling that a search engine allowing users to access copyrighted works through cached links is transformative because it serves a different and socially important purpose than that served by the original works); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1165 (9th Cir. 2007) (explaining that a search engine's display of thumbnail images of copyrighted works was transformative because it provided social benefit as an electronic reference tool); A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 634 (4th Cir. 2009) (ruling that the reproduction of student works by an anti-plagiarism technology system, Turnitin, for the purpose of evaluating originality is transformative fair use); Authors Guild v. Google, Inc., 804 F.3d 202, 207 (2d Cir. 2015) (explaining that the reproduction of copyrighted works for Google Books, including snippet view, was transformative because it enabled users to search for books relevant to their needs and interests, and this use transformatively provided valuable information about the original work, rather than replicating protected expression in a manner that provided a meaningful substitute for the original); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 96 (2d Cir. 2014) (ruling that digitization of copyrighted works to permit full-text searching of works, and to create accessible formats for print-disabled users is transformative fair use).

and (4) the effect of the use upon the potential market for, or value of, the copyrighted work.[54]

In relation to the first factor, there are several sub-factors to be considered. The first is whether Francis and Howe engaged in commercial activities. It is possible that they generated some commercial benefit from the display of the Kardashian deepfake in the Annka Kultys Gallery, and they may have earned advertising revenue from the video if it had remained on YouTube. But Francis has consistently described his goal as artistic, namely, to use deepfakes to "subvert" the cultural authority of celebrities, and to expose the vulnerability of personal data to emerging technologies of power.[55] Francis says his artwork is influenced by Simulationism, Jean Baudrillard's concept of hyperreality, postmodern semiotics, and René Magritte's 1929 Surrealist painting The Treachery of Images.[56] The second sub-factor is an inquiry into the transformativeness of the impugned use.[57] Although the deepfake is unlikely to qualify as a parody of the original work,[58] its appropriation of celebrity likeness to demonstrate our broader societal vulnerability to data exploitation and misinformation achieves a different, socially

---

54 17 U.S.C. § 107.

55 Charlotte Pyatt, *The Art of Interrogation: An Interview with Bill Posters*, JUXTAPOZ (June 19, 2020), https://www.juxtapoz.com/news/street-art/the-art-of-interrogation-an-interview-with-bill-posters/ [https://perma.cc/M8TK-WHSJ].

56 Bill Posters (@bill_posters_uk), INSTAGRAM (June 9, 2020), https://www.instagram.com/p/CBOIEh3lhPr/ [https://perma.cc/56Y5-E2KF].

57 Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) (explaining that the first factor in a fair use enquiry evaluates whether the new work "adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message; it asks, in other words, whether and to what extent the new work is 'transformative.' Although such transformative use is not absolutely necessary for a finding of fair use, the goal of copyright, to promote science and the arts, is generally furthered by the creation of transformative works. Such works thus lie at the heart of the fair use doctrine's guarantee of breathing space within the confines of copyright…the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use.").

58 *Id.* at 579–81 (explaining that parody "has an obvious claim to transformative value" but "like any other use, has to work its way through the relevant factors and be judged case by case in light of the ends of the copyright law."); Rogers v. Koons, 960 F.2d 301, 310 (2d Cir. 1992) ("[T]hough the satire need not be only of the copied work and may, as appellants urge of 'String of Puppies,' also be a parody of modern society, the copied work must be, at least in part, an object of the parody, otherwise there would be no need to conjure up the original work.").

valuable purpose than the original work does.[59] Additional sub-factors, such as customary use and potential bad faith by the defendant, are not applicable in this case and are increasingly ignored in fair use analyses.

The second fair use factor—the nature of the copyrighted work—shifts the focus from the defendant to the plaintiff. A work that is unpublished, or is more creative than factual, is considered more deserving of copyright's protection—courts are less likely to accept a fair use defense in these cases.[60] Here, the original work was a video interview with Kim Kardashian, shot in her home in Hidden Hills, California.[61] It follows the traditional format of *Vogue*'s "73 Questions" interviews, which are designed to provide a raw, unfiltered portrayal of a celebrity at home.[62] The camera follows the celebrity as they walk around their house in a single continuous take, rattling off responses to rapid-fire questions.[63] The importance of the appearance of spontaneity (as opposed to the

---

[59] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 454 (1984) ("[T]o the extent time-shifting expands public access to freely broadcast television programs, it yields societal benefits."); Cariou v. Prince, 714 F.3d 694, 706 (2013) ("If 'the secondary use adds value to the original—if [the original work] is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings—this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society.'") (quoting Castle Rock Ent., Inc. v. Carol Publ'g Grp., Inc., 150 F.3d 132, 142 (2d Cir. 1998)); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 96 (2d Cir. 2014) ("A use is transformative if it does something more than repackage or republish the original copyrighted work….[A] use does not become transformative by making an "invaluable contribution to the progress of science and cultivation of the arts."…Added value or utility is not the test: a transformative work is one that serves a new and different function from the original work and is not a substitute for it.").

[60] Blanch v. Koons, 467 F.3d 244, 256 (2d Cir. 2006) ("Two types of distinctions as to the nature of the copyrighted work have emerged that have figured in the decisions evaluating the second factor: (1) whether the work is expressive or creative, such as a work of fiction, or more factual, with a greater leeway being allowed to a claim of fair use where the work is factual or informational, and (2) whether the work is published or unpublished, with the scope for fair use involving unpublished works being considerably narrower.").

[61] *Kim Kardashian West on Her Growing Family, Law School, and Her Hidden Hills Home*, VOGUE (Apr. 11, 2019), https://www.vogue.com/video/watch/73-questions-with-kim-kardashian-west [https://perma.cc/7TEF-CED9].

[62] Emilia Petrarca, *28 Pressing Questions for the Vogue's '73 Questions' Guy*, THE CUT (Aug. 2, 2019), https://www.thecut.com/2019/08/vogue-73-questions-voice-behind-the-scenes.html [https://perma.cc/TBQ5-7BH3].

[63] *Id.*

interview feeling staged or rehearsed) means that the editing is minimal and no special effects are used. The more creative aspects of the video—the lighting, pacing, choreography, and interview questions—are not the elements that have been reproduced in Francis and Howe's deepfake. Given that the video is more factual than creative, this factor is likely to weigh in favor of a finding of fair use.

The third factor—the amount and substantiality of the portion used—is also likely to weigh in favor of fair use. The Kardashian deepfake appropriates roughly twenty-two seconds of the eleven-minute *Vogue* interview, largely reproducing Kim's visual appearance, but altering her speech and some of her facial movements.[64] This segment represents roughly three percent of the original work.[65] Additionally, the portion used does not go to the heart of the original: Kim merely answers the same kinds of trivial questions that feature in the remainder of the interview.[66] Finally, the amount taken was necessary for the transformative purpose of the use; the interview needed to appear sufficiently professional in order for the deepfake to be convincing.[67]

The fourth and final factor assesses the effect on the potential market for, or value of, the copyrighted work. The Kardashian deepfake, released in June 2019, is unlikely to have diminished the audience for the original *Vogue* video, which has collected over fifty-

---

[64] Bill Posters (@bill_posters_uk), INSTAGRAM, https://www.instagram.com/p/ByKguKlP4C/ [https://perma.cc/DC6P-AGUH].

[65] The original video is eleven minutes and seventeen seconds long, whereas the deepfake is only twenty-two seconds long. *See supra* notes 61, 64.

[66] Baxter v. MCA, Inc., 812 F.2d 421, 425 (9th Cir. 1987) ("Even if a copied portion be relatively small in proportion to the entire work, if qualitatively important, the finder of fact may properly find substantial similarity."); Apple Computer, Inc. v. Microsoft, Corp., 821 F. Supp. 616, 624 (N.D. Cal. 1993) ("[Q]uantitatively insignificant infringement may be substantial if the material is qualitatively important to plaintiff's work.").

[67] *See, e.g.*, Leibovitz v. Paramount Pictures Corp., 137 F.3d 109, 114–16 (2d Cir. 1998) ("'[T]he parody must be able to 'conjure up' *at least* enough of the original to make the object of its critical wit recognizable.'…'"[O]nce enough has been taken to assure identification,' as plainly occurred here, the reasonableness of taking additional aspects of the original depends on the extent to which the 'overriding purpose and character' of the copy 'is to parody the original,' and 'the likelihood that the parody may serve as a market substitute for the original.' That approach leaves the third factor with little, if any, weight against fair use so long as the first and fourth factors favor the parodist. Since those factors favor fair use in this case, the third factor does not help Leibovitz, even though the degree of copying of protectable elements was extensive.").

three million views on YouTube since it was released in April 2019.[68] Fans seeking an intimate portrayal of Kim Kardashian's home and family life would not be discouraged from watching *Vogue*'s interview by Francis and Howe's deepfake. Conversely, given the revenue generated by their celebrity coverage, Condé Nast is unlikely to ever develop a market for unflattering celebrity deepfakes, nor is this a use that they would likely license. Accordingly, the negligible effect of the Kardashian deepfake on the potential market for, or value of, *Vogue*'s interview, would likely weigh in favor of a finding of fair use. Given the four factors analyzed above, Francis and Howe would likely be protected by fair use in any copyright infringement proceeding brought by Condé Nast.

C. Jay-Z/Billy Joel Deepfake

A similar copyright claim was filed against another celebrity deepfake in early 2020. In April, Jay-Z's agency, Roc Nation LLC, filed copyright claims against two YouTube videos containing deepfake audio of Jay-Z reciting a Shakespearean soliloquy ("To Be or Not To Be")[69] and Billy Joel's "We Didn't Start the Fire."[70] The creator of the deepfake audio is an artist known as Vocal Synthesis whose YouTube channel is dedicated to the creation of deepfake audio generated from unlikely audio-textual pairings (for example, George W. Bush performing 50 Cent's "In Da Club").[71] Each deepfake audio is created by feeding a corpus of audio samples and transcripts into Google's open-source neural network system, Tacotron 2.[72] Constructing a training set for a new synthetic voice

---

[68] Vogue, *73 Questions with Kim Kardashian West (ft. Kanye West)*, YOUTUBE (Apr. 11, 2019), https://www.youtube.com/watch?v=QaZ93sibpk0&ab_channel=Vogue [https://perma.cc/8c2Z-9RFG].

[69] Andy Baio, *With Questionable Copyright Claim, Jay-Z Orders Deepfake Audio Parodies off YouTube*, WAXY BLOG (Apr. 28, 2020), https://waxy.org/2020/04/jay-z-orders-deepfake-audio-parodies-off-youtube/ [https://perma.cc/6E4N-HX83].

[70] *Id.*

[71] *Id.*; *see also* Nick Statt, *Jay Z Tries to Use Copyright Strikes to Remove Deepfaked Audio of Himself from YouTube*, THE VERGE (Apr. 28, 2020, 6:38 PM), https://www.theverge.com/2020/4/28/21240488/jay-z-deepfakes-roc-nation-youtube-removed-ai-copyright-impersonation [https://perma.cc/7BGA-5LPE].

[72] Baio, *supra* note 69. For a technical description of Tacotron 2, see JONATHAN SHEN, ET AL., NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS (2018).

and training the model to generate it requires over twelve hours of work, depending on the quality of the audio and the transcript.[73] YouTube initially removed both videos, but later reinstated them after reviewing Roc Nation's DMCA takedown requests and finding them to be incomplete.[74]

The Vocal Synthesis case bears many of the same markings as the Kardashian deepfake, but is complicated by the entirely synthetic nature of the audio output, and the additional copyrightability of the input text. Shakespeare's soliloquy is firmly in the public domain, however Billy Joel's "We Didn't Start the Fire" is still protected by copyright.[75] Accordingly, even if Vocal Synthesis had paid the mechanical licensing fee required to produce a cover of "We Didn't Start the Fire,"[76] he would also require a synchronization license in order to combine his synthetic audio with visual media on YouTube.[77] For now, we will focus on the viability of the copyright claim made by Roc Nation LLC. It is not known precisely which Jay-Z songs were fed to the algorithm, but it seems safe to assume that Roc Nation LLC owns the copyright, for each song, in both the composition and the sound recording.[78] Importantly, none of the songs are directly reproduced; once the algorithm has learned Jay-Z's speech patterns, it generates entirely new, synthetic audio.[79] In other words, the "sound" of Jay-Z's voice is digitally simulated.[80] The reproduction of each song within the algorithm's training

---

73 Baio, *supra* note 69.

74 *See* Statt, *supra* note 71.

75 *Hamlet, Prince of Denmark* was published by William Shakespeare in 1603; under U.S. copyright law, works published before 1925 are generally in the public domain. Billy Joel's "We Didn't Start the Fire" was released in 1989, and the copyright term under U.S. law is the life of the author plus seventy years. 17 U.S.C. § 302.

76 17 U.S.C. § 115.

77 The "cover license" available under 17 U.S.C. § 115(a)(2) allows the licensee to make and distribute a sound recording of the licensed musical composition upon the payment of a small royalty but it does not include the right to synchronize the composition with visual media (as occurs on YouTube), and the new arrangement is not permitted to "change the basic melody or fundamental character of the work."

78 *See, e.g.*, Waite v. UMG Recordings, Inc., 450 F. Supp. 3d 430, 438 n. 50 (S.D.N.Y. 2020) ("[N]oting that since sound recordings earned copyright protection in 1972, 'virtually all contracts' between artists and recording companies include 'work made for hire' provisions." (quoting NIMMER ON COPYRIGHT § 5.03 (2019))).

79 Baio, *supra* note 69.

80 *Id.*

dataset would likely fall within the text and data mining exception to copyright infringement.[81] And the exclusive rights in sound recordings do not extend to the creation of sound recordings which imitate or simulate the original.[82] In order for Roc Nation to establish copyright infringement (in the form of comprehensive nonliteral similarity) it would first need to establish the copyrightability of Jay-Z's speech patterns, which seems unlikely.[83]

Even if Roc Nation could establish copyright infringement, the highly transformative nature of the deepfakes would likely militate against any minimal commercial benefit derived from ad revenue on YouTube.[84] Certainly, the second and third factors would weigh against a finding of fair use; Jay-Z's songs are highly creative and we can assume that the model was trained on a large corpus of full-length songs.[85] However, the fourth factor, like the first, would weigh in favor of a finding of fair use. Neither of the Jay-Z deepfakes would adversely affect the potential market for, or value of, the original Jay-Z songs and they do not represent a market that Roc Nation would likely develop (or license) in the future. Given these

---

81 Posters, *supra* notes 43–44.

82 17 U.S.C. § 114(b).

83 *See, e.g.*, 17 U.S.C. § 102(b) for a summary of the idea-expression dichotomy, which renders facts uncopyrightable; *see also* Sobel, *supra* note 52.

84 Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) ("[T]he more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use."); Dr. Seuss Enter., L.P. v. ComicMix LLC, 256 F. Supp. 3d 1099, 1106, 1109 (S.D. Cal. 2017) ("[I]n the present case there is no question that Defendants created their work for profit. Although this weighs against Defendants in this factor, its weight is slight given both the transformative nature of the work…and the fact that *Boldly* does not supplant the market for *Go!* or the other relevant Dr. Seuss works.") ("This case presents an important question regarding the emerging 'mash-up' culture where artists combine two independent works in a new and unique way…if fair use was not viable in a case such as this, an entire body of highly creative work would be effectively foreclosed.").

85 Authors Guild v. Google, Inc., 804 F.3d 202, 221 (2d Cir. 2015) ("Notwithstanding the reasonable implication of Factor Three that fair use is more likely to be favored by the copying of smaller, rather than larger, portions of the original, courts have rejected any categorical rule that a copying of the entirety cannot be a fair use. Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original…. As with *HathiTrust*, not only is the copying of the totality of the original reasonably appropriate to Google's transformative purpose, it is literally necessary to achieve that purpose.").

considerations, Vocal Synthesis' Jay-Z deepfakes would likely be protected by fair use.

D. Democratizing Creative Production

The evolution of fair use jurisprudence—and its focus on transformativeness—seems consistent with the protection of the artistic deepfakes described above. However, the more difficult question is whether the protection of deepfakes is consistent with the overarching policy goals of the copyright regime. Although the Constitution gave Congress the power to distribute time-limited monopolies to authors and inventors to "promote the progress of science and useful arts,"[86] opinions on the *proper* purpose of copyright still vary widely. There are those who emphasize the importance of rewarding creators with property interests (fairness theory), while others prioritize the protection of the psychic bond between creator and creation (personality theory).[87] Some scholars emphasize the importance of incentivizing the production and distribution of intellectual products as public goods (welfare theory), and still others believe that copyright's purpose is to sustain a just and attractive culture that contributes to human flourishing (cultural theory).[88]

Of these groups, proponents of the cultural theory of copyright law are most likely to advocate for the protection of artistic deepfakes such as those described above. The ability of artists like Francis, Howe, and Vocal Synthesis to subvert and remix popular culture promotes diverse self-expression, and public discourse. Rather than remain passive consumers of cultural works, digital technology allows these individuals to become active co-creators by transforming copyrighted works in unconventional ways.[89] This decentralization

---

[86] U.S. CONST. art. I, § 8, cl. 8.

[87] *See generally* William W. Fisher, *Theories of Intellectual Property*, *in* NEW ESSAYS IN THE LEGAL AND POLITICAL THEORY OF PROPERTY (Stephen Munzer et al. eds., 2001).

[88] *Id*.

[89] *See generally* Niva Elkin-Koren, *Cyberlaw and Social Change: A Democratic Approach to Copyright Law in Cyberspace*, 14 CARDOZO ARTS & ENT. L.J. 215, 236 (1996); Stacey Lantagne, *Mutating Internet Memes and the Amplification of Copyright's Authorship Challenges*, 17 VA. SPORTS & ENT. L.J. 221 (2018); Cathay Smith, *Beware the Slender Man: Intellectual Property and Internet Folklore*, 70 FL. L. REV. 601 (2018); Niva Elkin-Koren, *Copyright in a Digital Ecosystem: A User Rights Approach, in* COPYRIGHT

of the meaning-making process disrupts the commercial paradigm of tightly controlled creativity,[90] and dismantles institutional hierarchies of knowledge production and ownership.[91] It also provokes important social discourse about who we are as a society, and who we want to be. All of these benefits would be chilled by the specter of a copyright infringement suit. And as evidenced by the removal of the Kardashian deepfake from YouTube, many creators will not challenge a wrongful takedown of their content. If the reproduction and manipulation of copyrighted content for the purpose of deepfake creation was not shielded by fair use, copyright infringement liability could present a nontrivial barrier to the creation and dissemination of artistic deepfakes.

II. Individual and Collective Harms

A. Our Vulnerability to Misinformation

As mentioned in the Introduction, it is important to distinguish between the *individual* and *collective* harms generated by the dissemination of deepfakes.[92] An individual whose likeness has been appropriated for deepfake pornography, or other similarly harmful content, may suffer severe emotional distress, psychological harm, and reputational injury. On the other hand, the *collective* harms generated by deepfakes tend to be exacerbations of existing social problems. Online harassment of women, for example, is as old as the

LAW IN AN AGE OF LIMITATIONS AND EXCEPTIONS (Ruth Okediji ed., 2017); LAWRENCE LESSIG, REMIX: MAKING ART AND COMMERCE THRIVE IN THE HYBRID ECONOMY (2008); Teresa Scassa, *Acknowledging Copyright's Illegitimate Offspring: User-Generated Content and Canadian Copyright Law*, *in* THE COPYRIGHT PENTALOGY: HOW THE SUPREME COURT OF CANADA SHOOK THE FOUNDATIONS OF CANADIAN COPYRIGHT LAW (Michael Geist ed., 2013).

90 *See, e.g.*, Debora Halbert, *Mass Culture and the Culture of the Masses: A Manifesto for User-Generated Rights*, 11 VAND. J. ENT. & TECH. L. 921, 924–26 (2009).

91 *See, e.g.*, David J. Gunkel, *What Does it Matter Who Is Speaking? Authorship, Authority, and the Mashup*, POPULAR MUSIC AND SOCIETY, at 71 (Feb. 22, 2012).

92 *See supra* Introduction. Note also the harms associated with the *threat* of such dissemination, for the purposes of blackmail and other kinds of exploitation.

internet itself,[93] but the ease and accessibility of deepfake pornography has increased the severity of this harassment and tied it explicitly to the exploitation and stigmatization of female sexuality. Such harassment ultimately reduces female participation in digital public spaces,[94] and degrades public discourse. Efforts to reduce such harassment require not only the removal of (individual) pornographic deepfakes, but broader measures aimed at the roots of systemic misogyny, and sexism.

Another collective harm associated with the dissemination of deepfakes is a reduction in our capacity to differentiate authentic from inauthentic content. Existing scholarship is rife with inflammatory claims that hyperrealistic deepfakes will undermine public safety, compromise international relations, and jeopardize national security.[95] Deepfakes are certainly capable of all these things, but so are many other kinds of misinformation. And it is *this root cause*—our societal vulnerability to misinformation—that must be addressed. Have we fostered such uncritical reliance on video footage that we would make significant decisions on the basis of uncorroborated evidence? Has societal trust in public institutions fallen so significantly that a single deepfake could cause mass unrest? And if this is the case, how did we get here? Like online harassment, our vulnerability to misinformation has deep roots. Human society has long grappled with inauthenticity in different forms. For as long as

---

93 *See, e.g.*, Ann Bartow, *Internet Defamation as Profit Center: The Monetization of Online Harassment*, 32 HARV. J. L. & GENDER 383 (2009); Alice E. Marwick & Robyn Caplan, *Drinking Male Tears: Language, the Manosphere, and Networked Harassment*, 18 FEMINIST MEDIA STUDIES 543 (2018); Jessica Vitak et al., *Identifying Women's Experiences with and Strategies for Mitigating Negative Effects of Online Harassment*, *in* PROCEEDINGS OF THE 20TH ACM CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK AND SOCIAL COMPUTING 1231 (2017).

94 For a description of incidents concerning Rana Ayyub, an investigative journalist in India, and media critic Anita Sarkeesian, see Drew Harwell, *Fake-Porn Videos are Being Weaponized to Harass and Humiliate Women: 'Everybody is a Potential Target,'* THE WASH. POST (Dec. 30, 2018), https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/ [https://perma.cc/ 3QGB-2LWU]; *see also* Marjan Nadim & Audun Fladmoe, *Silencing Women? Gender and Online Harassment*, SOC. SCI. COMPUT. REV. 245 (2019).

95 *See* Danielle Citron & Robert Chesney, *Deepfakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1784 (2019) ("Foreign policy could be changed in response to convincing deep fakes and forgeries.").

people have valorized certain *things*—original artwork, designer handbags, wild ginseng[96]—inauthentic versions have existed to meet excess demand,[97] and copyright law has been used to remove some of those counterfeits, as discussed in Part I. As it turns out, knowledge of the external world is also a highly-valued commodity, and so we are plied with different versions of it, each competing for ascendance.[98] Ultimately, this competition is about power—about *who* gets to decide what is "real" and what is "fake"—and once we understand this, we can recognize deepfakes as simply the latest weapon in that struggle for epistemological supremacy. To overcome our vulnerability to misinformation, then, we have to combat not only deepfakes, but the modern conditions that have nurtured them, including: shortened attention spans, a 24-hour news cycle, widespread dependence on digital networks (heightened by coronavirus quarantine restrictions), increasing social isolation, low levels of media literacy, social media echo chambers, political polarization, and a loss of trust in both scientific and political institutions.[99] Strengthening our capacity to identify (and counter) misinformation will require engagement on *all* of these fronts, not just the sporadic take-down of individual deepfakes.

---

[96] *See, e.g.*, Q. Lu et al., *Study on Nondestructive Discrimination of Genuine and Counterfeit Wild Ginsengs Using NIRS*, 59 EUR. PHYSICAL J. APPLIED PHYSICS, July 2012, at 1.

[97] *See, e.g.*, David Lowenthal, *Counterfeit Art: Authentic Fakes*, 1 IJCP 79, 79 (1992) ("[C]ounterfeiting is relative; measures of truth vary with time and place. What is fake or forged for some is 'real' or authentic for others…. The roots of originality and counterfeiting, of truth and falsehood, are inextricably intertwined."); John Henry Merryman, *Counterfeit Art*, 1 IJCP 27, 28–29 (1992) (discussing whether a perfect counterfeit, if indistinguishable from the original, is as good as the original, and should be valued as such).

[98] For philosophical meditations on metaphysical uncertainty, see *e.g.*, RENÉ DESCARTES, MEDITATIONS ON FIRST PHILOSOPHY (Hackett Publ'g Co. 3d ed. 1993); Peter Unger, *Ignorance: A Case for Scepticism*, 87 PHIL. REV. 154 (1978); Manley Thompson, *Book Reviews*, 94 ETHICS 143 (1983) (reviewing HILARY PUTNAM, REASON, TRUTH AND HISTORY (1981)).

[99] *See, e.g.*, Stephan Lewandowsky et al., *Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era*, 6 J. APPLIED RSCH. IN MEMORY & COGNITION 353 (2017); Alexei Abrahams & Gabrielle Lim, *Repress/Redress: What the "War on Terror" Can Teach Us About Fighting Misinformation*, THE HARV. KENNEDY SCH. MISINFORMATION REV., July 22, 2020, at 1(3).

B. The History of Ocularcentrism

Another factor which allows deepfakes to exploit our vulnerability to misinformation is the persistence of the normative claim that *seeing is believing*.[100] The epistemological priority of sight is deeply embedded in the Western history of ocularcentrism. The ancient Greeks prioritized sight over other senses due to its simultaneity and perceived objectivity; the ability to avoid direct interaction with the object of your gaze was believed to enhance the neutrality of your perception.[101] The ocularcentrism of early Greek thought was also present in medieval Christian society, which used images to convert new believers and educate the faithful.[102] In a largely illiterate society, biblical events were visually represented in stained glass windows, bas-reliefs, frescoes, altarpieces, and wooden carvings.[103] During the Renaissance, these lessons about the persuasive quality of visual representation were reapplied for secular purposes.[104] Vision "became the dominant sense in the modern world, even as it came to serve new masters."[105]

Renaissance artists used the illusion of perspective to render three-dimensional space on a flat, two-dimensional canvas.[106] The flattened compositions of medieval art were replaced by the illusion of depth,[107] and the many vantage points of medieval scenes were replaced by a single, sovereign eye.[108] This "monocular" or fixed point of beholding the world obscured the bodies of the painter and the viewer, thereby removing their emotional involvement in the depicted scene and giving it the illusion of detached reality.[109] Perspectival art was so widely adopted that its technique of visual

---

100 *See, e.g.*, Christopher J. Buccafusco, *Gaining/Losing Perspective on the Law, or Keeping Visual Evidence in Perspective*, 58 U. MIAMI L. REV. 609, 646 (2004) ("Once the public has become convinced of the transparency of a given medium, it will no longer scrutinize the products of that medium for inconsistencies and biases.").

101 JAY, *supra* note 4, at 23.

102 *Id.* at 30.

103 *Id.* at 30.

104 *Id.* at 32.

105 *Id.* at 32.

106 *Id.* at 35.

107 Lauwrens, *supra* note 4, at 30–31.

108 JAY, *supra* note 4, at 36.

109 *Id.* at 36.

representation became synonymous with vision itself.[110] Cartesian perspectivalism reflected the Western Enlightenment ideals of empirical, scientific observation of the external world, detached from the corrupting influence of the sensual body.[111] Sight was critical for the scientific project of the early naturalists, meticulously observing, measuring, and classifying different specimens so that 'seeing' from a distance become synonymous with scientific knowledge.[112] Probing vision characterized the early scientific revolution, liberating humans from "blind obedience" to the "voices of the past" (the interpreters of religious texts), and allowing them to *observe* the natural world for themselves, especially with the invention of optical instruments.[113]

In the early nineteenth century, the invention of the camera further secured the primacy of vision. The introduction of the daguerreotype in 1839 produced a "cult of images" which flooded mass advertising, in addition to artistic and scientific books.[114] This led to the "democratization of visual experience," or the incorporation of "low" subjects in the canon of what could be visually represented.[115] The daguerreotype was heralded as a direct transcription of reality, "produced by the operation of natural laws and not by the hand of man."[116] The "natural" chemical process of sensitizing a silver-coated copper plate to light through iodine and bromine exposure (and suspending the light-exposed plate over a dish of heated

---

110 *Id.* at 54; *see, e.g.*, Buccafusco, *supra* note 100, at 639, 645 ("The use of linear perspective, as standardized by the Renaissance painters, has disembodied the creative mechanism of image construction and presented images as direct transcriptions of the externally visible world. By erasing the human creator, the process of image creation looks less like a system of communication than a natural process for the gathering of visual data, and the image created is thought of not as a sign, but as a perception….A work in linear perspective is assumed to be a direct and truthful depiction created by an automatic and natural process the success of which need not be questioned."); *see also* Lev Manovich, *The Automation of Sight: From Photography to Computer Vision*, *in* ELEC. CULTURE 230 (Timothy Druckery ed., 1996).

111 JAY, *supra* note 4, at 67; *see also* Lauwrens, *supra* note 4, at 29–30.

112 Lauwrens, *supra* note 4, at 30.

113 JAY, *supra* note 4, at 39-40.

114 *Id.* at 73.

115 *Id.*

116 Jennifer L. Mnookin, *The Image of Truth: Photographic Evidence and the Power of Analogy*, 10 YALE J. L. & HUMAN 1, 16 (1998).

mercury) allowed Nature to simply "reproduce herself."[117] But it did not take long for the manipulability of photography to be realized. At the 1855 Universal Exposition in Paris, the audience was shocked to learn that photographs could be retouched or combined to form a composite image.[118] In 1869, photographer William H. Mumler was charged with fraud for selling "spirit" photographs of deceased individuals that had been produced using double exposures.[119] In 1870, Eugène Appert's fabricated photographs of violent anti-government protestors ("Communards") were used as state propaganda.[120] And in 1899, *Le Siècle* published a frontpage article describing "The Lies of Photography."[121]

Towards the end of the nineteenth century, the hegemony of Cartesian perspectivalism and other "spectatorial" epistemologies began to wane in favor of alternative approaches that exposed the "culturally mediated" nature of sight.[122] Twentieth-century France was characterized by "antiocular discourse" and the "antiretinal" art of Marcel Duchamp.[123] Philosophers began to problematize the unmediated, atemporal, and decorporealized notion of perspective,[124] arguing instead that every viewpoint was value-laden, rather than detached, and projective, rather than merely receptive.[125] The declining ontological primacy of sight was accelerated by the First World War and the Western Front's "bewildering landscape of indistinguishable, shadowy shapes, illuminated by lightning flashes of blinding intensity, and then obscured by phantasmagoric, often gas-induced haze."[126] The unreliability of sight, and the ease of visual disorientation, heightened the importance of non-ocular senses. The Surrealists, who were deeply affected by their wartime experiences,[127] sought to suppress the rational self through "sensual

---

117 *Id.*
118 JAY, *supra* note 4, at 75.
119 Mnookin, *supra* note 116, at 14.
120 JAY, *supra* note 4, at 81.
121 *Id.* at 75.
122 *Id.* at 95.
123 *Id.*
124 *Id.* at 112.
125 *Id.* at 113.
126 *Id.* at 130.
127 *Id.* at 138.

derangement."[128] They challenged the integrity of visual experience through techniques such as collage, frottage, decalcomania, and fumage.[129]

Increasing awareness of the subjectivity of vision was reflected in Jean-Paul Sartre's phenomenology of sight, which displayed a deep distrust of visual illusions and the "treachery" of being defined by the gaze of others.[130] Sartre emphasized the way in which the objectification of the gaze sustained racist and imperialist domination.[131] Similarly, Maurice Merleau-Ponty rejected the epistemology of the "objective spectator" whose vision was "wholly independent of his constitutive powers."[132] Rather, visual experience was subjective and constructed from "orders of signification" in which humans were deeply embedded.[133] Merleau-Ponty's phenomenology of subjectivity insisted that the observer was always *part of* the observed—embodied in the world, rather than disembodied and disinterested.[134] Similarly, Jacques Lacan emphasized the social construction of the visual field, and the sum of discourses that occupied the space between the subject and the world.[135] Meanwhile, Michel Foucault and Guy Debord highlighted the ways in which ocularcentrism promoted institutions of surveillance, spectacle, and social control.[136] In *Discipline and Punish*, Foucault described the ubiquitous, disciplining gaze of the Panopticon, and its ceaseless surveillance.[137] Looking was a form of power, exerted by the subject upon the object, with detached epistemological authority.[138]

---

128 *Id.* at 141.
129 *Id.* at 145.
130 *Id.* at 166.
131 *Id.* at 174.
132 *Id.* at 178.
133 *Id.*
134 Lauwrens, *supra* note 4, at 37.
135 *Id.* at 32.
136 JAY, *supra* note 4, at 229.
137 *Id.* at 240.
138 Lauwrens, *supra* note 4, at 29.

In the late 1960s and early 1970s, antiocular discourse shifted its attention to the "material apparatuses" of photography and film.[139] Roland Barthes understood the allure of photography: "You are the only one who can never see yourself except as an image; you never see your eyes unless they are dulled by the gaze they rest upon the mirror or the lens . . . even and especially for your own body, you are condemned to the repertoire of its images."[140] Barthes railed against its distorting effects: "[O]nce I feel myself observed by the lens, everything changes: I constitute myself in the process of 'posing,' I instantaneously make another body for myself, I transform myself in advance into an image."[141] Barthes emphasized the paradoxical contrast between the photograph's denotative capacity to imitate the world (its "analogical perfection"), and its second-order connotative capacity to signify (its semiotic overlay).[142]

Like photography, film also promoted the "ideology of the visible," or the hegemony of the eye.[143] By the early 1960s, the Barthesian "death of the author"[144] was mirrored in the "death of the *auteur"* within film, and semiological efforts to expose the devices of cinema's "reality effect."[145] Part of film's analogical power stemmed from the movement of images and its simulation of "atemporal instantaneity."[146] The monocular perspective of the camera mimicked the Cartesian perspectivalism of traditional painting, which privileged the "fixed, monologic eye," and disincarnated the painter and viewer in order to give the appearance of detached

---

139 JAY, *supra* note 4, at 260; *see also* Mnookin, *supra* note 116, at 2 (noting that the photograph has "long been perceived to have a special power of persuasion, grounded both in the lifelike quality of its depictions and in its claim to mechanical objectivity"); ROLAND BARTHES, CAMERA LUCIDA (Richard Howard trans., 1st American ed. Hill & Wang 1981); Walter Benjamin, *The Work of Art in the Age of Mechanical Reproduction*, *in* ILLUMINATIONS (Harry Zohn trans., Hannah Arendt ed., 1969); Lorraine Daston & Peter Galison, *The Image of Objectivity*, 40 REPRESENTATIONS, SPECIAL ISSUE: SEEING AUTUMN 81 (1992); SUSAN SONTAG, ON PHOTOGRAPHY (1977).

140 JAY, *supra* note 4, at 265.

141 *Id.* at 267.

142 *Id.* at 262.

143 *Id.* at 273.

144 ROLAND BARTHES, IMAGE, MUSIC, TEXT (Stephen Heath trans., Hill & Wang 1978).

145 JAY, *supra* note 4, at 271.

146 *Id.* at 303.

observation.[147] Rather than conveying multiple, and perhaps different, points of view, the moving camera erased any "dispersed and contradictory subjectivities" to produce a singular, disembodied gaze.[148] The viewer identified with the omniscient camera eye, producing a hyperreal sense of reality in which subject and object collapse into a "state of oneness with the world."[149] The cinematic screen functioned as a mirror, reinforcing the viewer's specular identity and the Lacanian role of mirror reflection in "the visual constitution of the self."[150] Film theorists such as Christian Metz, who worked to expose the ideological underpinnings of the cinematic apparatus, emphasized the object's disavowal of awareness of being viewed as a key feature of film's "reality effect."[151] By the end of the twentieth century, the eye had been deconstructed as an "innocent" medium of knowledge,[152] and philosophers emphasized the "culturally mediated" nature of visual perception.[153]

The history of Western ocularcentrism helps to explain both the persistence of the normative claim that *seeing is believing*, and the

---

147 *Id.* at 275.

148 *Id.* at 275.

149 *Id.* at 277.

150 *Id.* at 205, 278.

151 *Id.* at 272.

152 *Id.* at 348.

153 *Id.* at 95. There is longstanding legal insecurity regarding the admissibility and evocative power of visual evidence as reflected in the iconography of the goddess Justicia—she is blindfolded to prevent the seduction of images and dispassionately deliver impartial verdicts. *See, e.g.*, *The Photograph as a False Witness*, 10 VA. L.J. 644, 645–46 (1886); H. Vogel, *Photography and Truth*, 6 PHILA. PHOTOGRAPHER 262, 262 (1869); H.J. Morton, *The Trials of the Photographer*, 2 PHILA. PHOTOGRAPHER 36, 36 (1865); *Judicial Photography*, 15 PHOTOGRAPHIC J. 107, 107 (1872); Buccafusco, *supra* note 100, at 617; Martin Jay, *Must Justice Be Blind? The Challenges of Images to the Law*, *in* LAW AND THE IMAGE, THE AUTHORITY OF ART AND THE AESTHETICS OF LAW 78 (Costas Douzinas & Lynda Nead eds., 1999); Craig Murphy, *Computer Simulations and Video Re-Enactments: Fact, Fantasy and Admission Standards*, 17 OHIO N.U. L. REV. 145, 163 (1990); John Selbak, Comment, *Digital Litigation: The Prejudicial Effects of Computer Generated Animation in the Courtroom*, 9 HIGH TECH. L.J. 337, 357 (1994); Elan E. Weinreb, Note, *'Counselor, Proceed With Caution': The Use of Integrated Evidence Presentation Systems and Computer-Generated Evidence in the Courtroom*, 23 CARDOZO L. REV. 393, 395–96 (2001).

variety of visual technologies that have exploited this norm.[154] Contextualizing deepfakes within this visual history helps us to conceive them as simply a new iteration of a very old problem. Yet, few techniques of visual manipulation have been met with such outrage and condemnation. So, what makes deepfakes different? The answer may lie in the synchrony of audio and visual elements. A Photoshopped image of an Iranian missile launch, for example, appeals only to a viewer's static sight.[155] But a deepfake of President Obama introducing a deep learning course at MIT,[156] or of Tom Cruise golfing on TikTok,[157] appeals both to our visual and auditory senses in a manner that seems to defy deception. We are *watching* them speak, as we are *hearing* their words. The synchrony of audio and visual elements implicates not just one sense, but two. Psychological research on the emotional responses produced by unimodal (auditory *or* visual) and bimodal (auditory *and* visual) sensory

---

[154] In 1935, Leni Riefenstahl transfigured the "reality" of the Third Reich, while apparently recording it, for her Nazi propaganda film, *Triumph of the Will*. In 2008, the media arm of Iran's Revolutionary Guard, Sepah News, published a photograph of the simultaneous launch of four missiles, which appeared on the cover of the Chicago Tribune, the Financial Times, and the Los Angeles Times. In fact, only three missiles launched that day; the fourth had been digitally added. Earlier this year, U.S. President Donald Trump posted an inauthentic video on Twitter that had been manipulated to falsely represent a CNN broadcast. These examples, while anecdotal, highlight the longstanding vulnerability of visual media to manipulation, even without the use of deepfake technology. *See, e.g.*, Ken Kelman, *Propaganda as Vision: Triumph of the Will*, LOGOS 2.4 (Fall 2003); Hany Farid, *Seeing Is Not Believing*, IEEE SPECTRUM (Aug. 1, 2009, 12:00 AM), https://spectrum.ieee.org/computing/software/seeing-is-not-believing [https://perma.cc/MFK8-ZYUK]; Kate Conger, *Twitter Labels Trump Tweet About 'Racist Baby' as Manipulated Media*, N.Y. TIMES (June 28, 2020), https://www.nytimes.com/2020/06/18/technology/trump-tweet-baby-manipulated.html [https://perma.cc/4E26-F5VA].

[155] *See, e.g.*, MIA FINEMAN, FAKING IT: MANIPULATED PHOTOGRAPHY BEFORE PHOTOSHOP 5 (2012); Mike Nizza & Patrick Whitty, *In Image of Iran's Power, There's Less Than Meets the Eye*, N.Y. TIMES (July 11, 2008), https://www.nytimes.com/2008/07/11/world/middleeast/11missile.html [https://perma.cc/93LD-RTG9].

[156] Alexander Amini, *Barack Obama: Intro to Deep Learning MIT 6.S191*, YOUTUBE (Feb. 10, 2020), https://www.youtube.com/watch?v=l82PxsKHxYc [https://perma.cc/V3AB-3JDW].

[157] Alex Hern, *'I Don't Want to Upset People': Tom Cruise Deepfake Creator Speaks Out*, THE GUARDIAN (Mar. 5, 2021), https://www.theguardian.com/technology/2021/mar/05/how-started-tom-cruise-deepfake-tiktok-videos [https://perma.cc/2RRH-FMEB]; James Vincent, *Tom Cruise Deepfake Creator Says Public Shouldn't Be Worried About 'One-Click Fakes*,*'* THE VERGE (Mar. 5, 2021), https://www.theverge.com/2021/3/5/22314980/tom-cruise-deepfake-tiktok-videos-ai-impersonator-chris-ume-miles-fisher [https://perma.cc/8EAD-PJXT].

experience suggests that the combination of sensory elements affects our response to individual stimuli.[158] Our perception of events or objects in the external world depends upon their stimulation of our senses; for example, we experience a passing car through sight and sound.[159] Deepfakes harness the hyperrealism of *multisensory* experience, and our familiarity with the co-occurrence of speech sounds, and visible changes in an individual's articulatory facial musculature.[160] The audiovisual congruence of deepfakes therefore increases their persuasive effect on the viewer, relative to unimodal techniques of visual media manipulation, such as Photoshop.

As deepfakes test the limits of society's tolerance for visual manipulation, they force us to confront our history of ocularcentrism and to interrogate the utility of the normative claim that *seeing is believing*. Our overreliance on the visual world as our primary source of knowledge and meaning has had far-reaching consequences.[161] For example, the overwhelming credibility of visual evidence has incentivized widespread state surveillance for evidence of wrongdoing,[162] as evidenced by the ubiquity of CCTV cameras.[163] The power to observe—*to make visible*—is the power to control, and modern surveillance technologies extend the dominance of the disembodied gaze.[164] Continuous state surveillance, in turn, erodes public trust in institutions and exacerbates our

---

158 Annabel J. Cohen, *How Music Influences the Interpretation of Film and Video: Approaches from Experimental Psychology*, *in* PERSPECTIVES IN SYSTEMATIC MUSICOLOGY 15–36 (R. A. Kendall & R. W. H. Savage eds., 2005).

159 *Id.* at 16.

160 *Id.*

161 Lauwrens, *supra* note 4, at 26.

162 *See* MICHEL FOUCAULT, DISCIPLINE AND PUNISH: THE BIRTH OF THE PRISON, LONDON 96 (Alan Sheridan trans., Vintage Books 2d ed. 1995) (1977); Apple Igrek, *Review: Gary Shaprio, Archaeologies of Vision: Foucault and Nietszche on Seeing and Saying*, 3 FOUCAULT STUD. 132 (2005) (reviewing GARY SHAPIRO, ARCHAEOLOGIES OF VISION: FOUCAULT AND NIETZSCHE ON SEEING AND SAYING (2005)).

163 *See, e.g.*, Kelly Gates, *The Cultural Labor of Surveillance: Video Forensics, Computational Objectivity, and The Production of Visual Evidence*, SOC. SEMIOTICS, Mar. 12, 2013, at 242 (2013).

164 Christopher Taylor, Visual Surveillance: Contemporary Sociological Issues at 227 (1997) (Ph.D. dissertation, University of Nevada) (on file with University Libraries, University of Nevada Las Vegas); *see also* Donncha Kavanagh, *The Limits of Visualisation: Ocularcentrism and Organization*, *in* THE ROUTLEDGE COMPANION TO VISUAL ORGANIZATION (Emma Bell et al., eds., 2013).

vulnerability to misinformation. Conversely, of course, ocularcentrism is also instrumentalized *against* the state by its citizens to expose wrongdoing by public officials.[165] The nationwide protests against police brutality triggered by the footage of George Floyd's murder reflect this phenomenon.[166] If bystander footage of this kind could be dismissed as a deepfake, criminal acts might go unpunished,[167] and social movements might falter. This alone provides a compelling justification for maintaining the credibility of visual evidence.[168]

And yet—it would be naïve to rest our hopes on visual evidence alone. From Selma's "Bloody Sunday" (1965),[169] to the beating of Rodney King (1991),[170] and the murders of Eric Garner (2015),[171] and George Floyd (2020),[172] the history of visual evidence *demonstrates* that filming police violence does not end police violence. The ubiquity of body cameras—and bystanders with smartphones—

---

165 *See, e.g.*, Sam Gregory, *Cameras Everywhere: Ubiquitous Video Documentation of Human Rights, New Forms of Video Advocacy, and Considerations of Safety, Security, Dignity, and Consent*, 2 J. HUM. RTS. PRAC. 191, 196–98 (2010).

166 *See, e.g.*, Wesley Morris, *The Videos That Rocked America. The Song That Knows Our Rage*, N.Y. TIMES (June 3, 2020), https://www.nytimes.com/2020/06/03/arts/george-floyd-video-racism.html [https://perma.cc/K9CS-YHC9]; Alex Altman, *Why the Killing of George Floyd Sparked an American Uprising*, TIME (June 4, 2020, 6:49 AM), https://time.com/5847967/george-floyd-protests-trump/ [https://perma.cc/V88P-KEUG].

167 *See, e.g.*, Jeannie Suk Gersen, *The Vital Role of Bystanders in Convicting Derek Chauvin,* The New Yorker (April 21, 2021), https://www.newyorker.com/news/our-columnists/the-vital-role-of-bystanders-in-convicting-derek-chauvin [https://perma.cc/8WJ3-6NEK].

168 For a discussion of "the Liar's Dividend," see Chesney & Keats Citron, *supra* note 3, at 888.

169 Aniko Bodroghkozy, *How the Images of John Lewis Being Beaten During 'Bloody Sunday' Went Viral*, THE CONVERSATION (July 23, 2020), https://theconversation.com/how-the-images-of-john-lewis-being-beaten-during-bloody-sunday-went-viral-143080 [https://perma.cc/3KH8-4LQK].

170 Frank Tomasulo, '*I'll See It When I Believe It': Rodney King and the Prison-House of Video*, *in* THE PERSISTENCE OF HISTORY: CINEMA, TELEVISION, AND THE MODERN EVENT 74 (Vivian Carol Sobchack ed., 1996).

171 *See, e.g.*, Katie Benner, *Eric Garner's Death Will Not Lead to Federal Charges for N.Y.P.D. Officer*, N.Y. TIMES (July 16, 2019), https://www.nytimes.com/2019/07/16/nyregion/eric-garner-case-death-daniel-pantaleo.html [https://perma.cc/GE8P-SXYJ].

172 *See, e.g.*, Evan Hill et al., *How George Floyd Was Killed in Police Custody*, N.Y. TIMES (May 31, 2020), https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html [https://perma.cc/UXN6-M4ET].

has not altered police behavior.[173] Nor has careful, frame-by-frame analysis of body camera footage overcome decades of qualified immunity jurisprudence.[174] As Ethan Zuckerman explains, the idea that police violence, like other information problems, could simply be solved by an increase in data flows is a "techno-utopian fantasy."[175] Individuals "armed with images" have *not* been able to effect systemic change.[176] This is not to say that credible video footage bears no utility whatsoever; rather, that focusing disproportionately on acquiring (and disseminating) visual evidence of social problems can distract us from dismantling the (less visible) structures of power that benefit from their persistence.[177] The science of climate change, for example, has been so fiercely repudiated by powerful industries that even the evidence of their own eyes (e.g., the charred Californian coastline) has not convinced climate change deniers otherwise.[178] In other words, information alone, unharnessed to structures of power, cannot effect social change.[179]

If deepfakes force us to interrogate our relationship with visual evidence, and the social utility of ocularcentrism, such conversations are long overdue. Are we still *willing* to accept the price of ocularcentrism, and if we are, what additional safeguards must be

---

173 Ethan Zuckerman, *Why Filming Police Violence Has Done Nothing to Stop It*, MIT TECH. REV. (June 3, 2020), https://www.technologyreview.com/2020/06/03/1002587/sousveillance-george-floyd-police-body-cams/ [https://perma.cc/2S46-2FX4].

174 *See id.*; *see, e.g.*, John P. Gross, *Qualified Immunity and the Use of Force: Making the Reckless into the Reasonable*, 8 ALA. C.R. & C.L. L. REV. 67, 71 (2017); Lindsay de Stefan, *No Man Is Above the Law and No Man Is Below It: How Qualified Immunity Reform Could Create Accountability and Curb Widespread Police Misconduct*, 47 SETON HALL L. REV. 543, 544 (2017); Joanna C. Schwartz, *The Case Against Qualified Immunity*, 93 NOTRE DAME L. REV. 1797, 1817 (2018).

175 Zuckerman, *supra* note 173.

176 *Id.*

177 For a discussion of the role of *invisibility* in human life, see, *e.g.*, Koji Komatsu, *Not Seeing is Believing: The Role of Invisibility in Human Lives*, 51 INTEGRATIVE PSYCH. & BEHAV. SCI. 14 (2017).

178 *See, e.g.*, Peter Baker et al., *As Trump Again Rejects Science, Biden Calls Him a 'Climate Arsonist*,*'* N.Y. TIMES (Sept. 14, 2020), https://www.nytimes.com/2020/09/15/us/elections/biden-calls-trump-a-climate-arsonist-as-the-president-denies-the-science-of-wildfires.html [https://perma.cc/J5KC-ZT5J].

179 Zuckerman, *supra* note 173.

introduced to preserve the credibility of visual evidence?[180] What investments must be made to improve media literacy,[181] and strengthen authentication protocols? How will we maintain the credibility of visual evidence amid the increasing ubiquity of visual manipulation technologies, such as video-editing filters on Instagram and Tik Tok?[182] Unfortunately, as Part III will demonstrate, current legislative initiatives avoid these existential questions, and focus instead on developing Band-Aid solutions to the distribution of specific deepfakes.

III. PROPOSED SOLUTIONS

The public and private sectors have proposed a variety of measures to address the unique challenges posed by deepfakes. The public sector offers legal remedies, while the private sector offers policy and technological solutions. Beginning with the public sector, the first thing to note is that any sweeping legislative prohibition

---

[180] Human society has long introduced safeguards to preserve the credibility of visual evidence, rather than dismiss the utility of visual evidence altogether. *See, e.g.*, Jacqueline Marks Bibicoff, *Seeing Is Believing? The Need for Cautionary Jury Instructions on the Unreliability of Eyewitness Identification Testimony*, 11 SAN FERN. VAL. L. REV. 95, 98 (1983).

[181] *See, e.g.*, Rachel Rodgers et al., *When Seeing Is Not Believing: An Examination of the Mechanisms Accounting for the Protective Effect of Media Literacy on Body Image*, 81 SEX ROLES 87, 87–96 (2019) (discussing how skepticism concerning the extent to which images portray reality can help protect female adolescents from the harmful effects of thin-ideal internalization); Colin C Barton, *Critical Literacy in the Post-Truth Media Landscape*, 17 POL'Y FUTURES IN EDUC. 1024 (2019).

[182] *See, e.g.*, Jiayang Fan, *China's Selfie Obsession*, THE NEW YORKER (Dec. 11, 2017), https://www.newyorker.com/magazine/2017/12/18/chinas-selfie-obsession [https://perma.cc/HY7T-U9PD] ("In the same way that you would point out to your friend if her shirt was misbuttoned, or if her pants were unzipped, you should have the decency to Meitu her face if you are going to share it with your friends…."); Jia Tolentino, *The Age of Instagram Face*, THE NEW YORKER (Dec. 12, 2019), https://www.newyorker.com/culture/decade-in-review/the-age-of-instagram-face [https://perma.cc/Z39K-CXE8]; Orla Pentelow, *FaceTime Timothée Chalamet Thanks to the Instagram Filter You've Been Dreaming Of*, BUSTLE (May 15, 2020), https://www.bustle.com/p/your-facetiming-timothee-chalamet-fantasy-just-came-true-via-this-instagram-filter-22906687 [https://perma.cc/G23F-P6WU]; Christianna Silva, *A Theater Student Gets Supersized Attention After Superhero Video Goes Viral*, NPR (July 5, 2020, 7:53 AM), https://www.npr.org/2020/07/05/887310065/a-theater-student-gets-supersized-attention-after-superhero-video-goes-viral [https://perma.cc/QM4C-WNP2].

on the dissemination of deepfakes would be incompatible with the free speech protections of the First Amendment.[183] Since 1964, the Supreme Court has only strengthened protection for falsehoods,[184] including intentional falsehoods, which cannot be restricted unless they cause serious harm that cannot be avoided through more speech-protective means—e.g., counter speech.[185] And as Cass

---

[183] There are exceptions from First Amendment protection for "obscenity," *See, e.g.*, Roth v. United States, 354 U.S. 476, 483–88 (1957) ("[T]he unconditional phrasing of the First Amendment was not intended to protect every utterance….[I]mplicit in the history of the First Amendment is the rejection of obscenity as utterly without redeeming social importance….We hold that obscenity is not within the area of constitutionally protected speech or press….However, sex and obscenity are not synonymous. Obscene material is material which deals with sex in a manner appealing to prurient interest."); Miller v. California, 413 U.S. 15, 24 (1973) ("The basic guidelines for the trier of fact must be: (a) whether 'the average person, applying contemporary community standards' would find that the work, taken as a whole, appeals to the prurient interest…(b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value."). However, most pornography (and thus deepfake pornography) would be unlikely to meet the (elusive) legal definition of obscenity. *See generally* Douglas Harris*, Deepfakes: False Pornography Is Here and the Law Cannot Protect You*, 17 DUKE L. & TECH. REV. 99–127 (2019); Spivak, *supra* note 28, at 364.

[184] *See, e.g.*, N.Y. Times Co. v. Sullivan, 376 U.S. 254, 271–81 (1964) (noting that the constitutional protection of the First Amendment "does not turn upon 'the truth, popularity, or social utility of the ideas and beliefs which are offered.'…That erroneous statement is inevitable in free debate, and…must be protected if the freedoms of expression are to have the 'breathing space' that they 'need to survive'….The constitutional guarantees require, we think, a federal rule that prohibits a public official from recovering damages for a defamatory falsehood relating to his official conduct unless he proves that the statement was made with 'actual malice'—that is, with knowledge that it was false or with reckless disregard of whether it was false or not."); Gertz v. Robert Welch, Inc., 418 U.S. 323, 341–348 (1974) ("The First Amendment requires that we protect some falsehood in order to protect speech that matters….[T]he States should retain substantial latitude in their efforts to enforce a legal remedy for defamatory falsehood injurious to the reputation of a private individual.…We hold that, so long as they do not impose liability without fault, the States may define for themselves the appropriate standard of liability for a publisher or broadcaster of defamatory falsehood injurious to a private individual.") (noting that private individuals will not have to meet the higher standard of "actual malice" that public officials must meet).

[185] United States v. Alvarez, 567 U.S. 709, 725–29 (2012) ("The First Amendment requires that the Government's chosen restriction on the speech at issue be "actually necessary" to achieve its interest….There must be a direct causal link between the restriction imposed and the injury to be prevented….The Government has not shown, and cannot show, why counterspeech would not suffice to achieve its interest….The remedy

Sunstein explains, there are many reasons to protect falsehoods: state characterizations of "truth" and "untruth" may not always be trustworthy; fear of punishment for falsity will inevitably chill some truthful expression; engagement with false statements can provoke more informed discourse; falsehoods reveal the spectrum of social perspective on an issue; and driving falsehoods underground can inadvertently increase their power.[186] After *Alvarez*, the Supreme Court is likely to permit only carefully tailored restrictions on certain kinds of harmful deepfakes.[187]

The legal remedies available to individual victims of deepfakes fall within civil or criminal liability regimes. With respect to civil liability, assuming the creator of the deepfake is identifiable and located within U.S. jurisdiction, the plaintiff may sue for defamation,[188] publicity in a false light, intentional infliction of emotional distress,[189] wrongful appropriation of another's likeness,[190] or the right of publicity.[191] Alternatively, if the defendant is unidentifiable,

---

for speech that is false is speech that is true….In addition, when the Government seeks to regulate protected speech, the restriction must be the 'least restrictive means among available, effective alternatives.'").

[186] Cass Sunstein, *Falsehoods and the First Amendment*, 33 HARV. J.L. & TECH. 388–426 (2020).

[187] Chesney & Keats Citron, *supra* note 3, at 889.

[188] With respect to defamation, most deepfake videos would likely constitute libel ("printed") rather than slander ("transitory") and would probably rise to the level of "actual malice" required for torts involving public figures. The type of proof necessary will depend on statutory libel requirements, which vary by state. *See* Spivak, *supra* note 28, at 367, 370.

[189] This generally requires proof of "extreme and outrageous conduct." *See, e.g.*, Taliani v. Resurreccion, 115 N.E.3d 1245, 1254 (Ill. App. 3d. 2018) ("To prevail on a claim of intentional infliction of emotional distress, the plaintiff must prove the following three elements: (1) that the defendant's conduct was truly extreme and outrageous, (2) that the defendant either intended that his conduct would cause severe emotional distress or knew that there was a high probability that his conduct would do so, and (3) that the defendant's conduct did in fact cause severe emotional distress."); Cantrell v. Forest City Publ'g Co., 419 U.S. 245, 248 (1974).

[190] For non-celebrities, the tort of wrongful appropriation of the name or likeness of another may be helpful, but the victim would need to demonstrate economic purpose, for example, use of the victim's likeness to endorse or advertise a product. Where no monetary value is derived, it may be difficult for a deepfake victim to satisfy the elements of wrongful appropriation. *See* Spivak, *supra* note 28, at 381.

[191] The right of publicity is generally only useful to celebrities and other public figures who can clearly demonstrate the commercial value in the exploitation of their name and likeness. *See, e.g.*, Hart v. Elec. Arts, Inc., 740 F. Supp. 2d 658, 664, 669 (D.N.J. 2010), and the many statutes that provide an exception for "newsworthy" material.

located outside U.S. jurisdiction, or financially unable to meet a judgment sum, the plaintiff might consider suing the platform on which the deepfake was distributed. Although platforms are largely shielded from liability for user-generated content under the "super immunity"[192] offered by Section 230 of the Communications Decency Act, there are exceptions to this immunity for content that violates federal criminal law, the Electronic Communications Privacy Act, or intellectual property law.[193] Some scholars have also advocated amending Section 230 to extend platform liability to harmful deepfakes, similar to the 2018 amendment for sex trafficking.[194]

Each of these civil remedies has strengths and weaknesses which have been extensively catalogued elsewhere[195] and do not bear repeating here. The only point I wish to make is that any successful use of copyright law to remove revenge pornography does *not* indicate its use in the removal of pornographic deepfakes.[196] The two

---

192 *See* Chesney & Citron, *supra* note 3, at 890.

193 *See* 47 U.S.C. § 230(e). With respect to the intellectual property exception, courts are still divided as to whether the text of the statute ("Nothing in this section shall be construed to limit or expand any law pertaining to intellectual property") refers to state *and* federal law, or just federal law. At multiple points in section 230(e), Congress specified whether it intended a subsection to apply to state or federal law, so some scholars argue that if Congress intended the statute to refer only to federal law, it would have specified as such. On this basis, they argue that the exception covers *all* IP laws, including state laws relating to the right of publicity. Accordingly, a deepfake victim could argue that the immunity provided by section 230 is pierced by a state right of publicity, as an IP right. *See* Almeida v. Amazon.com, Inc., 456 F.3d 1316, 1323 (11th Cir. 2006) ("[T]here appears to be no dispute that the right of publicity is a type of intellectual property right."); *see also* Spivak, *supra* note 28, at 394–95.

194 *See, e.g.*, Citron & Chesney, supra note 95, at 1799 ("Section 230 should be amended to allow a limited degree of platform liability relating to deep fakes.").

195 *See, e.g.*, Harris, *supra* note 183, at 99; Spivak, *supra* note 28, at 364; Elizabeth Caldera, *Reject the Evidence of Your Eyes and Ears: Deepfakes and the Law of Virtual Replicants*, 50 SETON HALL L. REV. 177, 178 (2019); Jessica Ice, *Defamatory Political Deepfakes and the First Amendment*, 70 CASE W. RES. L. REV. 417, 419 (2019); Marc Jonathan Blitz, *Lies, Line Drawing, and Deep Fake News*, 71 OKLA. L. REV. 59, 62 (2018); Rebecca Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, 88 FORDHAM L. REV. 887, 891 (2019).

196 *See, e.g.*, Amanda Levendowski, *Using Copyright to Combat Revenge Porn*, 3 N.Y.U. J. INTELL. PROP. & ENT. L. 422 (2014); Kaitlan M. Folderauer, *Not All Is Fair (Use) in Love and War: Copyright Law and Revenge Porn*, 44 U. BALT. L. REV. 321 (2015); Ann Bartow, *Copyright Law and Pornography*, 91 OR. L. REV. 1 (2012).

situations are very different. Revenge pornography generally involves the nonconsensual release of explicit photographs, previously shared between intimate sexual partners.[197] If the victim took those photographs, they generally own the copyright in them.[198] Pornographic deepfakes, on the other hand, involve substantial alterations to existing footage in order to create a completely new work.[199] This type of transformative use is likely permitted under copyright law, despite the fact that it is obviously egregious in the pornographic context.[200] This is because emotional, psychological, and/or reputational harms are not the intended targets of copyright relief; copyright's function as the engine of free expression is to *promote* the creation and publication of expressive works by protecting the *commercial* interests of the author.[201] Habitual filing of copyright claims to remove pornographic deepfakes would, over time, normalize the misuse of copyright law to remove any unwanted deepfake, and the specter of a copyright claim would chill the creation of beneficial deepfakes. To address the very real and serious harms caused by pornographic deepfakes, Congress should create new statutory remedies specifically designed to help individual victims. At the federal level, a bill has been introduced that would create private rights of action for individual deepfake victims.[202] At the state level, both California and New York have passed legislation

---

197 Levendowski, *supra* note 196.

198 *Id.*

199 Harris, *supra* note 183, at 109.

200 *See, e.g.*, Bleistein v. Donaldson Lithographing Co., 188 U.S. 239, 251–52 (1903) ("It would be a dangerous undertaking for persons trained only to the law to constitute themselves final judges of the worth of pictorial illustrations, outside of the narrowest and most obvious limits. At the one extreme, some works of genius would be sure to miss appreciation. Their very novelty would make them repulsive until the public had learned the new language in which their author spoke.…At the other end, copyright would be denied to pictures which appealed to a public less educated than the judge. Yet if they command the interest of any public, they have a commercial value—it would be bold to say that they have not an aesthetic and educational value- —and the taste of any public is not to be treated with contempt.").

201 *See, e.g.*, Garcia v. Google, Inc., 786 F.3d 733, 744–45 (9th Cir. 2015).

202 Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, H.R. 3230, 116th Cong. § 2(g) (2019).

allowing individual victims of nonconsensual deepfake pornography to sue for damages.[203]

With respect to criminal liability, assuming the availability of both law enforcement resources and prosecutorial will, a deepfake creator could be prosecuted for violating federal cyberstalking laws,[204] certain impersonation crimes,[205] incitement,[206] election-related deception,[207] or the impersonation of public officials, or candidates for office.[208] At the federal level, a bill has been introduced that would criminalize the production and distribution of pornographic deepfakes, as well as deepfakes designed to cause violence or physical harm, incite conflict, interfere in an election, or facilitate criminal conduct.[209] At the state level, Virginia and Texas have passed laws criminalizing nonconsensual deepfake pornography,[210] and deepfakes that interfere with elections,[211] respectively. Massachusetts has proposed the criminalization of the creation or distribution of deepfakes intended for use in otherwise criminal or tortious conduct.[212] Many of these bills require proof of intent, for example, that the defendant *intended* to "humiliate or otherwise harass" the victim of the pornographic deepfake.[213] Such requirements may shield these bills from First Amendment scrutiny.[214]

---

203 *See* A.B. 602, 2019-2020 Reg. Sess. (Cal. 2019); A.B. 5605-C, 2019-2020 Leg., Reg. Sess. (N.Y. 2019).

204 *See* 18 U.S.C. § 2261A and analogous state statutes.

205 *See, e.g.*, N.Y. PENAL LAW § 190.25(4) (2008) ("Impersonates another by communication by internet website or electronic means with intent to obtain a benefit or injure or defraud another, or by such communication pretends to be a public servant in order to induce another to submit to such authority or act in reliance on such pretense.").

206 *See* 18 U.S.C. § 2101(a)(1).

207 *See, e.g.*, COLO. REV. STAT. § 1-13-109(1)(a) ("No person shall knowingly make, publish, broadcast, or circulate or cause to be made, published, broadcasted, or circulated in any letter, circular, advertisement, or poster or in any other communication any false statement designed to affect the vote on any issue submitted to the electors at any election or relating to any candidate for election to public office.") and analogous state statutes.

208 *See, e.g.*, 18 U.S.C. § 912; H.R. 3230.

209 H.R. 3230 § 2(f).

210 VA. CODE ANN. § 18.2-386.2 (West 2019).

211 TEX. ELEC. CODE ANN. § 255.004(d) (West 2019).

212 H. 3366, 191st Gen. Ct. (Ma. 2019).

213 *See, e.g.*, H.R. 3230; *see also* VA. CODE ANN. § 18.2-386.2.

214 *See, e.g.*, Louis Tompros et al., *The Constitutionality of Criminalizing False Speech Made on Social Networking Sites in a Post-Alvarez, Social Media-Obsessed World*, 31 HARV. J.L. & TECH. 65, 88 (2017).

Meanwhile, the private sector response has largely taken the form of policy and technological initiatives. Given the significant portion of social and political discourse that occurs on digital platforms, the policies enacted by these platforms (unconstrained by the First Amendment) are highly consequential. In January 2020, Facebook announced that it would remove any "misleading manipulated media" that had been edited or synthesized "in ways that aren't apparent to an average person and would likely mislead someone" *and* was produced by "artificial intelligence or machine learning."[215] Facebook clarified that this policy would not extend to "parody or satire, or video that has been edited solely to omit or change the order of words."[216] There are three problems with this approach. First, a blanket ban of this kind would chill the creation and distribution of artistic deepfakes of the kind produced by Barnaby Francis, Daniel Howe, and Vocal Synthesis, unless they were classified *by Facebook* as parody or satire, and it's not clear that Facebook should wield such normative authority. Secondly, for those concerned with reducing the spread of misinformation, the limited application of the policy to deepfakes produced by "artificial intelligence or machine learning" seems both arbitrary and ineffective. As we have seen, "cheap fakes" such as the slurred video of Nancy Pelosi also carry significant capacity for harm. Thirdly, reliance on the civic-minded whims of private enterprise to staunch the spread of misinformation has proven to be an unsustainable strategy.[217] Facebook has consistently been unwilling to sacrifice profit in order to play the arbiter of truth.[218] As its user base increasingly trends

---

[215] Monika Bickert, *Enforcing Against Manipulated Media*, FACEBOOK (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/F55T-FSSW].

[216] *Id.*

[217] *See, e.g.*, Sheera Frenkel & Davey Alba, *Trump's Disinfectant Talk Trips Up Sites' Vows Against Misinformation*, N.Y. TIMES (Apr. 30, 2020), https://www.nytimes.com/2020/04/30/technology/trump-coronavirus-social-media.html [https://perma.cc/3C5J-6X3A].

[218] *See, e.g.*, Greg Bensinger, *Does Zuckerberg Understand How the Right to Free Speech Works?*, N.Y. TIMES (July 8, 2020), https://www.nytimes.com/2020/07/08/opinion/facebook-civil-rights-audit.html [https://perma.cc/6Z2P-B26K].

conservative,[219] Facebook has been reluctant to remove any conservative misinformation, even going so far as to label climate change denial "opinion" rather than factually inaccurate.[220]

Twitter's deepfake policy is more nuanced. In February 2020, the platform announced that it may *label* any media that had been "significantly and deceptively altered or fabricated," and may *remove* such media if it had been shared in a deceptive manner, and was likely to impact public safety or cause serious harm.[221] Twitter did not limit the application of its policy to deepfakes that had been created using artificial intelligence or machine learning, thereby facilitating the removal of cheap fakes. It also listed several factors that it would consider in its evaluation of the likelihood that the content would cause harm: threats to physical safety of a person or group; risk of mass violence or widespread civil unrest; and threats to privacy or free expression or participation in civic events.[222] The limited application of the policy to content likely to cause harm protects artistic and other beneficial deepfakes. Like Twitter, YouTube's approach to manipulated media is also conditioned on the risk of harm. Rather than imposing a blanket ban on deepfakes, the platform prohibits content "that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm."[223]

---

[219] *See, e.g.*, Nick Bilton, *How Facebook Became the Social Media Home of the Right*, VANITY FAIR (June 5, 2020), https://www.vanityfair.com/news/2020/06/how-facebook-became-the-social-media-home-of-the-right [https://perma.cc/AW5F-HDBA].

[220] Veronica Penney, *How Facebook Handles Climate Disinformation*, N.Y. TIMES (July 14, 2020), https://www.nytimes.com/2020/07/14/climate/climate-facebook-fact-checking.html [https://perma.cc/L2K3-YWW6]; *see also* Craig Silverman et al., *'I Have Blood on My Hands': A Whistleblower Says Facebook Ignored Global Political Manipulation*, BUZZFEED NEWS (Sept. 14, 2020, 3:36 PM), https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo [https://perma.cc/8E3E-FGFF].

[221] Yoel Roth & Ashita Achuthan, *Building Rules in Public: Our Approach to Synthetic & Manipulated Media*, TWITTER BLOG (Feb. 4, 2020), https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [https://perma.cc/V7A6-44PK].

[222] *Id.*

[223] *Spam, Deceptive Practices & Scams Policies*, YOUTUBE HELP, https://support.google.com/youtube/answer/2801973?hl=en&ref_topic=9282365 [htpps://perma.cc/8TGH-5ADW].

Artistic deepfakes that acknowledge their inauthenticity will not be affected by this policy.

The private sector has also developed a variety of technological means to detect deepfakes, for example, by searching for spatial artifacts (e.g., blending boundary, fore/background contrast, inconsistent head poses), and temporal artifacts (e.g. emotional discrepancies, irregular blinking patterns or pulse signal, viseme/phoneme asynchrony, flickers and jitter).[224] Microsoft's Video Authenticator, launched in September 2020, is designed to help news outlets and political campaigns identify, using a confidence score, the likelihood that media has been artificially manipulated, by detecting the blending boundary of a deepfake and subtle fading or greyscale elements.[225] Artifact-based detection can be evaded, however, by mitigating individual flaws within a deepfake, just as deep learning detection methods can be overcome by adversarial machine learning.[226] For this reason, deepfake detection technologies should be supplemented by anti-tampering protection measures, and content provenance and authenticity frameworks.[227] Blockchain and other distributed ledger technologies may be able to guarantee the provenance, authenticity, and traceability of digital content using smart contracts.[228] Microsoft has developed authentication tools that enable content producers to add digital hashes and certificates to content metadata, which can be read and verified by browser extensions.[229] Congress has also proposed legislation that would increase funding for efforts by the Department of Defense to counter manipulated media content,[230] award prizes for the development of deepfake

---

224 *See* Mirsky & Lee, *supra* note 17, at 26–27.

225 Tom Burt, *New Steps to Combat Disinformation*, MICROSOFT BLOG (Sept. 1, 2020), https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/ [https://perma.cc/6CAZ-4697].

226 *See* Mirsky & Lee, *supra* note 17, at 26, 28.

227 *Id.* at 28.

228 Paula Fraga-Lamas & Tiago M. Fernández-Caramés, *Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality, Computers and Society*, 22 IT PROF'L 53, 55 (2019); Haya. R. Hasan & Khaled Salah, *Combating Deepfake Videos Using Blockchain and Smart Contracts*, 7 IEEE ACCESS 41596, 41597 (2019).

229 Burt, *supra* note 225.

230 National Defense Authorization Act for Fiscal Year 2020, H.R. 2500, 116th Cong. § 256 (2019).

detection technology,[231] and require the National Science Foundation ("NSF") to support research on digital forensic tools designed to detect deepfakes.[232]

CONCLUSION

It has not been my intention to downplay the severity of the harms caused by deepfakes, nor to undermine the legitimacy of public and private sector responses to the unique challenges they pose. Rather, my goal is to temper alarmist claims of "epistemological anarchy"[233] by reminding readers, firstly, that visual evidence has been vulnerable to manipulation for as long as visual technology has existed, and secondly, that our insistence that *seeing should be believing* has harmful as well as beneficial consequences. The advent of deepfakes has exposed the vulnerabilities of an ocularcentric society and reinforced the importance of building trust in public institutions, authenticating and corroborating sources, and investing in media literacy and education. The social issues that are exacerbated by deepfakes—including misogyny and misinformation—have deep roots, and short-term efforts to stifle the distribution of deepfakes should not distract us from the larger project of dismantling the (invisible) social structures that support them. Nor should widespread condemnation of pornographic and other harmful deepfakes prompt the misuse of copyright law to remove deepfakes from digital platforms. Deepfake technology bears significant capacity for social good, and this capacity should not be stifled by unfounded fearmongering.

---

[231] Damon Paul Nelson and Matthew Young Pollard Intelligence Authorization Act for Fiscal Years 2018, 2019, and 2020, H.R. 3494, 116th Cong. § 707 (2019).

[232] Identifying Outputs of Generative Adversarial Networks Act, H.R. 4355, 116th Cong. § 3 (2019); *see also* H.R. 3230, 116th Cong. § 7 (2019).

[233] William Galston, *Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics*, BROOKINGS (Jan. 8, 2020), https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/ [https://perma.cc/M8MR-QCK5].