**“Digital Humanities in the Deepfake Era”**

forthcoming chapter in *Debates in the Digital Humanities 2022*, edited by Gold and Klein (University of Minnesota Press, 2022)

Abraham Gibson

The video is unmistakable and unforgettable. The grainy footage shows President Nixon taking a seat and steeling his nerves. It was to be the most important speech of his career, one that he had hoped he would never deliver. “My fellow Americans,” he begins in a somber tone, “fate has ordained that the men who went to the moon to explore in peace will stay on the moon to rest in peace.” The implications were clear enough. Something had gone terribly wrong with Apollo 11, the first NASA mission to attempt a lunar landing, and astronauts Neil Armstrong and Buzz Aldrin were dead. “For every human being who looks up at the moon in the nights to come will know that there is some corner of another world that is forever mankind.” With these words, Nixon sought to console a grieving nation during one of the most tragic moments in its history.

Except the tragedy never happened. To be sure, the text of the speech was real. It had been written by Nixon’s speechwriter, William Safire, in 1969, in case something went wrong. Thankfully, the mission succeeded, and the speech wasn’t needed. It went unread until 2019, when researchers at the MIT Center for Advanced Virtuality resurrected the script (and the President’s visage) for a special screening at the 2019 International Documentary Film Festival. The short film, “In Event of Moon Disaster,” was produced on a computer using AI-powered video dialogue replacement. The filmmakers sought to emphasize both the power and the danger of synthetic audiovisual media, otherwise known as “deepfakes” (Panetta and Burgund 2019).

In general usage, the word “deepfake” refers to any video that has been altered using machine-learning algorithms to produce hyper-realistic videos that show actual people saying and doing things they never said or did. Some videos seamlessly transplant one person’s face

onto another person's body. Others create an entire person from scratch. In both cases, the algorithmic alterations are difficult, if not impossible, to detect with the naked eye. There are now several mobile apps that help users with no programming skills produce synthetic videos with face-swapping, artificial lip-synching, and more, but most researchers do not consider these videos deepfakes since they do not use machine learning (Paris and Donovan 2019). In late 2019, cybersecurity officials identified more than 14,000 deepfakes on the internet, a 100% increase in just seven months (Ajder, *et al.* 2019). The number has no doubt risen even higher since then.

The first deepfake to gain widespread attention was released in 2018. The video appeared to show Barack Obama saying, "Donald Trump is a total and complete dipshit." In fact, the video was an effective PSA from actor and director Jordan Peele about the dangers of deepfakes. Acting as an invisible marionette, Peele used the Obama avatar to caution viewers that deepfakes could lead to "some kind of fucked-up dystopia" (Mack 2018). In the years since, several other deepfakes have also gone viral. Perhaps you've seen the one that shows Bill Hader morphing into Arnold Schwarzenegger, or the one that places Will Smith's face on Cardi B.'s body. It isn't just videos, by the way. Last year, a British energy company was duped out of nearly a quarter-million dollars when a scammer used AI to imitate the CEO's voice and request a wire transfer over the phone (Stupp 2019). Some have expressed fear that deepfakes herald nothing less than the "information apocalypse" (Silverman 2020), the very "collapse of reality" (Foer 2018).

Efforts to counter the deepfake menace have so far fallen into one of two categories. First, many in the tech sector promote technological solutions, especially automated detection. There are practical problems with this approach, however. After all, the automated detectors train the automated generators to fix their mistakes, which means that subsequent deepfakes may evade detection altogether. Other potential technological solutions include provenance stamps,

geolocation tags, blockchain verification, and mandatory registration of content creators (Knight 2018, Paris and Donovan 2019, Chesney and Citron 2019). Second, there are legal options. For example, China recently banned deepfakes outright. Germany has also passed stiff laws penalizing media companies that fail to remove racist or threatening content from their platform in an expeditious manner, and some believe that a similar tactic might work with deepfakes. At the present, however, most places have no laws to handle deepfakes, which means that legal challenges to deepfakes will be handled in a costly and inefficient case-by-case basis (Citron and Chesney 2019, Woollacott 2019).

Technological solutions and legal challenges are limited because they are post-hoc fixes to larger, more complex sociotechnical systems. By comparison, the digital humanities are uniquely interested in both the social and the technical. Practitioners are trained to describe complicated systems, including their impact on society and culture, and can thus help support structural change. Ten years ago, the digital humanities might not have been ready to meet the deepfake challenge. The field was still young, still coalescing, still largely preoccupied with what qualifies as the digital humanities. The field entered a second phase around five years ago, as scholars sought to establish connections between the digital humanities and other fields of inquiry across campuses and throughout the academy. We have now entered a third phase, and not a moment too soon. This latest iteration insists that the digital humanities must embrace social activism, that we must move beyond the campus, and that we must engage the public at all costs. Rather than framing the digital humanities around advances in digital scholarship, Safiya Umoja Noble writes, perhaps we should interrogate how digital humanities too often obscure important features of the social, political, and economic landscape (Noble 2019). Matthew K. Gold and Lauren Klein are more explicit still: “Our work within the digital humanities is enabled

by larger social, political, and technological systems. In the present moment, we need work that exposes the impact of our embeddedness in those larger systems and that brings our technical expertise to bear on the societal problems that those systems sustain" (Gold and Klein 2019).

First and foremost, scholars in the digital humanities should broadcast our core belief that technologies are best understood as sociotechnical systems. This broader view acknowledges that technology cannot be divorced from its social context. It encourages the citizenry to expect and, if necessary, demand answers about the provenance of any given technology. Rather than trying to identify and stamp out every deepfake on the web, digital humanists can help identify the deeper structural issues that allow tens of thousands of deepfakes to proliferate in the first place. We can advocate for reform, rejecting false creeds like "techno-solutionism" that promote technology as humanity's savior (Morozov 2013). As Catherine D'Ignazio and Lauren Klein explain, "we must look to understand and design systems that address *oppression* at the structural level" (D'Ignazio and Klein 2020).

We can also help demystify deepfakes by showing that previous technologies engendered similar concerns (Pyne 2019). Take the printing press, for example. Several writers have shown that the rise of pamphlet culture in seventeenth-century England facilitated widespread disinformation during the first English Civil War (White 2019, Peacey 2013). The invention of photography was no less revolutionary, and it too facilitated disinformation. Famous examples from the American Civil War include the transposition of Lincoln's head on to Calhoun's body, the insertion of Grant into a battle scene near City Point, and Matthew Brady's decision to re-arrange dead bodies of soldiers to set up visual tableaux (Trachtenberg 1989). A century later, Stalin famously ordered the mass alteration of photographs in Soviet Russia, an effort to purge his enemies from the history books (King 1997). Meanwhile, the ability to deceive using

audiovisual (AV) manipulation was, until recently, the stuff of Hollywood. *Jurassic Park* and *Forrest Gump* were among the first, using AV manipulation to resurrect dinosaurs and Presidents, respectively. The visage of the late Peter Cushing was recently used in *Rogue One*, and there are plans to cast James Dean in a drama about the Vietnam War (Ritman 2019).

These famous examples notwithstanding, the overwhelming majority of extant deepfakes are non-consensual pornography. In fact, the word "deepfake" was coined in November 2017, when an anonymous Reddit user named "deepfakes" shared a software toolkit that would allow anyone to make synthetic videos replacing one person's face with another. To demonstrate, the user posted a manipulated video that appeared to show actress Gal Gadot in a pornographic film (Rothman 2018). Since then, deepfake porn has exploded in popularity. In fact, one recent census found that approximately 96% of deepfakes on the internet are pornographic, and that approximately 99% of these fake videos target women (Ajder, *et al*. 2019).

None of this will come as a surprise to scholars in the humanities. As is well known, the porn industry has often served as a catalyst for technological change. It has also been one of the most reliable early adopters for any new medium. Porn played a crucial role in development and/or expansion of Polaroid cameras, handheld camcorders, VCRs, cable TV, and the internet (Barss 2010, Coopersmith 1998). Deepfakes might seem like just another connection between porn and tech, but they are much different. When it comes to deepfake pornography, the "participants" don't even know it is happening. Celebrities are not the only target, either. Deepfakes can now be weaponized against anyone, including ex-lovers, professional rivals, and even random strangers, none of whom consented to "star" in a pornographic video.

Countless studies in the digital humanities and beyond have shown that technology has a disproportionately negative impact on women and other minoritized groups. Scholars have

demonstrated time and time again that digital technologies are more likely to reify existing power structures than undermine them, that digital infrastructures invariably have an outsized impact on underrepresented groups, and that the digital humanities can help us elucidate structural inequalities in the non-digital world (Benjamin 2019, Noble 2018, Broussard 2018, Eubanks 2018, Losh and Wernimont 2018, Hicks 2017, O'Neill 2016, Gallon 2016). As digital humanists, we should advocate for equity, dignity, and structural change.

Deepfakes will also create challenges that are unique to specific domains within DH. Consider their impact on digital history, for example. Some outside the field have suggested that deepfakes will be good for historians. Legal experts Danielle Citron and Bobby Chesney both herald the potential pedagogical value that deepfakes offer historians, writing that deepfakes make it possible to "manufacture videos of historical figures speaking directly to students" (Citron and Chesney 2019). Sure enough, programmers are already using OpenAI's Generative Pre-trained Transformer-3 (GPT-3) to simulate correspondence with historical personalities from any era.[1] Once again, the practical applications are not limited to famous people. For example, Microsoft recently filed a patent that would allow the company to digitally "resurrect" the recently departed in the form of chatbots and eventually deepfake avatars (Smith 2021).

Even so, the preponderance of evidence suggests there are legitimate reasons for concern. First, malicious actors *already* mischaracterize the historical record in service of their agenda. Studies have shown that white supremacists routinely co-opt and misrepresent research on genetic genealogy, classical antiquity, and Viking genomics to suit their racist worldview (Panofsky and Donovan 2019, Nelson 2016, Strand and Källén 2021). One can only imagine how neo-Confederates and neo-Nazis might like to revise the historical record with help from

[1] https://aiwriter.app/

deepfakes. Second, deepfakes will generate different reactions depending on the context in which they're viewed. Consider the Nixon deepfake. While it was first displayed within the context of a museum exhibit, the film is now available on YouTube. Given that millions of Americans already doubt the moon landings ever took place, it is easy to imagine someone thinking that the deepfake is real, especially when it is divorced from its original context. Third, many of the strategies to combat deepfakes, from litigation to "life-logging" (Citron and Chesney 2019, Eggers 2013), presume that the target is still alive, but who speaks for the dead? Peter Cushing never consented to star in *Rogue One*., and President Nixon never eulogized Apollo 11. Did each surrender rights to his visage when he died? Does everyone? These questions are ripe for humanistic analysis of agency, power, authorship, and intellectual property, among other topics.

Historians who study pre-twentieth-century history, when moving-image technology was largely non-existent, might think that they will be spared the worst effects of deepfakes, but generative adversarial networks, or GANs, leaves *all* media vulnerable to digital manipulation.[2] This includes everything from photographs to cave art. In fact, the first piece of AI-generated art (a blurry portrait of a non-existent person, created by GANs that were trained on more than 15,000 portraits) recently sold at Christies for more than $400,000 (Cohn 2018). GANs can also read and manipulate digitized texts, which are the foundation of modern historical research. Combining machine learning with image processing techniques, researchers have built OCR engines that can read typeface and, increasingly, handwritten text (Mermon *et al* 2020). In fact, AI allows machines to not only read handwriting, but also produce handwriting in any language. How will historians discern genuine primary sources from fake ones? Ironically, in an

[2] GANs employ two sets of algorithms: one, the generator, creates content that is modeled on existing source data, while the other, the discriminator, works to constantly identify imperfections in the false images. Working in tandem, the algorithms train against one another to produce an ever more realistic image.

increasingly digitized world, we will be called upon to vouchsafe the physical archives and safeguard authenticity like never before.

This is more than an academic debate. To quote George Orwell's prescient aphorism, "Who controls the past controls the future, who controls the present controls the past." Authoritarian regimes will not hesitate to suppress and misrepresent the historical record to retain power. Glenn D. Tiffert explains how officials in communist China recently asked Cambridge University Press and Springer Nature to hide more than a thousand potentially offensive articles on Chinese history from users in China. Keen to remain in business with the most populous nation on Earth, the publishers quietly acquiesced. For subscribers in China, the items simply disappeared from search results, as if they never existed. Such large-scale censorship would have been impossible a generation ago. Now, it can be done with a few clicks (Tiffert 2019).

Tiffert drew his evidence from communist China, but the past few years have shown that democracies are also vulnerable to authoritarianism. In the United States, Donald Trump actively promoted widespread disinformation throughout his presidency. He routinely called fake news real and real news fake. He often retweeted "cheapfakes," relatively crude AV manipulations produced with mobile apps rather than artificial intelligence, and, given that his only criterion for truth was whether or not the information flattered him, he would have undoubtedly promoted any deepfake that cast him in a positive light. Trump was eventually kicked off Twitter after he incited a deadly riot at the U.S. Capitol, but democracies remain vulnerable to disinformation in general and deepfakes in particular. Malicious actors are known to use fake profiles on social media using AI-generated faces (Gleicher 2019), and research has shown that micro-targeted deepfakes can have real and demonstrable effect on political attitudes (Dobber, *et al*. 2020).

To properly confront the deepfake challenge, DH scholars may find it necessary to utilize both computational methods and traditional methods of scholarly research. There is a rapidly growing body of scholarship on the computational humanities in general (Johnson, *et al.* 2021, Mullen 2018-2020, Afanador-Llach, *et al*. 2017, Graham, *et al.* 2016, Arnold and Tilton 2015), and the "visual turn" in DH in particular (Tilton 2019). Recent advances in machine learning and computer vision have transformed how researchers engage with visual media. Algorithms for machine vision can detect, identify, and qualify features within an image, and then build predictive models that it can apply to future datasets. A growing number of scholars utilize both computational and humanistic methods in their analyses of visual media, thereby signaling potential ways of engaging with deepfakes from a DH perspective (Arnold and Tilton 2019, Wevers and Smits 2020, Lee 2020, Mittell 2019). Meanwhile, just as DH scholars must be willing to utilize computational tools, so too must we insist that fields like computer science integrate humanistic perspectives. Doing so would bring critical perspectives to bear on technological advances. We need, in other words, more multidisciplinary research in the technology sector. DH intervention doesn't necessarily mean that every DHer should start researching deepfakes. Instead, we should advocate for embedding humanities perspectives in the tech world to anticipate problematic aspects of tech research.

In closing, DH scholars should bear three lessons in mind as we brace for potentially ubiquitous deepfakes. First, we should promote **digital literacy** at every turn. We don't necessarily need to learn code, but we *do* need to understand how algorithms influence every part of our daily lives if we are going to successfully navigate the cacophony of disinformation (McPherson 2012, Schmidt 2016). Second, DH scholars must also embrace **advocacy**. Deepfakes threaten everything from interpersonal relationships to international relations, but

they are especially injurious toward women and other minoritized groups. Rather than resigning ourselves to nihilism, we can advise our neighbors to seek answers when confronted with a provocative video on social media. What is the source of this video? Who might have created it and why? Finally, we should recommit to **values** that have long defined the digital humanities. DH scholars have suggested several core principles to help unify the field, including openness, collaboration, collegiality, connectedness, diversity, experimentation, hope, compassion, and empathy (Nowviskie 2015, Spiro 2012, Noble 2019, Klein and Gold 2019). Keeping the faith will not be easy. Deepfakes breed distrust and fascism weaponizes credulity (Tiffert 2019), but that only underscores the importance of the DH perspective. If we're earnest about our values, then the digital humanities can help the larger public stay resilient, rebellious, and real.

**BIBLIOGRAPHY**

Afanador-Llach, Maria José, *et al*. (eds.). *The Programming Historian*, 2nd edition (2017) <programminghistorian.org>.

Ajder, Henry, *et al*. "The State of Deepfakes: Landscape, Threats and Impact," *Deeptrace Labs* (September 2019).

Ajder, Henry. "Deepfake Threat Intelligence: a statistics snapshot from June 2020," Sensity (March 7, 2020).

Arnold, Taylor and Lauren Tilton. "Distant Viewing: Analyzing Large Digital Corpora," *Digital Scholarship in the Humanities* (2019): <doi.org/10.1093/digitalsh/fqz013>.

Arnold, Taylor and Lauren Tilton, *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text* (Springer, 2015).

Barss, Patchen. *The Erotic Engine: How Pornography Has Powered Mass Communication from Gutenberg to Google* (Ancho Canada, 2011).

Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code* (John Wiley and Sons, 2019).

Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, 2018)

Chesney, Robert and Danielle K. Citron. “Deepfakes and the New Disinformation War,” *Foreign Affairs* (January/February 2019).

Citron, Danielle K. and Robert Chesney. “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” *California Law Review* (2019).

Cohn, Gabe. “AI art at Christie’s sells for $432,500,” *New York Times* (October 25, 2018).

Coopersmith, Jonathan. “Pornography, Technology, and Progress,” *Icon* 4 (1998): 94-125.

Coppin, McKay. “The Billion-Dollar Disinformation Campaign to Reelect the President,” *The Atlantic* (March 2020)

Ctrl Shift Face, “Bill Hader impersonates Arnold Schwarzenegger [Deepfake],” YouTube, May 10, 2019: <youtube.com/watch?v=bPhUhypV27w>.

D’Ignazio, Catherine and Lauren F. Klein, *Data Feminism* (MIT Press, 2020).

Dobber, Tom, *et al*. “Do (microtargeted) deepfakes have real effects on political attitudes?” *International Journal of Press/Politics* 25 (2020): 1-23

Eggers, Dave. *The Circle* (Knopf, 2013).

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin’s Publishing, 2018)

Foer, Franklin. “The Era of Fake Video Begins,” *The Atlantic* (May 2018).

Gallon, Kim. “Making a Case for the Black Digital Humanities,” in *Debates in the Digital Humanities 2016,* edited by Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2016)

Gleicher, Nathaniel. “Removing Coordinated Inauthentic Behavior from Georgia, Vietnam and the US,” *Facebook Newsroom* (December 20, 2019).

Gold, Matthew K. and Lauren F. Klein, “A DH That Matters,” in *Debates in the Digital Humanities 2019,* edited by Matthew K. Gold and Lauren F. Klein (Minneapolis: University of Minnesota Press, 2019)

Shawn Graham, *et al. Exploring Big Historical Data: The Historian’s Macroscope* (London: Imperial College Press, 2016) <themacroscope.org/2.0>.

Hicks, Marie. *Programmed Inequality: How Britain Discarded Women and Lost Its Edge in Computing* (MIT Press, 2017)

Johnson, Jessica Marie, *et al.*, (eds.). *Computational Humanities* (Minneapolis: University of Minnesota Press, 2021).

King, David. *The Commissar Vanishes: The Falsification of Photographs and Art in Stalin's Russia* (Henry Holt, 1997).

Knight, Will. "The Defense Department has produced the first tools for catching deepfakes," *MIT Technology Review* (August 7, 2018)

Losh, Elizabeth and Jacqueline Wernimont. *Bodies of Information: Intersectional Feminism and the Digital Humanities* (University of Minnesota Press, 2019)

Mack, David. "This PSA about Fake News from Barack Obama is Not What It Appears," *Buzzfeed* (April 17, 2018).

McPherson, Tara. "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation," in *Debates in the Digital Humanities 2012,* edited by Matt Gold (University of Minnesota Press), 139-160.

Mermon, J., et al. "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *IEEE Access* 8 (2020): 142642-142668.

Mittell, Jason. "Videographic Criticism as a Digital Humanities Methods," in *Debates in the Digital Humanities 2019,* edited by Matthew Gold and Lauren Klein (University of Minnesota Press, 2019).

Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism* (Public Affairs, 2013).

Mullen, Lincoln A. *Computational Historical Thinking: With Applications in R* (2018-2020) <dh-r.lincolnmullen.com>.

Nelson, Alondra. *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome* (Beacon Press, 2016).

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018)

Noble, Safiya Umoja. "Toward a Critical Black Digital Humanities," in *Debates in the Digital Humanities 2019,* edited by Matthew Gold and Lauren Klein (University of Minnesota Press, 2019).

Nowviskie, Bethany. "Digital Humanities in the Anthropocene," *Digital Scholarship in the Humanities* 30 (December 2015): i4-i15.

O'Neill, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Penguin Books, 2016).

Panetta, Francesca and Halsey Burgund. "In Event of Moon Emergency (short film)," MIT Center for Advanced Virtuality, 2019.

Panofsky, Aaron and Joan Donovan. "Genetic ancestry testing among white nationalists: from identity repair to citizen science," *Social Studies of Science* 49 (2019): 653-681.

Paris, Britt and Joan Donovan. "Deepfakes and Cheap Fakes," *Data & Society* (September 18, 2019).

Peacey, Jason. *Print and Public Politics in the English Revolution* (Cambridge University Press, 2013).

Pyne, Lydia. *Genuine Fakes: How Phony Things Teach Us About Real Stuff* (Bloomsbury 2019).

Ritman, Alex. "James Dean Reborn in CGI for Vietnam War Action-Drama," *Hollywood Reporter* (November 6, 2019).

Rothman, Joshua. "In the Age of A.I., is Seeing Still Believing?" *New Yorker* (November 12, 2018).

Schmidt, Benjamin. "Do Digital Humanists Need to Understand Algorithms?" in *Debates in the Digital Humanities 2016,* edited by Matt Gold and Lauren Klein (University of Minnesota Press, 2016).

Silverman, Craig. "The information apocalypse is already here, and reality is losing," *Buzzfeed* (May 22, 2020).

Smith, Adam. "Microsoft patent shows plans to revive dead loved ones as chatbots," *The Independent* (January 20, 2021).

Spiro, Lisa. "'This Is Why We Fight': Defining the Values of the Digital Humanities," in *Debates in the Digital Humanities 2012,* edited by Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2012).

Strand, Daniel and Anna Källén, "I am a Viking! DNA, popular culture and the construction of geneticized identity," *New Genetics and Society* (2021).

Stupp, Catherine. "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case," *Wall Street Journal* (August 30, 2019).

Tiffert, Glenn D. "Peering down the Memory Hole: Censorship, Digitization, and the Fragility of Our Knowledge Base," *American Historical Review* 124 (April 2019): 550-568.

Tilton, Lauren. “The Visual Turn in DH,” Keynote at the Digital Humanities and the Visual World Symposium (October 12, 2019).

Trachtenberg, Alan. *Reading American Photographs: Images as History* (Hill and Wang 1989).

Wevers, Melvin and Thomas Smits, “The visual digital turn: Using neural networks to study historical images,” *Digital Scholarship in the Humanities* 35 (2020): 194-207.

White, William. “Parliament, Print, and the Politics of Disinformation, 1642-1643,” *Historical Research* 92 (November 2019): 720-736.

Woollacott, Emma. “China Bans Deepfakes in New Content Crackdown,” *Forbes* (November 30, 2019).