Artificial Intelligence as a Disinformation Engine

Revision #76

Dylan W. Wheeler

University of New Hampshire

Philosophy Thesis – In Review

May 13, 2020

ACKNOWLEDGMENTS

This research would not have been possible without the mentorship and direction given by Nicholas J. Smith, Ph.D. His undergraduate classes in topics of science, technology, and society originally inspired me to consider these pressing problems. Appreciation is due to Brandon Bryant for his expertise in cryptography and decentralized currencies. I am grateful for the wisdom of Tom Haines and Jaed Coffin, whom each helped me crystalize the broad scope of disinformation and its implications. My heartfelt thanks to my parents for their unwavering support that empowered me to enjoy so many opportunities.

Table of Contents

Introduction

New forms of artificial intelligence (A.I.) can autonomously replace people in existing images or videos with the likeness of others. The resulting media are called deepfakes. These replacements could take any form—anything from basic face-swapping (think: Snapchat augmented reality filters) to full-body mapping, where a person's body can be mapped onto videos of professional dancers to make it appear that they, too, are highly coordinated. People of the Internet are now using various kinds of A.I. that are excellent at face-swapping to deepfake actor Nicholas Cage into various movies that did not cast him. These algorithms are allowing anyone in the world to easily make it look like a target is doing or saying anything.

In the wrong hands, these innocent algorithms enable the mass manufacture of disinformation (or "fake news"), false information spread deliberately to deceive harmfully. If we are not careful, we could enter a post-truth future, where the truth about issues becomes overshadowed by emotionally charged headlines and discourse, making it even more challenging to identify truthful and genuine media. With little ways to determine which images and videos have been synthesized, the most sensationalist disinformation will go viral and attract the most attention, usually at the expense of a population or demographic. In this paper, I argue that AI-powered disinformation tools are becoming increasingly ubiquitous. I argue that the best way to address disinformation is by facilitating widespread digital literacy while building new technologies that provide increased context for news articles and social media posts. My central recommendations will be advocating for digital literacy education, investments in technologies to provide ample context and transparency, and to pursue new technologies that deter inauthentic behavior.

Contemporary Disinformation

Disinformation exists in many forms but is generally characterized by a deliberate attempt to mislead and deceive people. American engineer and YouTuber Destin Sandlin interviewed Renée DiResta, a 2019 Mozilla Fellow in Media, Misinformation, and Trust, who said that perpetrators of disinformation can be both corporations and individuals and have two primary motivations: financial and ideological.[1] That is, they usually seek to extract advertisement revenue from their manufactured content, or try to manipulate public opinion (usually for financial or political gain). This reality is not surprising, given that money is power in our global economic system. Disinformation is not new, either, for its earlier forms were often government-sponsored propaganda. Propaganda has taken various forms, with earlier examples being pamphlets posted in town squares. This model's cost per impression ratio is high; governments would design and manufacture these pamphlets for the masses, but only people who were walking by and happened to stop would see them. Computers and computer graphics have helped pave the way for the next generation of disinformation: synthetic media, or algorithmically created/modified media. Contemporary disinformation is far more dangerous because it is created fully autonomously using A.I. that can synthesize and spread it without a single human interaction. No longer must curious people seek out and read propaganda; it is now being delivered to peoples' inboxes and newsfeeds, without consent, and backed by social features that make engaging with it easier than ever before. Disinformation can now reach more people with comparatively little cost and effort.

Industry-grade tools like Adobe Photoshop (released in 1988) and Adobe After Effects (released five years later in 1993) gave humanity the ability to manipulate images and videos at

[1] Sandlin, Destin W. 2019. *Manipulating the YouTube Algorithm.*

large scale for the first time. No longer was it as expensive or time-consuming to cut and paste physical pieces of paper for mass production—we could now produce near-perfect computer-generated imagery (CGI) from the digital realm because all the computationally-intensive rendering took place by a computer. However, the use of Adobe's robust software programs is still costly and offers an overwhelmingly vast selection of features. Only a team of talented graphic design artists can, with time and money, learn how to doctor images and videos convincingly. This phenomenon is why CGI is typically solely used by large-scale production companies with massive budgets like Marvel Studios, whose budget for *Avengers: Endgame* was $356 million.[2]

As with all forms of nascent technology, the landscape is rapidly changing. With the advent of community-driven content-generating algorithms, the process of creating disinformation is automated. Developers from all over the world are researching, refining, and optimizing these algorithms and are making them free to the public. Instead of needing an expensive computer graphics company to alter the visual data with proprietary tools, we can have free algorithms do the job just as well—if not better. FaceSwap, for example, is an open-source algorithmic face-swapping tool that anyone with an internet connection can utilize.[3] The software is maintained by over sixty contributors who have built an infrastructure that takes this highly technical process and simplifies it to a matter of (1) gathering image data for a subject and (2) metaphorically pressing "go." Real-Time Voice Cloning is another open-source tool that can not only clone a voice from mere minutes of a sample sound but can also make that cloned voice

[2] n.d. Avengers: Endgame.

[3] https://github.com/deepfakes/faceswap

speak whatever the author desires.[4] The software has received rave reviews[5] and spawned over 2,800 open-source variations. Research and progress in these spaces are booming. For little-to-no cost, anyone can use these algorithms to produce deepfakes that look or sound utterly genuine to our senses.

We are observing a shift in resources away from expensive creativity and talent toward inexpensive training data and processing power. Data and computational speed are resources that are becoming more accessible among individuals and corporations all over the world. Data volumes and processing speed are, in theory, boundless resources that only increase with every technological innovation. Moore's law, for example, is the observation that computational speed roughly doubles every two years. This observation has remained true since the 1970s. As history suggests, processing speed and the quantity and quality of data are only improving with every technological innovation. Further, cloud services like Amazon Web Services allow individuals without tremendous computing power to rent computing power from data centers that do. We can be sure that deepfakes will continue to be invented more rapidly and with higher fidelity.

The Role of Artificial Intelligence

As these algorithms improve, they have become, what many would consider, "artificially intelligent." A.I. (sometimes called machine intelligence) is intelligence demonstrated by machines, contrasting the "natural" intelligence exhibited by humans and other animals. Since machines have been demonstrating infant forms of intelligence since the first calculator, people colloquially use the term "A.I." to describe machines or algorithms that mimic cognitive

[4] https://github.com/CorentinJ/Real-Time-Voice-Cloning

[5] Arik, Sercan O., Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. *Neural Voice Cloning with a Few Samples.*

functions that we typically associate with fellow humans, including learning and problem-solving abilities.

The full scope of A.I. can be broken down into artificial narrow intelligence (ANI, or "weak A.I.") and artificial general intelligence (AGI, or "strong A.I."). ANI refers to a machine or algorithm's ability to perform a single task extremely well. Most contemporary A.I. takes this form with examples including the Global Positioning System GPS, automated content curation systems, spam filters, speech to text applications, mobile check depositing, fraud prevention, image recognition, and plagiarism detectors. AGI is a concept based on the premise that machines or algorithms can be made to think like humans and function similarly to the human brain. Examples of AGI might include more evolved autonomous virtual assistants (e.g., Google Assistant, Alexa, or Siri) and fully autonomous driving systems.[6] Disinformation-generating software is classified as ANI, where the focus of all these technologies is in single task execution with excellent user satisfaction. The focus of AGI, on the other hand, is to build technologies that can take the place of humans where complex decision making is required and narrowly focused systems are insufficient.

Algorithms that can learn and problem solve have quickly shown massive potential in the field of computer science and its associated consumer products. Companies from nearly every industry are jumping at any opportunity they can to incorporate A.I. into their processes. Advancements in ANI have led to smart speakers, smart lights, smart refrigerators, and even

[6] Note that there are no true examples of AGI because it does not currently exist. Researchers generally believe that an AGI would possess the transferability of knowledge to tasks other than a narrow domain. A genuine AGI would be able to competently play a game of chess while it runs an autonomous vehicle and derives new mathematical proofs, for example.

smart frying pans. This trend has led A.I. to become a buzzword in these innovation spaces, and marketing departments are quick to brand their products and services with it.

Stripping away the sensationalism, A.I. is intrinsically nothing more than a set of algorithms that take input data, perform a series of mathematical operations on it, and return a result. Further, and potentially the "secret sauce," if the output is not what its engineers desire, the algorithm tweaks its calculations to yield a better outcome next time through a process called backpropagation.[7] While certainly simple in practice, recent innovations in processing power have enabled these algorithms to work far better than they could have before.

When people combine ever-smarter A.I. with CGI, we find that its creation becomes far easier. Much of the skill and talent that comes along with years of training in graphics design has been automated away by algorithms that can crunch numbers faster than the best designers can create. The primary type of A.I. used in generating graphics is called a generative adversarial network (GAN). GANs are relatively new, as Apple Inc. Director of Machine Learning Ian Goodfellow invented the method in 2014.[8] Goodfellow thought that by having two adversarial algorithms contest with each other to generate realistic graphics, they could rapidly train each other to perform ever-better while circumventing any need for comparatively slower human interaction.

Consider an example GAN whose task is to generate convincing (fake) images of cats. The "generator" A.I. tries its best to generate images of cats from random input noise. This input

[7] Backpropagation is a commonly used tactic to train neural networks. The algorithm computes the gradient of the "loss function," a function that, in this case, calculates the error between the generated imagery and the authentic samples (since the goal is to come as close to authenticity as possible). The loss function is used with respect to the weights and biases within the neural network's nodes while prioritizing efficiency to quickly update weights and minimize loss (i.e., gradient descent). Backpropagation strictly refers to the algorithm that computes the optimization gradient—not how the gradient is used.

[8] Goodfellow, Ian J, et al. 2014. *Generative Adversarial Nets.*

data acts as a “seed” by which the generation takes place. Every output cat image corresponds to a specific input, and even slightly changing that input causes an entirely different cat to generate. The generator has not been told what a cat is supposed to look like, so its first few attempts to create them are wildly inaccurate. However, the “discriminator” A.I. has access to a database of cat images. The discriminator knows what cats are supposed to look like and coaches the generator while evaluating its outputs (through backpropagation). Not only does the discriminator try to guess whether the generated images are real cats, but it also produces an error score which the generator can use to infer how it can tweak itself to get closer to generating photorealistic cats. This zero-sum game[9] continues until the discriminator cannot distinguish the generator’s synthetic images apart from real ones. Since there is no human involved, the network is free to play against itself as quickly as its processors allow. This style of learning is called unsupervised learning since it requires no human supervision and can scale exponentially faster than any human-guided system could.[10]

Unsupervised learning is the same strategy employed by Google DeepMind’s AlphaZero. In December 2017, AlphaZero defeated the world chess champion Stockfish 8 program. Stockfish 8 was programmed without A.I., where its developers instead relied on centuries of human experience plus decades of computer experience to train it. Conversely, AlphaZero was built only to understand the rules and objectives of chess. Instead of teaching it from prior experience like Stockfish 8, AlphaZero learned strategies by playing against itself. After only four hours of training on Google’s supercomputers, AlphaZero did not lose once, beating

[9] In game theory, a zero-sum game is a mathematical representation of a situation where each participant’s gain or loss is balanced by the losses or gains of the other participants; a net change of zero. Poker is one example of a zero-sum game since the sum of the winnings equal the sum of the losses. Any two-player game, where there is one winner and one loser, is also zero-sum.

[10] Goodfellow, Ian J, et al. 2014. *Generative Adversarial Nets.*

Stockfish 8 with a record of 28-72-0. However, the resulting strategy is even more bizarre. Because AlphaZero had learned nothing from any human, many of its winning moves seemed unconventional to us. What chess masters initially thought to be mistakes turned out to be what we might call “genius” strategies.[11] This example demonstrates the staggering potential of these types of systems; in a matter of hours, AlphaZero went from utter ignorance to creative mastery without any human assistance.

The Role of Social Media

GANs have exploded in popularity due to their efficiency and are no longer being confined to generating imagery. GANs can now generate synthetic text, audio, or video, giving rise to “deepfakes,” a term which is a portmanteau of “deep learning” and “fake.” ThisPersonDoesNotExist.com is the result of a GAN that has generated images of human faces—the generated people do not exist (see Figure 1 for an example image).[12] American comedian Jordan Peele had his production company use GANs to create a deepfake of former U.S. President Barack Obama with Peele’s voice impersonation as a public announcement to help make people more aware of the emerging technology.[13] Unfortunately, not everyone is as good-intentioned as Peele, and many look to use GANs nefariously without transparency. Bad actors are prominently relying on social media as the means of delivering synthetic media to large audiences.

The choice to use social media is not accidental and relates to the underlying principles of the modern “attention economy.” These economics treat human attention as a scarce commodity,

[11] Harari, Yuval N. 2018. *Why Technology Favors Tyranny.*

[12] Horev, Rani. 2019. *Style-based GANs – Generating and Tuning Realistic Artificial Faces.*

[13] Vincent, James. 2018. *Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news.*

given that people have a limited supply of attention. Because social media companies tend to be free to use (to reach the broadest possible audience), they must be creative about how they make revenue. Selling user data to advertisers has become the primary source of income for companies like Google, YouTube, and Facebook because of how useful those insights can be when serving relevant advertisements to people. The more time a person spends engaging with a social media platform, the more user data that platform can collect and the more profitable that user becomes. Hence, social media companies typically seek to get users onto their platform and keep them there (often for as long as possible). They accomplish this goal with a series of algorithms that evaluate peoples' past activity to deliver new content that it thinks they will like. As of five years ago, these algorithms only needed as few as 300 of your "likes" to understand your personality traits better than anyone—even your spouse.[14] These technologies are what has made social media binges so frequent as well as a general addiction to social media by Internet-connected people.

Given our world's usage and reliance on social media, it has become critical to how the public consumes and shares current events. Since the goal of these platforms is to serve captivating—not necessarily truthful—content, this reality becomes problematic when the content curation algorithms serve up disinformation. Even after years of refinement from thousands of engineers from around the world, these algorithms will still sometimes curate disinformation if the content has enough engagement. Stepping back, we realize that social media platforms like Facebook, Twitter, and YouTube were never built to host an informed debate about the news; instead, they reward virality. The most sensationalist media will capture the most amount of attention. Curation algorithms notice these opportunities and will show the

[14] Quenqua, Douglas. 2015. *Facebook Knows You Better Than Anyone Else.*

media to even more people to maximize attention, which (tying in attention economics) translates to more profit.

To be successful at spreading disinformation, perpetrators must take advantage of virality by running massive operations of computer-guided (automated) user accounts to engage with their content: a practice called artificial engagement. Because social media algorithms use heuristics (like the number of "likes," comments, followers, and shares) to determine which pieces of content are popular, illegal groups with thousands of fake or stolen user accounts coordinate behavior on enormous scales to simulate engagement. To put this scale into perspective, in only three months of 2019, Facebook announced it had caught and removed 2.2 billion fake accounts.[15] Given that the company has identified roughly 2.5 billion monthly active (organic) users as of December 2019[16], there are almost as many fake accounts engaging in social discussions as there are real ones. These groups upload many instances of synthetic content, point them to each other, and have fake accounts and "click farms"[17] engage with the content.[18] The artificial engagement feeds into manufactured amplification, which is a phenomenon where the content curation algorithms are fooled into thinking that a piece of content is organically popular. The manufactured amplification eventually makes a piece of content popular enough that it rises above the general noise in the social media space and starts getting shown to real humans. From there, authentic people engage with the content, trust that the content is credible from context, including "like" counts, comments, and shares, and the

[15] Soto Reyes, Mariel. 2019. *Facebook removes 2.2 billion fake accounts in three months.*

[16] Facebook, Inc. 2020. *Facebook Reports Fourth Quarter and Full Year 2019 Results.*

[17] A click farm is a form of fraud, where a commercial enterprise employs many people to repeatedly click on and interact with online content to artificially inflate statistics of traffic or engagement.

[18] Sandlin, Destin W. 2019. *Manipulating the YouTube Algorithm.*

manufactured content goes viral. Over forty-eight hours, Twitter user Benjamin Strick identified an army of Chinese and Russian bots that were attacking a Chinese businessman who had a critical role in China's response to the 2019 novel coronavirus outbreak (COVID-19).[19] There are many more examples of this phenomenon on YouTube.[20]

From an engineering perspective, it is harder than ever to detect disinformation once real people begin engaging with it because their authentic engagement covers up the work done by the automated accounts.[21] Moreover, for every counter-measure, a social media platform invents to fight people gaming their system, it is not long before these people develop counter-counter-measures, which prompts the platform engineers to develop counter-counter-counter-measures, *ad infinitum*. To offer a brief example, social media companies realized that click farms could be tracked and thwarted through location detection; having a thousand new interactions from the same city in a matter of minutes is suspicious and flagged accordingly. To get around this, click farms now incorporate location-spoofing technologies to make their devices appear to come from all over the world, even if they are truly in a single basement. In this arms race, there is no winning. If there is a way to produce viral videos organically, there will always be a way to game that system synthetically.

Researchers at New York University recently identified hundreds of groups of Instagram users that systematically exchange "likes" and comments to similarly game the system.[22] They dubbed these groups "pods" and noted that they straddle the line between authentic and

[19] https://twitter.com/BenDoBrown/status/1255547411201691651

[20] The beginning of Sandlin's video highlights one example, and the rest of the video further explains more generally how these deceptive videos work: https://www.youtube.com/watch?v=1PGm8LslEb4

[21] Sandlin, Destin W. 2019. *Manipulating the YouTube Algorithm.*

[22] New York University Tandon School of Engineering. 2020. *"Researchers use machine learning to unearth underground Instagram "pods"."*

inauthentic engagement because, on the one hand, real people are interacting with each other, but, on the other hand, those interactions are not authentic—they interact with strangers' posts merely to have the favor returned some day. The practice of mutual engagement is called reciprocity abuse, and social media platforms have struggled against it for years. It has created a moral grey area that is particularly hard to combat due to it being difficult to determine whether two people engage with each other because they are close friends or if they merely want their content boosted.

Fortunately, there are companies like Astroscreen trying to fix this problem. In April 2019, the company successfully raised $1 million to detect social media manipulation. The company purportedly uses "coordinated activity detection, linguistic fingerprinting and fake account and botnet detection."[23] At this stage, however, it has not produced anything tangible for social media giants to utilize. The company's website consists of a thorough explanation of the problem with promises of a solution one day. Astroscreen's progress well-represents the whole industry: everyone is working quickly with no solution in sight.

Fuel for the Engine

A.I.'s ability to generate convincing imagery opens many creative avenues for artists and videographers. As CGI companies automate their processes, I believe movie budgets will decrease while the imagery quality increases. This lower barrier to entry will allow smaller groups or individuals to compete with massive corporations. With more players on the field, we could see the art industry radically transform. The $135 billion video game industry[24] would likely also see unprecedented levels of innovation in graphics.

[23] Butcher, Mike. 2019. *Astroscreen raises $1M to detect social media manipulation with machine learning.*

[24] Batchelor, James. 2018. *Global games market value rising to $134.9bn in 2018.*

I foresee a future where media distributors like Netflix could allow their users to put themselves into any movie or show. People would be able to give Netflix access to photos and videos of themselves (training data), and Netflix would use that data and a GAN to insert them into any scene they would like. They could even become the main characters!

Face-licensing celebrity endorsements could also become popular. Instead of tracking down and paying a celebrity to physically come into a studio and take photos or record commercials for product endorsements, the company could pay the celebrity for privileges to deepfake their faces onto bodies of other, low-budget actors. Of course, this future opens the door to personalized advertising, too. Instead of seeing an unfamiliar person model clothing or star in an infomercial, companies could use your face or the faces of people you recognize to better market to your interests. See Appendix A for a detailed list of other positive applications of GANs.

However, technological innovations tend to be neutral in the way they affect people in that they are tools. A hammer, for example, is not objectively useful or harmful; rather, it depends on how the user chooses to use the hammer—hammers can build and destroy. While AI-generated imagery has some arguably positive applications, its invention has opened the door for misuse. Many of the motivations to further automated CGI are grounded in ethical use cases (see Appendix A), but a growing number of people are becoming motivated by the anonymous, unethical uses: namely, involuntary pornography and ideological warfare.

Shortly after the invention of FakeApp, the A.I. face-swapping tool used to generate the viral Obama deepfake,[25] Reddit user "u/deepfakes" posted several pornographic videos built with the software to the subreddit "r/deepfakes." These videos depicted celebrities engaging in lewd

[25] Vincent, James. 2018. *Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news.*

acts without their consent and led to a series of community-sourced deepfaked pornography with celebrities including Daisy Ridley, Gal Gadot, Emma Watson, Katy Perry, Taylor Swift, and Scarlett Johansson.[26] Celebrities are a frequent target due to their popularity. The fact that there are thousands of photos and videos of these people makes for readily available training data, which then makes the generation far more realistic. However, everyday people should be worried about being targets too. With images and videos more frequently shared than ever before, people are finding their most intimate photos and videos leaked online. “Revenge porn” is the term used when intimate imagery of someone is distributed without their consent—and it is an epidemic with one in five Australians[27] and one in eight Americans[28] affected, according to recent reports. Although many countries might convict offenders, there is currently no federal law against revenge porn in the United States, mainly in part due to the U.S. Constitution’s First Amendment, which prevents the government from infringing on an individual’s right to freedom of speech and press.[29]

In the same way that celebrities are common targets for deepfakes, politicians are also vulnerable. Peele’s production company made it clear to the public that the Obama video was a deepfake, and they were transparent about how they created the video. The intention behind that deepfake was education and awareness. Not everyone will be this forthcoming, especially those

[26] Hawkins, Derek. 2018. *Reddit bans ‘deepfakes,’ pornography using the faces of celebrities such as Taylor Swift and Gal Gadot.*

[27] Henry, Nicola, Anastasia Powell, and Asher Flynn. 2017. *Not Just ‘Revenge Pornography’: Australians’ Experiences of Image-Based Abuse*, 4.

[28] Eaton, Asia A, Holly Jacobs, and Yanet Ruvalcaba. 2017. *2017 Nationwide Online Study of Nonconsensual Porn Victimization and Perpetration*, 11.

[29] This amendment is the basis for Facebook’s argument that its platform should continue to never require that its users’ content be truthful.

with malintent. This reality is why United States lawmakers say A.I. deepfakes “have the potential to disrupt every facet of our society.”[30]

Senator Marco Rubio believes that the ability to produce synthetic media is “the next wave of attacks against America and Western democracies,” citing a hypothetical situation where a deepfake depicting a political figure gets quickly promulgated by the media (and digested by our culture that is already susceptible to bias and believing “outrageous things”) that influences an election before authorities can identify the media as fake.[31] I believe that Senator Rubio’s worries are not ill-founded. One might easily imagine a scenario where manufactured content drops the day of a significant election that sways the final vote. Alternatively, and perhaps more gravely, consider that the United States government can launch a nuclear weapon in mere minutes,[32] and there are currently no deepfake detection tools operating on that time scale. If a deepfake of President Donald Trump declaring war on North Korea went viral, would we be able to react in time?

Shutting Down the Engine

Given that disinformation is such a complicated technical problem to solve, many argue for the cessation of the development of AI-powered disinformation tools. Their arguments follow the “just because we can, does not mean we should” mentality. While arguments for cessation intrinsically have merit, it is essential to evaluate them within the context of stopping or even

[30] Vincent, James. 2018. *US lawmakers say AI deepfakes ‘have the potential to disrupt every facet of our society’.*

[31] United States Senate. 2018. *At Intelligence Committee Hearing, Rubio Raises Threat Chinese Telecommunications Firms Pose to U.S. National Security.*

[32] Ludacer, Rob. 2018. *Here's how easy it is for the US president to launch a nuclear weapon.*

slowing A.I. development in this area. I argue that AI-powered disinformation campaigns are impossible to stop, and I will use nuclear weapons as an analogy to help illustrate my stance.

Under the direction of theoretical physicist J. Robert Oppenheimer, July 16, 1945, marked the detonation of the first-ever atomic bomb. With mushroom clouds 40,000 feet high and toxic levels of radiation that would linger for years, the world quickly learned how deadly they are. Paradoxically, that mere fact encouraged the mass development of these weapons from global superpowers. Why? Control and protection. If the world knows that the United States has nuclear weapons, they better make sure they also have them to protect their people if the United States were to attack. While they are at it, they ought to build more of them than any other country for the sake of controlling the technology’s power. After all, having control and protection are two, quite primal, motivators that have strong ethical arguments in favor of them: who does not want to be protected? Seeking control and protection are the reasons the Defense Advanced Research Projects Agency (DARPA) is investing millions of dollars into researching and creating deepfakes. Many lawmakers like Senator Rubio share the sentiment that disinformation can soon become a national security threat to our government and democracy and desire to get ahead of the problem by incubating, studying, and protecting against its use.

AI-powered disinformation campaigns are even more dangerous than nuclear weapons in many ways. Nuclear weapons require exorbitantly expensive, massive, and sophisticated machinery to create, operate, and detonate. They have also required a cohort of talented engineers and technicians to manage the infrastructure, though this is changing too, as more sophisticated ANI is performing better than its human counterparts. Nuclear weapon manufacturing, thus, requires a multi-million-dollar investment with oversight at every level. These characteristics are critical to note because they have made nuclear weapons somewhat feasible to regulate, although with great effort. The automated tools that make disinformation possible on a global scale, on the other hand, are algorithms that can be transported and copied millions of times without the slightest degradation. They are open source tools and libraries available for anyone to download and play with, young and old alike, regardless of their underlying hardware. Anyone with an internet connection can experiment with GANs, and we can only hope they choose to use the tools for good. For these reasons, I believe that it is sufficiently clear that the development of AI-powered disinformation campaigns cannot be stopped, slowed, or even regulated.

Any attempt to build a deepfake detection algorithm is similarly ill-fated. Because deepfakes are generated with GANs, the generator algorithm stops getting better when the discriminator can no longer identify the synthetic creation as fake and, thus, can no longer provide coaching for the generator. If you can make the discriminator better at detection, it is trivial for the generator also to get better; you merely swap out the old discriminator with the new and improved one and retrain the network until the generator has outsmarted the detector once more. Therefore, it will be impossible to solve the disinformation problem by analyzing the content directly. A more likely strategy to combat disinformation is through targeting the sources themselves and evaluating the context of the disinformation. However, as Sandlin and DiResta note, content delivery platforms like Google and Facebook may be equally unequipped to solve this problem because contextual clues are fickle.[33]

Google, whose mission is to “organize the world’s information and make it universally accessible and useful,”[34] plays an enormous role in monitoring and controlling the spread of

[33] Sandlin, Destin W. 2019. *Manipulating the YouTube Algorithm.*

[34] Google LLC. 2019. *About.*

disinformation. In a white paper published in February 2019 titled, "How Google Fights Disinformation," the company states that since the field of synthetic media is fast-paced and hard to predict, it is investing in research to understand how A.I. can help detect synthetic content as it emerges.[35] Since the production of disinformation is automated, the only sensible way to combat the problem with more automated systems. Google looks to fight disinformation on three fronts: making quality count, counteracting malicious actors, and giving users more context. In summary, this means that the company strives to continue delivering only the most relevant and authoritative content while developing counter-measures to detect malicious activity using different "signals." These signals are company trade secrets and change over time.[36] Lastly, it is giving users more context to show all sides of the story, including a balance of views and detailed source information. When you search for a current event or controversial topic, for example, Google aims to show you related news articles from other journalists to capture and offer a wide range of perspectives. I assume that the company refrains from disclosing specific defense strategies to keep malicious actors guessing; however, this is unclear and merely my speculation. Though I am excited to follow Google's progress in this war, I am not convinced its efforts will bear fruit. The company regularly removes content from its services and subsidiary companies like YouTube to comply with its company policies, legal demands, and government censorship laws,[37] yet Google still finds itself in the middle of international conflicts.[38]

[35] Google LLC. 2019. *How Google Fights Disinformation.*

[36] Ibid.

[37] Rosen, Jeffrey. 2008. *Google's Gatekeepers.*

[38] Conger, Kate, and Daisuke Wakabayashi. 2018. *Google Employees Protest Secret Work on Censored Search Engine for China.*

The Washington Post reports on researchers who have designed algorithms that analyze videos for “telltale indicators of a fake” such as specific light, shadows, and movement patterns.[39] Despite their progress, they claim that they remain overwhelmed. Hany Farid, a computer science professor and digital forensics expert, reports, “the number of people working on the video-synthesis side, as opposed to the detector side, is 100 to 1.”[40] The researchers note that although high-definition photos and videos are easiest to spot due to there being more opportunities for flaws to reveal themselves (i.e., there are more pixels that the A.I. has to get correct), most social media platforms compress photos and videos into smaller resolutions to make them faster to share (less data is faster to transmit). The telltale signs we might use to identify fakes, in other words, can be smudged, doctored, and even faked, both when they are created and as they evolve across platforms.

Considering these challenges, some researchers are trying other strategies by investigating cryptographic authentication systems that would fingerprint a photo or video the moment it is captured. That solution could work but is a tall order since it would require compliance from camera and microphone manufacturers around the globe.[41] Furthermore, only the original (source) material could be protected. Any postproduction work like lighting enhancements, cropping, or audio quality boosting would render the fingerprint worthless. I will revisit this proposed solution later in this paper in my “Recommendations” section.

[39] Harwell, Drew. 2019. *Top AI researchers race to detect ‘deepfake’ videos: ‘We are outgunned’.*

[40] Ibid.

[41] Ibid.

Many other one-off projects exist to combat the disinformation crisis. Shallow[42] is an open-source deepfake detection tool open for all to help improve. However, the project appears to have been abandoned, since, at the time of writing, activity stopped on May 22, 2018. FALdetector[43] attempts to detect photoshopped faces by scripting Adobe Photoshop. Photoshop has features to edit a person's face by warping the photo, and FALdetector tries to detect warping as well as provides suggestions on how it thinks the original image looked. The results of this project are promising but still experimental.

The Implications

Social Media

Being a brand-new technology, no one is sure how to respond to the sudden wave of manufactured content. Each social media platform has its philosophy regarding which types of content is and is not allowed to thrive on their platforms—examining each of the significant players sheds visibility into the current state of the response from the industry.

Reddit responded to the revenge porn deepfakes by banning the "r/deepfakes" subreddit altogether, as involuntary pornography is against their terms of use agreement. This subreddit, however, was also home to a plethora of positive use cases and research. Alas, Reddit deemed it necessary to censor the entire subreddit. These types of community-driven platforms rarely see censorship on this level. The move sparked a massive controversy, with many people wondering what makes deepfaked pornography different than traditional look-alike nude pictures and videos of celebrities.

[42] https://github.com/mvaleriani/Shallow

[43] https://github.com/peterwang512/FALdetector

Facebook, on the other hand, launched a home-grown program to fight revenge porn.[44] They ask worried people to send their intimate photos to Facebook so they can register and block them from ever getting posted onto the platform. All of Facebook's privacy scandals aside, people who still choose to engage will have their photos "hashed." Hashing is a one-way mathematical function that produces a numerical fingerprint for the input data. Any piece of data can be hashed relatively quickly, but the reverse is computationally infeasible. If a new photo is uploaded, and it matches one of the banned hashes, Facebook's algorithms will automatically detect and remove it.

There are many problems with this program. Since the resulting hash is a numerical fingerprint specific to a piece of input data, changing that input data even the slightest amount results in a completely new hash. In a photograph with millions of pixels, changing a single pixel will have no visual difference and create a brand-new hash. The same is true for cropping or applying filters or edits to a photo. Facebook or anyone cannot store every permutation of a photo and, thus, is trivial for malicious actors to break Facebook's system.

At the time of writing, there are no safeguards in place preventing people from uploading compromising content (and doctoring it) to target an individual. Even once the content is deemed fake, there is no guarantee that Facebook will remove it. On June 11, 2019, Vice News reported the existence of deepfaked videos of Mark Zuckerberg, Kim Kardashian, and President Donald Trump. Facebook, in a statement, said that it would not remove the fake videos because "we don't have a policy that stipulates that the information you post on Facebook must be true."[45] Instead, the company said it would treat those videos "the same way we treat all misinformation

[44] O'Brien, Sara Ashley. 2018. *Facebook's controversial 'revenge porn' pilot program is coming to the US, UK.*

[45] Harwell, Drew. 2019. *Top AI researchers race to detect 'deepfake' videos: 'We are outgunned'.*

on Instagram.[46] If third-party fact-checkers mark it as false, we will filter it from Instagram's recommendation surfaces like Explore and hashtag pages."[47]

However, Facebook's stance notably weakened when it faced pressure to remove COVID-19-related misinformation. On April 16, 2020, the company announced new initiatives to help its users find credible COVID-19 information while reducing the prevalence of bad actors.[48] Of those initiatives, the company is partnering with fact-checking groups from all over the world to flag potentially problematic posts and reduce their distribution. The company's response is surprising, given how reluctant it had been to remove other instances of prior misinformation. Despite its claims, the company evidently does put a limit on free speech. Facebook will continue working with fact-checkers to enforce the accuracy of information on specific topics, and it is up to the rest of the world to dictate what those topics are.

Governments

The United States government is taking a different type of preventative action. On June 12, 2019, Representative Yvette Clarke (D-NY) proposed the Defending Each and Every Person from False Appearances by Keeping Exploitation Subject (DEEPFAKES) to Accountability Act of 2019 to the House of Representatives.[49] The legislature takes steps to criminalize synthetic media by requiring anyone who creates it to disclose somehow that the content is fake. The bill suggests using "embedded digital watermarks" and "clearly readable text" appearing on said fake

[46] Facebook owns Instagram

[47] Shieber, Jonathan. 2019. *Facebook will not remove deepfakes of Mark Zuckerberg, Kim Kardashian and others from Instagram.*

[48] Rosen, Guy. 2020. *An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19.*

[49] Clarke, Yvette. 2019. *H.R.3230 - Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019.*

imagery. Per Seattle-based writer and photographer Devin Coldewey's report, "the act would create a task force at the Department of Homeland Security that would form the core of government involvement with the practice of creating deep fakes, and any counter-measures created to combat them."[50] The bill could also be used to criminalize tampering with deepfake disclosures, meaning that anyone caught removing these disclosures would be penalized.

The DEEPFAKES Accountability Act is a good start, and I am glad to see Congress taking preventative action. However, the bill suffers significant limitations in practicality. It seeks to criminalize the production of synthetic media without transparency from its creator that the media is intended to be fake. This transparency would take the form of a watermark or metadata. This law is unenforceable because anyone creating these fake photos and videos for nefarious use will not attach their name to it, making these people no easier than they are currently to track down. Further, watermark- and metadata-based markers are trivial to remove. If someone has the computing power to render someone else's face onto a body, they likely can edit over a disclosing watermark or crop the deepfaked media altogether to hide such disclosures. As Coldewey puts it, "as soon as [a] piece of media leaves the home of its creator, it is out of their control and very soon will no longer be in compliance with the law."[51]

Ideological Challenges

Until social media platforms and governments can solve the problem of disinformation, democratic societies should be anxious. German philosopher Immanuel Kant argued that we are all rational creatures and that we ought to use our reasoning when acting in life instead of letting others think for us. However, a rational agent cannot act rationally when equipped with false

[50] Coldewey, Devin. 2019. *DEEPFAKES Accountability Act would impose unenforceable rules — but it's a start.*

[51] Ibid.

information. While many agree that democracies are ethical because they grant non-felon citizens of the requisite age a right to vote, the system quickly becomes unethical when the victims of disinformation exercise their rights to vote and do so with a poisoned knowledge base. In this situation, Kant would argue, the person could not act rationally because they were unknowingly influenced by bias or an otherwise misrepresentation of ideas. Indeed, these new disinformation-producing technologies create fundamental challenges to the core assumptions of liberal democracies.

Even non-democratic governments are at risk of disinformation. Presently, every world government is run by a group of people, and any person can fall victim to disinformation or perpetuate its use as a propaganda machine. We are often quick to highlight all the ways everyday people will struggle against disinformation in their newsfeeds, but disinformation has no biases—it will affect everyone, including members of government, regardless of skin color, gender, ethnicity, or cultural background. Disinformation is a global problem and will likely require a global solution.

When Facebook says that it will not remove disinformation from its platform, it cites issues of free speech, noting that, in the United States, people have the right to express things that are not true. Though this is not entirely true, as some expression is criminal, for example, referencing bombs in public places like airports and schools can result in serious legal consequences due to infringing on the safety and security of others. To this end, some authoritarian states might have an easier time controlling disinformation compared to places like the United States. Though, the primary risk for those governments' citizens is that the government can also use disinformation to its advantage by controlling the public narrative.

Perhaps the conversation we should be having is if disinformation ever justifies restrictions on free speech and creative expression?

Recommendations

While synthetic media generation is still in its infancy, we are all moderately safe. However, this status could change overnight, especially with technology pioneer Hao Li predicting that “perfectly real” deepfakes will come to fruition in one or two years.[52] Will we reach a point as a society where political information becomes so corrupt that democratic processes become illegitimate? I believe that we are dangerously close to this point and must do everything we can to fight disinformation to protect epistemology. Given that the development of automated disinformation campaigns is unstoppable and getting better by the day, there must be increased technological investment from social media companies, including Google, Facebook, and Twitter, to provide additional context and transparency to the consumer. Tools that capture our material world for digital representation, including cameras and microphones, can also undergo technological investment for transparency. Lastly, we can also invest in new technologies that increase the cost of spam to deter nefarious use. In the meantime, the first step to overcoming the challenges outlined in this paper is merely becoming aware that disinformation exists. I will use this section to outline each of my recommendations, beginning with raising awareness.

[52] Stankiewicz, Kevin. 2019. *'Perfectly real' deepfakes will arrive in 6 months to a year, technology pioneer Hao Li says.*

Awareness Through Digital Literacy

Based on the evidence provided in this paper, I do not believe that any amount of platform-specific moderation, censorship, or deepfake detection will stop people from effectively spreading disinformation. Therefore, I believe that our most significant line of defense of disinformation is our awareness of its existence. Sandlin believes that all forms of disinformation provoke us "to fight with our neighbors."[53] He understands that the world-class engineers behind these social media companies are working as hard as they can to combat this problem with every tool at their disposal. However, until we change as a society toward political grace and away from being quick to fight, we will continue getting fooled.[54] If you see an article, image, audio clip, or video whose message is to pit groups of people against each other, treat it as a red flag that there is likely more to the picture subject to equal consideration. Take care to ensure that you understand the full 360-degree view of an issue before forming or defending strong opinions.

I am advocating for digital literacy training through this awareness. When I refer to "digital literacy," I include learning how to leverage our technologies to find, evaluate, create, and communicate information while also being able to identify disinformation and other harmful digital practices (e.g., spam, phishing, malware, and scams). One form of digital literacy found in companies all over the world is employee training that teaches and encourages people to identify suspicious emails, links, or attachments while forwarding (or reporting) to the information technology department before opening. Outcomes from this digital literacy training would include the practiced familiarity of regularly and frequently cross-referencing information. By habitually entertaining and investigating new perspectives, we will become a more sensitive and

[53] Sandlin, Destin W. 2019. *Manipulating the YouTube Algorithm.*

[54] Ibid.

informed society that is better equipped to coexist with pervasive disinformation. Those most likely to object to such training aspirations include people seeking to increase the amount of fear, insecurity, and conspiracy within a group of people. These people prefer an ignorant and susceptible population that can easily be controlled with fear, for fear is an especially compelling motivator due to its deep roots in our survival-driven biology.

For any form of digital literacy to be effective, it must be taught to young people in primary and secondary education. As the world continues to go evermore digital, it will be increasingly important for humans to become adept at navigating the noisy and confusing virtual world. Further, those skills can be refined as necessary through regular training in private businesses. Company executives ought to find it beneficial to incorporate some form of digital literacy training into the workplace, where the company's employees can learn how to recognize and report disinformation and other harmful or deceptive digital practices. If we can teach the next generation of humans how to anticipate and navigate the complexities of the virtual world, identify biased or sensationalist media, and appreciate divergent thought, we will be better equipped to pursue truth over harmful ideological manipulation.

Transparency Through New Technologies

Beyond the awareness of disinformation, the dogma that the digital realm is a representation of the physical world must be eroded. A bridge that perfectly connects digital information to its representation within the physical world may never be built; therefore, we ought to understand that anything in digital form in no way guarantees that it accurately represents the material world. The creation of this hypothetical bridge is one such technological advancement that I would welcome innovation within, although I foresee no way to accomplish this feat.

Though my first recommendation regarding digital literacy training is feasible in theory, it will suffer challenges in practice. Awareness in people can only spread so quickly, and human behavior is notoriously difficult to change. It could be many years before the general public is aware of disinformation and actively screens each piece of content before consumption. While digital literacy should be our chief goal, we should invest resources into making social platforms smarter about recognizing and flagging polarizing media in the meantime. All opinions regarding sociopolitical situations fall on a spectrum, and showing that spectrum in-line with the media would help the layperson understand the full scope of ideas. As these ANI algorithms get smarter, it might one day be possible to show where on the spectrum the media falls and provide context for why.

Ultimately, context is nuanced with many layers and intricacies. Our current best forms of context relate to numerical attributes about the media: namely, how many "likes," comments, and shares the content has as well as its publisher's "like," subscriber, or follower count. The problem with these contextual identifiers is that they can be purchased. Each of the preceding trust metrics can be artificially farmed, which means that our current tools for understanding context are easily subjected to monetary influence. To solve this problem, we must turn to contextual clues that cannot be purchased. We must invest in and build new technologies that try to capture the abstract contextual clues that cannot be purchased, including the journalist's potential biases (those relating to their career, ideologies, demographic, socioeconomic status, and political influence), the publishing company's biases (those relating to the company's organization structure, financial supporters, customer base, political influence), their parent or partner (if applicable), and any financial motivations that any of those people or organizations may have.

Work done by human and algorithmic fact-checkers should continue to be highlighted so that the consumer understands what has and has not been vetted. YouTube, for example, has recently brought fact-check panels to searches and videos in the United States.[55] The idea is to make factual evidence for controversial or rapidly evolving topics more apparent and accessible. WhatsApp, a product of Facebook, has added in-app features to allow people to swiftly search the web for the texts, images, or videos that they have received for more context.[56] Google, too, recently decided to make its advertisers verify their identities through a rigorous process.[57] While we have come far, there is still plenty of opportunity and research in this field to show ample context and transparency to consumers.

Increased transparency could take the form of reporters releasing unedited footage as a footnote to the polished news report. Original photos, audio clips, and video recordings could be provided alongside its recording device's cryptographic fingerprint (recall this concept from earlier). Anyone could cross-reference the key to the purported original documentation to confirm details about the physical device that captured the data. To address the counterargument that I proposed earlier in this paper, manufacturers of recording equipment would need to build this cryptography feature into each of their devices. Unlike a seamless software update that can be retrofitted for any device, this upgrade would need to be built into the hardware to ensure maximum security. While this idea is certainly not viable globally across all recording equipment, these cryptographic features could come with the devices that journalists would be most likely to use (i.e., professional-grade recording equipment). In the short term, we can all

[55] Newton, Casey. 2020. *YouTube brings fact-check panels to searches in the United States.*

[56] https://techcrunch.com/2020/03/21/whatsapp-search-web-coronavirus/

[57] https://www.blog.google/products/ads/advertiser-identity-verification-for-transparency/

benefit from having cryptographic proof that a photo or video genuinely came from a physical recording device without postproduction alterations. As this technology becomes more practical, it might reach its way into amateur-grade cameras and smartphones. When employed thoughtfully, this transparency could help form the foundation of the bridge connecting our material world to the digital realm.

Increase the Cost of Spam

As I alluded to earlier in this paper, part of the reason that social media platforms are the targets of disinformation campaigns is because of how cost-effective it is for malicious actors to reach a broad audience. Therefore, another solution to limiting coordinated inauthentic behavior is to increase the cost of spam. If it becomes too expensive for automated accounts to engage with disinformation, it will fizzle out. How might a computer system disincentivize spam?

In 1975, Israeli biologist Amotz Zahavi proposed a hypothesis now known as the handicap principle to explain how evolution could provoke honest and reliable signaling between animals having survival motivations to deceive each other.[58] The principle suggests that the most reliable signals must be the costliest to the signaler and require something that could not be afforded by an individual with less of a particular trait. Receivers of said signals have confidence that they indicate quality because signalers with inferior qualities cannot afford to produce such wastefully extravagant signals. For example, luxuriously loud peacock feathers make the bird more vulnerable to predators (e.g., easier to spot and harder to evade) and could, therefore, be a handicap. However, the peacock is signaling to potential peahen mates that it has survived despite having luxurious feathers and hence must be more fit and attractive than others. We do

[58] Zahavi, Amotz. 1975. *Mate selection—a selection for a handicap.* Journal of Theoretical Biology, 205–214.

this as humans, too, when we purchase extraordinarily expensive items to signal that we are desirable because we can afford to spend our wealth on such items. In the digital realm, there are two primary ways to signal quality through scarcity: computational work and digital currency (or cryptocurrency).

Bitcoin and other decentralized cryptocurrencies are perfect examples to illustrate costly signaling. Bitcoin is a decentralized ledger that keeps track of which virtual wallets own how much of the cryptocurrency. Because there is no central authority to ensure the reliability of the leger, Bitcoin uses proof of computational work (proof-of-work, or PoW) to signal reliability. Machines from all over the world are running simple hashing functions (recall from earlier) to digitally “sign” authenticity to a list of transactions.[59] This process, referred to as “mining,” is extremely slow, and, as of March 7, 2020, the machines on the bitcoin network were performing 123 terahashes (123,000 billion hashes) per second—its highest ever.[60] Even though mining presently requires over 77 terawatt-hours[61] of energy to sustain (more than some countries), it is computationally trivial to prove that the work took place.[62] In Bitcoin’s case, the transaction list with the most amount of work proved is considered the “true” list. For someone to attack the network, they would need to control the majority share of the network’s processing power to prove that they did the most work. At the time of writing, it would cost over $800,000 per hour[63]

[59] Live blockchain demonstration for further understanding: https://andersbrownworth.com/blockchain/block

[60] https://www.blockchain.com/charts/hash-rate

[61] https://digiconomist.net/bitcoin-energy-consumption

[62] For the sake of brevity within this paper, I am grossly oversimplifying the beautiful intricacies that make consensus over decentralized networks possible. For an in-depth look into the technology and further clarification, consider watching 3Blue1Brown’s 26-minute video: https://www.youtube.com/watch?v=bBC-nXj3Ng4.

[63] https://www.crypto51.app/

to rent the processing power needed for this attack. Bitcoin has effectively made nefarious use too costly to be financially viable.

Digital currencies can also be used to signal importance and authenticity of online media. For example, Baemail (stands for “before anything else mail”) is an email service that allows its users to attach digital currency to emails they send.[64] Recipients can see how much currency is attached to each of their incoming mail and get paid by engaging with the message. Following the handicap principle, the more expensive emails signal increased importance and authenticity (proof-of-worth). Spam, on the other hand, is less likely to be opened, read, or responded to, since the digital payment reward is lower or zero. Of course, this methodology only works if the digital currency has perceived value because it is intrinsically worthless (the same is true for bitcoin).

These technologies exist and have been proving their worth in the outskirts of mainstream media. Social media platforms can take a page from this book and implement similar incentive structures around posting on their platforms. However, these technologies are not silver bullets. In the case of proof-of-work, exorbitant amounts of electricity are consumed, contributing to substantial carbon emissions, which may prevent it from global adoption. In the case of proof-of-worth, content context is still purchasable, which means it is still subject to political influence. Increased investment in these areas will grant researchers the opportunity to identify other, more economically and environmentally sustainable approaches to increasing the cost of spam. Rewarding urgent or authentic content has translational benefits in reducing disinformation.

[64] https://baemail.me/

Conclusion

Disinformation has always existed in various forms throughout generations—from pamphlets in town halls to newspapers and social media newsfeeds. Whether creating deepfakes or misrepresenting media through incorrect context, disinformation lurks. Deepfakes, being the most significant advancement in the production of synthetic media, pose grave concerns for people and governments around the globe. With deepfaking technologies so accessible to the average person, its creation and research are unstoppable. We must act globally now to prevent catastrophic issues of epistemology in the foreseeable future.

As I have outlined in this paper, detection is not an ultimate solution to these emerging problems alone. We must prepare for a future when any article, photo, or video can look just as real as any other by building the infrastructure to allow for this reality to occur safely. This infrastructure is two-faceted: awareness from consumers and transparency from providers. People must be aware that they cannot safely take anything at face value and must be able to discover and consume other viewpoints easily. Companies must invest in better tools for consumers to verify the authenticity of the media and continue to thwart malicious actors seeking to take advantage of unsuspecting victims.

With AI-powered disinformation being such a new phenomenon, the field is rapidly changing. I suspect that this paper may already be out of date. There is much more work to be done to identify further promising technological investments that can help people see and understand critical contextual clues behind any article or post.

I have made an underlying assumption throughout this paper that our society wants to know when something is falsely portrayed as real. Though I do believe this assumption to be the case, I understand that individual people might not be so concerned. When leisurely scrolling through social media feeds, people want to feel comfortable and at ease. The average person tends to believe what they want to believe based on what makes sense in their conception of the world. Those who seed disinformation understand these tendencies and use our data from social media platforms to determine what genres of disinformation we will be most vulnerable to believing and sharing. Considering that there is mild interest in truth going against far more powerful incentives toward untruth, digital literacy can only be so helpful. Social media platforms have a social responsibility to limit the spread of harmfully deceptive information such that our collective society is better off. By developing new features and tools that show ample context behind articles and posts, these companies might be able to tip the scale in society's favor: toward truth.

Figures

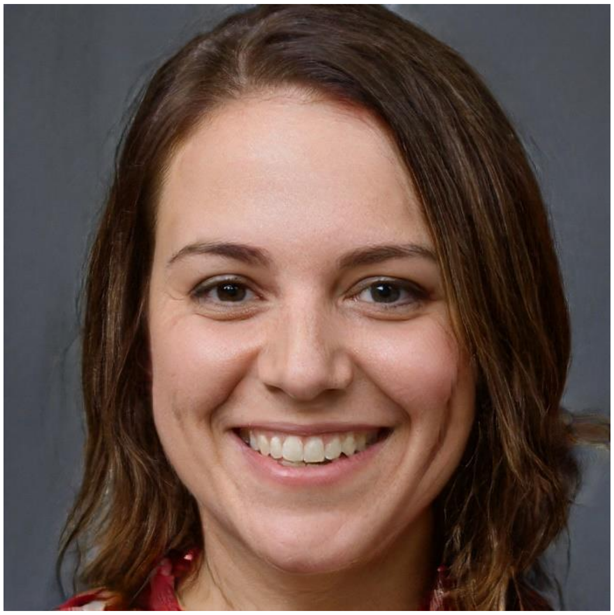

Figure 1: GAN-generated image by thispersondoesnotexist.com.

Appendix A

A longer list of positive use-cases for GANs from Jason Brownlee (Brownlee 2019):

- Generate Examples for Image Datasets
- Generate Photographs of Human Faces
- Generate Realistic Photographs
- Generate Cartoon Characters
- Image-to-Image Translation
    - Translation of semantic images to photographs of cityscapes and buildings.
    - Translation of satellite photographs to Google Maps.
    - Translation of photos from day to night.
    - Translation of black and white photographs to color.
    - Translation of sketches to color photographs.
    - Translation from photograph to artistic painting style.
    - Translation of horse to zebra.
    - Translation of photograph from summer to winter.
    - Translation of satellite photograph to Google Maps view.
    - Translation of painting to photograph.
    - Translation of sketch to photograph.
    - Translation of apples to oranges.
    - Translation of photograph to artistic painting.
- Text-to-Image Translation
- Semantic-Image-to-Photo Translation
    - Cityscape photograph, given semantic image.
    - Bedroom photograph, given semantic image.
    - Human face photograph, given semantic image.
    - Human face photograph, given sketch.
- Face Frontal View Generation
- Generate New Human Poses
- Photos to Emojis
- Photograph Editing
- Face Aging
- Photo Blending
- Super Resolution
- Photo Inpainting
- Clothing Translation
- Video Prediction
- 3D Object Generation

Bibliography

Arik, Sercan O., Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. *Neural Voice Cloning with a Few Samples.* Department of Computer Science, Cornell University, Ithaca: Cornell University. https://arxiv.org/abs/1802.06006.

n.d. *Avengers: Endgame.* IMDb.com, Inc. Accessed May 10, 2019. https://www.boxofficemojo.com/movies/?id=marvel2019.htm.

Batchelor, James. 2018. *Global games market value rising to $134.9bn in 2018.* Gamer Network. December 18. Accessed May 12, 2019. https://www.gamesindustry.biz/articles/2018-12-18-global-games-market-value-rose-to-usd134-9bn-in-2018.

Bishop, Katie. 2020. *A.I. in the adult industry: porn may soon feature people who don't exist.* Guardian News & Media Limited. February 7. Accessed April 10, 2020.

Brownlee, Jason. 2019. *18 Impressive Applications of Generative Adversarial Networks (GANs).* Machine Learning Mastery Pty. Ltd. July 12. Accessed July 18, 2019. https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/.

Butcher, Mike. 2019. *Astroscreen raises $1M to detect social media manipulation with machine learning.* April 18. Accessed August 2, 2019. https://techcrunch.com/2019/04/18/astroscreen-raises-1m-to-detect-social-media-manipulation-with-machine-learning/.

Clarke, Yvette. 2019. *H.R.3230 - Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019.* The United States of America. June 12. Accessed June 16, 2019. https://www.congress.gov/bill/116th-congress/house-bill/3230.

Coldewey, Devin. 2019. *DEEPFAKES Accountability Act would impose unenforceable rules — but it's a start.* June 13. Accessed June 14, 2019. https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/.

Conger, Kate, and Daisuke Wakabayashi. 2018. *Google Employees Protest Secret Work on Censored Search Engine for China.* The New York Times Company. November 16. Accessed May 4, 2019. https://www.nytimes.com/2018/08/16/technology/google-employees-protest-search-censored-china.html.

Eaton, Asia A, Holly Jacobs, and Yanet Ruvalcaba. 2017. *2017 Nationwide Online Study of Nonconsensual Porn Victimization and Perpetration.* Department of Psychology, Florida International University, Miami, Florida: Cyber Civil Rights Initiative, Inc., 11. Accessed August 16, 2018.

Facebook, Inc. 2020. *Facebook Reports Fourth Quarter and Full Year 2019 Results.* January 29. Accessed January 31, 2020. https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-Fourth-Quarter-and-Full-Year-2019-Results/default.aspx.

Goodfellow, Ian J, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." *Neural Information Processing Systems Foundation.* University of Montreal. June 10. Accessed May 10, 2019. https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Google LLC. 2019. *About.* February 26. https://about.google/.

—. 2019. "How Google Fights Disinformation." *Google.* February 16. https://www.blog.google/documents/33/HowGoogleFightsDisinformation.pdf.

Harari, Yuval N. 2018. *Why Technology Favors Tyranny.* The Atlantic Monthly Group. August 30. Accessed May 7, 2019. https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/.

Harwell, Drew. 2019. *Top A.I. researchers race to detect 'deepfake' videos: 'We are outgunned'.* The Washington Post. June 12. Accessed June 18, 2019. https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/.

Hawkins, Derek. 2018. *Reddit bans 'deepfakes,' pornography using the faces of celebrities such as Taylor Swift and Gal Gadot.* The Washington Post. February 8. Accessed May 4, 2019. https://www.washingtonpost.com/news/morning-mix/wp/2018/02/08/reddit-bans-deepfakes-pornography-using-the-faces-of-celebrities-like-taylor-swift-and-gal-gadot/.

Henry, Nicola, Anastasia Powell, and Asher Flynn. 2017. "Not Just 'Revenge Pornography': Australians' Experiences of Image-Based Abuse." RMIT University, Melbourne, Australia, 4. Accessed August 15, 2019.

Horev, Rani. 2019. "Style-based GANs – Generating and Tuning Realistic Artificial Faces." *Lyrn.AI.* December 26. Accessed March 3, 2019. https://www.lyrn.ai/2018/12/26/a-style-based-generator-architecture-for-generative-adversarial-networks/.

Ludacer, Rob. 2018. *Here's how easy it is for the U.S. president to launch a nuclear weapon.* Insider Inc. November 14. Accessed May 1, 2019. https://www.businessinsider.com/nuclear-bomb-launch-procedure-us-government-president-2017-11.

New York University Tandon School of Engineering. 2020. "Researchers use machine learning to unearth underground Instagram "pods"." Accessed May 1, 2020. https://engineering.nyu.edu/news/researchers-use-machine-learning-unearth-underground-instagram-pods.

Newton, Casey. 2020. *YouTube brings fact-check panels to searches in the United States.* Vox Media, LLC. April 28. Accessed April 30, 2020. https://www.theverge.com/2020/4/28/21239792/youtube-fact-check-panels-videos-united-states-misinformation-covid-coronavirus.

O'Brien, Sara Ashley. 2018. *Facebook's controversial 'revenge porn' pilot program is coming to the U.S., U.K.* Turner Broadcasting System, Inc. March 23. Accessed August 14, 2018. https://money.cnn.com/2018/05/23/technology/facebook-revenge-porn/index.html.

Quenqua, Douglas. 2015. *Facebook Knows You Better Than Anyone Else.* January 19. Accessed December 8, 2019. https://www.nytimes.com/2015/01/20/science/facebook-knows-you-better-than-anyone-else.html.

Rosen, Guy. 2020. *An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19.* April 16. Accessed April 17, 2020. https://about.fb.com/news/2020/04/covid-19-misinfo-update/.

Rosen, Jeffrey. 2008. *Google's Gatekeepers.* The New York Times Company. November 28. Accessed May 9, 2019. https://www.nytimes.com/2008/11/30/magazine/30google-t.html.

Sandlin, Destin W. 2019. *Manipulating the YouTube Algorithm - (Part 1/3) Smarter Every Day 213.* Google LLC. March 31. Accessed April 3, 2019. https://www.youtube.com/watch?v=1PGm8LslEb4.

Shieber, Jonathan. 2019. *Facebook will not remove deepfakes of Mark Zuckerberg, Kim Kardashian and others from Instagram.* June 11. Accessed June 13, 2019. https://techcrunch.com/2019/06/11/facebook-will-not-remove-deepfakes-of-mark-zuckerberg-kim-kardashian-and-others-from-instagram/.

Soto Reyes, Mariel. 2019. *Facebook removes 2.2 billion fake accounts in three months.* Insider, Inc. May 28. Accessed December 18, 2020. https://www.businessinsider.com/facebook-removed-22-billion-fake-accounts-2019-5.

Stankiewicz, Kevin. 2019. *'Perfectly real' deepfakes will arrive in 6 months to a year, technology pioneer Hao Li says.* CNBC LLC. September 20. Accessed September 21, 2019. https://www.cnbc.com/2019/09/20/hao-li-perfectly-real-deepfakes-will-arrive-in-6-months-to-a-year.html.

United States Senate. 2018. *At Intelligence Committee Hearing, Rubio Raises Threat Chinese Telecommunications Firms Pose to U.S. National Security.* May 15. Accessed February 18, 2019. https://www.rubio.senate.gov/public/index.cfm/press-releases?ID=B913F422-DC4F-4F19-A664-D9CE70559F87.

Vincent, James. 2018. *U.S. lawmakers say A.I. deepfakes 'have the potential to disrupt every facet of our society'.* Vox Media, Inc. September 14. Accessed June 19, 2019. https://www.theverge.com/2018/9/14/17859188/.

—. 2018. *Watch Jordan Peele use A.I. to make Barack Obama deliver a PSA about fake news.* Vox Media, Inc. April 14. Accessed May 8, 2019. https://www.theverge.com/tldr/2018/4/17/17247334.

Zahavi, Amotz. 1975. *Mate selection—a selection for a handicap.* Journal of Theoretical Biology, 205–214. doi:10.1016/0022-5193(75)90111-3.