Artificial Intelligence and Its Disinformation Campaigns

Dylan W. Wheeler

University of New Hampshire

Summer Undergraduate Research Fellowship

September 30, 2019

Table of Contents

Introduction

The term that describes how we can be confident in what we know is epistemology. It is the complete theory of knowledge from its methods to its validity and scope. Epistemology is the exploration of what separates justified belief from opinion. Ignorance, on the other hand, is usually defined as the lack of knowledge or information. However, ignorance can be strategically manufactured. In 1995, Stanford University professor Robert Proctor and linguist Iain Boal coined the term "agnotology" to describe the strategic and purposeful production of ignorance.[1] Today, our world is full of agnotology. We are living in a society that barrages social media giants like Facebook, Twitter, YouTube, and others with complaints about fake news and demands that these companies need to improve content management. Such complaints are justified, given malicious actors have figured out how to game the system and trick the benignly-engineered algorithms into propagating fake news and other forms of manufactured (synthetic) media. In this paper, I will argue that artificial intelligence (AI) is expediting the manufacture of synthetic information, which presents a danger to our world's political systems and informational integrity. I will analyze the present state of AI, explore how these algorithms facilitate the generation of synthetic information, evaluate their uses, argue that its development cannot be stopped, and forecast the social and political implications of this reality.

Artificial Intelligence Has Many Forms

AI (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

[1] Boyd, Danah. 2019. *Agnotology and Epistemological Fragmentation.* Data & Society Research Institute. April 26.

colloquially use the term “AI” to describe machines or algorithms that mimic cognitive functions that we typically associate with fellow humans, including learning and problem-solving abilities. 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 AI into their processes. This trend has led AI 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, AI is intrinsically nothing more than one or more 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. While certainly simple in practice, recent innovations in processing power have enabled these algorithms to work far better than they could have before.

There are so many different “artificially-intelligent” algorithms that researchers have broken them down into subcategories: machine learning (ML) and deep learning. ML is a subset of AI and is a technique to achieve AI through algorithms trained with data. Deep learning is a subset of ML inspired by the structure of the human brain. These data structures are typically called artificial neural networks (ANN) for this reason. Examples of AI include automated content curation, spam filters, voice to text applications, product recommendations, mobile check depositing, fraud prevention, pattern and image recognition, and plagiarism detectors.

Artificial Neural Networks

ANNs, like our brains, are made up of neurons. These neurons are arranged in layers, which are used to process data. The three types of layers are input, hidden, and output. Input data is broken down and mapped to individual neurons in the first (input) layer of the ANN. At the other side of the network is the last (output) layer of neurons, which are mapped to more complex concepts. Hidden layers reside in between the input and output layers and vary in size and number.

For example, imagine an image classification system where you give an ANN a square image of a handwritten digit, 28 pixels wide, where the objective is to identify which digit (0-9) it is. The ANN maps each of the 784 (28^2) pixels of that image to a different neuron in the input layer. The output layer, in this example, would contain only ten neurons (one for each digit) and only one will "light up" at the end of processing, corresponding to the digit that the algorithm thinks it has identified.

Along the way from the input layer to the output layer are hidden layers of neurons. Each layer is interconnected to its neighboring layers through "channels."[2] Neurons and their associated channels will activate their neighboring neurons based on a mathematical "activation function" that will activate if a combination of biases and weights cause it to trigger.[3]

These biases and weights are arbitrary numbers that the algorithm continuously adjusts during training. When a network is created, all the neurons of the hidden layers are configured with a random bias and weight. Training, in this example, consists of teaching the network what handwritten numbers are by feeding it many example images and indicating what the result should be for each of them. The algorithm makes calculated optimizations to the hidden layers' neurons' biases and weights to get the output layer of neurons to fire correctly. After enough

[2] Simplilearn. 2019. *Deep Learning In 5 Minutes.* June 3.

[3] Ibid.

training, the algorithm can properly perform image recognition on new images that were not contained in the training data. See Figure 1 for a topology of the ANN in this example.

While ANNs are currently the most efficient way to process unstructured data, they suffer limitations: namely, they require a substantial volume of data to train since there are anywhere from dozens to millions of neurons to adjust, each with varying biases and weights. For this example network, training might require thousands of 28x28 handwritten digits before the network is accurate enough to process new data reliably. All these images, neurons, biases, and weights require an enormous amount of computation, translating to expensive computing hardware and time—anywhere from a few hours to many months. By scaling this example up to a fully self-driving car ANN, one could imagine the sheer volume of data, processing power, and time would take to build an accurate[4] network. With each batch of training, the network's randomly adjusted neurons get better at hitting the right outcomes. This simulated evolution of a neural network lends appreciation for the billions of years of evolution our brains went through before their neurons became as capable as they are today.

Generative Adversarial Networks

Another AI technique and form of deep learning is generative adversarial networks (GANs). GANs are relatively new, as Apple Inc. Director of Machine Learning Ian Goodfellow created the concept in 2014, improving on some of the limitations of ANNs.[5] Goodfellow noticed that with a traditional ANN, humans must supply training data to the network. This process becomes a problem at scale because humans must carefully prepare this training data, as

[4] Acceptable levels of accuracy will vary per ANN. The accuracy expected from a full self-driving car ANN would be considerably higher than an experimental character recognition ANN due to increased stakes.

[5] Goodfellow, Ian J, et al. 2014. "Generative Adversarial Nets." *Neural Information Processing Systems Foundation.* University of Montreal. June 10.

any mistake will throw off the network. The larger the scale, the greater the volume of data the engineers must prepare. This task, of course, takes time and effort. Goodfellow sought after a way for the algorithm to generate its own labeled training data without human supervision.

The solution is adversarial machine learning. The objective of adversarial ML is to have two adversarial algorithms contest with each other to complete an objective (see Figure 2 for a network topology). The idea is to have one algorithm attempt to fool another algorithm by generating false, yet convincing, training data. Adversarial ML involving generation in this way are categorized as GANs.

Consider an example GAN whose task is to generate convincing (fake) images of cats. One algorithm, the "generator," tries its best to generate images of cats from random input noise. The input layer of this network is random, and the output layer of neurons needs to output color values of an image that resembles a cat. The network does not know what a cat is supposed to look like, so its first few attempts are wildly inaccurate. However, the other algorithm, the "discriminator," has access to the internet, knows what cats are supposed to look like, and coaches the generator by evaluating its outputs. 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 its neurons' biases and weights to get closer to generating photo-realistic cats.

This zero-sum game[6] 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 processor allows. This style of learning is called unsupervised

[6] 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.

learning since it requires no human supervision and can scale faster than any human-guided system could.[7] For perspective, Google’s AlphaZero chess algorithm beat the world’s former-best computer, which had access to centuries of human experience and strategies, in a mere four hours—all because it used adversarial ML to play against itself without the help of any human.[8]

In addition to images, GANs can also generate synthetic text, audio, or video, giving rise to “deepfakes.” The term is a portmanteau of “deep learning” and “fake,” since GANs are a type of deep learning used to produce artificial outputs. ThisPersonDoesNotExist.com is the result of a GAN that has generated images of human faces—the generated person does not exist (see Figure 3 for an example image).[9] Comedian Jordan Peele’s production company used GANs to create a deepfake of former U.S. President Barack Obama with Peele’s voice impersonation as a public announcement to make people more aware of the emerging technology.[10]

FaceSwap[11] is an active, open-source project specializing in face-swapping deepfakes, which over sixty people have contributed to date. The project has a multitude of guides explaining how to set up the algorithm to superimpose someone’s face onto someone else’s body in both photo and video rendering. Real-Time-Voice-Cloning[12] is another active and open-source project that can clone a voice in real-time. The project boasts its ability to take a short input

[7] Goodfellow, Ian J, et al. 2014. "Generative Adversarial Nets."

[8] Harari, Yuval N. 2018. *Why Technology Favors Tyranny.* The Atlantic Monthly Group. August 30.

[9] Horev, Rani. 2019. "Style-based GANs – Generating and Tuning Realistic Artificial Faces." *Lyrn.AI.* December 26.

[10] Vincent, James. 2018. *Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news.* Vox Media, Inc. April 14.

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

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

voice sample, immediately train itself, and produce high fidelity outputs. Both projects use GANs and are written in programming language Python.

The Benefits of Augmenting Computer Graphics with AI

Since we have had the ability to manipulate photos and videos for decades—with industry-grade tools like Adobe Photoshop (released in 1988) and Adobe After Effects (released five years later in 1993)—why has synthetic media suddenly become such a concern? The answer is a shift in the resources required to manufacture synthetic information. The use of Adobe's robust software programs is expensive and only a team of talented graphic design artists can, with time and money, doctor photos and videos convincingly. This phenomenon is why computer-generated imagery (CGI) is typically used by large-scale production companies with massive budgets like Marvel Studios whose budget for *Avengers: Endgame* was $356 million.[13] With the advent of GANs, this process is becoming automated. Instead of needing an expensive computer graphics company to alter the visual data, a free algorithm can do the job just as well—if not, better. We are observing a shift in resources away from expensive creativity and talent toward inexpensive training data and processing power. Data and speed are resources that are becoming more ubiquitous among individuals and corporations all over the world—and only increases 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.

The ability for AI to generate convincing imagery opens many creative avenues for artists and videographers. As CGI companies automate their processes, I believe movie budgets will

[13] n.d. *Avengers: Endgame.* IMDb.com, Inc.

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[14] would likely also see unprecedented levels of innovation in graphics.

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.

The Drawbacks of Augmenting Computer Graphics with AI

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 photos and videos could be helpful to society, its invention opens the door for misuse. I will explore

[14] Batchelor, James. 2018. *Global games market value rising to $134.9bn in 2018.* Gamer Network. December 18.

how AI is automating ideological manipulation via involuntary pornography and political disinformation.

Involuntary Pornography

Shortly after the invention of FakeApp, the AI face-swapping tool used to generate the viral Obama deepfake,[15] Reddit user u/deepfakes posted several pornographic videos built with the software to the subreddit r/deepfakes. These videos depicted celebrities engaging in lewd 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.[16] 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. Reddit responded to these posts by banning the r/deepfakes subreddit altogether, as involuntary pornography is against their terms of use agreement. Save pornography, this subreddit was 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, and the move sparked a massive controversy with many wondering what makes deepfaked pornography different than traditional look-alike nude pictures and videos of people.

Celebrities are not the only ones in trouble; everyday people should be worried too. With photo and video sharing more frequent than ever before, people are finding their most intimate photos and videos leaked online. “Revenge porn” is the term used when intimate photos or videos of someone are distributed without their consent—and it is an epidemic with one in five

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

[16] Hawkins, Derek. 2018. *Reddit bans ‘deepfakes,’ pornography using the faces of celebrities such as Taylor Swift and Gal Gadot.* The Washington Post. February 8.

Australians[17] and one in eight Americans[18] affected, according to recent reports. With no federal law against revenge porn in the United States, Facebook took the situation into their own hands with their pilot to fight revenge porn.[19]

The program is simple: send your intimate photos to Facebook so they can register and block them from getting posted onto the platform. With all the privacy scandals and struggles that Facebook has been having in recent years, sending your most private data to the company would seem risky to the average person. Should people actually do this, their photos will be "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. The idea is 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 particular 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. It is impossible for Facebook to store every permutation of a photo and trivial for malicious actors to break Facebook's system.

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

[18] 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.

[19] O'Brien, Sara Ashley. 2018. *Facebook's controversial 'revenge porn' pilot program is coming to the US, UK.* Turner Broadcasting System, Inc. March 23.

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 they would not remove the faked videos because "we don't have a policy that stipulates that the information you post on Facebook must be true."[20] Instead, the company said it would treat those videos "the same way we treat all misinformation on Instagram.[21] If third-party fact-checkers mark it as false, we will filter it from Instagram's recommendation surfaces like Explore and hashtag pages."[22]

Political Disinformation

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 with malintent who wish to spread disinformation; false information deliberately intended to mislead public opinion or obscure the truth. This reality is why United States lawmakers say AI deepfakes "have the potential to disrupt every facet of our society."[23]

[20] Harwell, Drew. 2019. *Top AI researchers race to detect 'deepfake' videos: 'We are outgunned'.* The Washington Post. June 12.

[21] Facebook owns Instagram

[22] Shieber, Jonathan. 2019. *Facebook will not remove deepfakes of Mark Zuckerberg, Kim Kardashian and others from Instagram.* June 11.

[23] Vincent, James. 2018. *US lawmakers say AI deepfakes 'have the potential to disrupt every facet of our society'.* Vox Media, Inc. September 14.

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.[24] I fear that Senator Rubio's worries are not ill-founded. The United States government can launch a nuclear weapon in mere minutes,[25] 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?

The United States government is not so sure we would, so they are taking 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 to Accountability Act of 2019 to the House of Representatives.[26] The DEEPFAKES Accountability Act takes steps to criminalize synthetic media by requiring anyone who creates synthetic media to disclose somehow that the content is fake. The bill suggests using "embedded digital watermarks" and "clearly readable text" appearing on said fake images and videos. Per Seattle-based writer and photographer Devin Coldewey's report, "the act would create a task force at the Department of

[24] United States Senate. 2018. *At Intelligence Committee Hearing, Rubio Raises Threat Chinese Telecommunications Firms Pose to U.S. National Security.* May 15.

[25] Ludacer, Rob. 2018. *Here's how easy it is for the US president to launch a nuclear weapon.* Insider Inc. November 14.

[26] 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.

Homeland Security that would form the core of government involvement with the practice of creating deep fakes, and any countermeasures created to combat them."[27]

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 criminalizes the production of synthetic media without transparency that the media is fake. This punishment is unenforceable because anyone creating these fake photos and videos for nefarious use will not attach their name to it. 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 watermark or crop the deepfaked media altogether. 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."[28]

Social Media Rewards Virality

Analysis of the drawbacks to AI-generated synthetic information points to the central problem this technology creates: uncertainty on a global scale. Malicious actors can use technologies like GANs to manufacture and distribute disinformation massively. I have laid the groundwork by explaining how synthetic information is created, analyzing the driving forces of this innovation by evaluating its use cases, and exploring its benefits and drawbacks. To frame my argument that the development of this technology is inevitable and unstoppable, I will explain how people are currently using disinformation on social media platforms like Facebook, Twitter, and YouTube.

[27] Coldewey, Devin. 2019. *DEEPFAKES Accountability Act would impose unenforceable rules — but it's a start.* June 13.

[28] Ibid.

Every software company has the same objective: get users onto the platform and keep them there—sometimes for as long as possible. Social media platforms accomplish this goal by a series of algorithms that evaluate your past activity to deliver new content that it thinks you will like. These technologies are what has made social media binges so frequent as well as a general addiction to social media by young people. Social media has become critical to how the public consumes and shares current events. This reality becomes problematic when these algorithms serve up disinformation without even knowing it, since the goal is to serve captivating content—not necessarily truthful content. Social media platforms like Facebook, Twitter, and YouTube were never built to host an informed debate about the news; instead, they reward virality.

American engineer Destin Sandlin found himself victim to disinformation and wanted to get to the bottom of it.[29] Sandlin spoke with an unnamed engineer at Google and Renée DiResta, a 2019 Mozilla Fellow in Media, Misinformation, and Trust, to begin his three-part YouTube video series about content manipulation. He is one of many people (colloquially known as "YouTubers") who rely on revenue from advertisements on his videos to support his family, so learning how YouTube's algorithms work is always advantageous for him. The more he understands how the platform will serve up his videos to his audience, the better he can tailor his content for the masses.

Perpetrators of disinformation can be both corporations and individuals. According to DiResta, they have two primary motivations: financial and ideological.[30] That is, they usually seek to extract advertisement revenue from their manufactured content, try to manipulate public opinion, or some combination thereof. To be successful at spreading their disinformation, they

[29] Sandlin, Destin W. 2019. *Manipulating the YouTube Algorithm - (Part 1/3) Smarter Every Day 213.* Google LLC. March 31.

[30] Ibid.

run a massive operation of manufactured content and use automated user accounts to engage with their content; a practice called artificial engagement. Since 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 can coordinate behavior on a massive scale to trick these algorithms. These groups upload many instances of synthetic content, point them to each other, and have fake accounts and "click farms"[31] engage with the content.[32] The artificial lift eventually makes a piece of content popular enough that it rises above the noise in the algorithm and starts getting shown to real humans. From there, authentic people engage with the content, believe the content has authentic credibility due to factors including like counts and comments, and the manufactured content goes viral.

From an engineering perspective, it is harder than ever to detect disinformation once real people begin engaging with it because, in a way, their authentic engagement covers up the work done by the automated accounts.[33] 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. In this arms race, there is no winning.[34] If there is a way to produce viral videos organically, there will always be a way to game that system synthetically.

[31] 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.

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

[33] Ibid.

[34] Ibid.

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.”[35] At this stage, however, it has not produced anything tangible for social media giants to utilize. Their 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.

Fighting Malicious Actors

Any attempt to build a deepfake detection algorithm is similarly ill-fated. Since 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. Therefore, it will be impossible to solve the disinformation problem by looking at the content directly. Zooming out, it may be possible to combat disinformation by targeting the sources themselves or 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.[36]

Google, whose mission is to “organize the world’s information and make it universally accessible and useful,”[37] plays an enormous role in monitoring and controlling the spread of

[35] Butcher, Mike. 2019. *Astroscreen raises $1M to detect social media manipulation with machine learning.* April 18.

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

[37] Google LLC. 2019. *About.* February 26.

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, they are investing in research to understand how AI can help detect synthetic content as it emerges.[38]

In other words, Google plans to fight algorithms with more algorithms. They look to fight disinformation on three fronts: make quality count, counteract malicious actors, and give users more context. The company strives to continue delivering only the most relevant and authoritative content. They want to continue developing counter-measures to detect malicious activity from analyzing different “signals.”[39] Lastly, they are giving users more context to show all sides of the story, including a balance of views and detailed source information. I assume they refrain from getting too specific to keep malicious users guessing. I am excited to follow Google’s progress on this war, but I am not convinced their efforts will bear fruit. The company regularly removes information from its services and subsidiary companies like YouTube to comply with its company policies, legal demands, and government censorship laws,[40] yet they still find themselves in the middle of international conflicts.[41]

The Washington Post reports on researchers who have designed algorithms that analyze videos for “telltale indicators of a fake” such as light, shadows, and movement patterns.[42] Despite all their progress, they say that they remain overwhelmed. Hany Farid, a computer-

[38] —. 2019. "How Google Fights Disinformation." *Google.* February 16.

[39] Ibid.

[40] Rosen, Jeffrey. 2008. *Google’s Gatekeepers.* The New York Times Company. November 28.

[41] Conger, Kate, and Daisuke Wakabayashi. 2018. *Google Employees Protest Secret Work on Censored Search Engine for China.* The New York Times Company. November 16.

[42] Harwell, Drew. 2019. *Top AI researchers race to detect ‘deepfake’ videos: ‘We are outgunned’.* The Washington Post. June 12.

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.”[43] The researchers note that, although high-definition photos and videos are easiest to spot due to more opportunities for flaws to reveal themselves, most social media platforms compress photos and videos into smaller formats to make them faster to share.

In light of this challenge, some researchers are investigating cryptographic authentication systems that would fingerprint a photo or video the moment it is captured (recall hashing from earlier). That solution could work but is a tall order since it would require compliance from camera and microphone manufacturers.[44] Furthermore, only the original (source) material could be protected. Any post-production work like lighting enhancements, cropping, or audio quality boosting would render the fingerprint worthless.

Many other one-off projects exist to combat the disinformation crisis. Shallow[45] 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[46] 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 what it thinks the original image looked like. The results of this project are promising but still experimental.

[43] Harwell, Drew. 2019. *Top AI researchers race to detect ‘deepfake’ videos: ‘We are outgunned’.* The Washington Post. June 12.

[44] Ibid.

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

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

The Case For AI-Powered Disinformation Tools

Given that disinformation is such a complicated technical problem to solve, many argue for the cessation of the development of AI-powered disinformation tools. These arguments follow the “just because we can, does not mean we should” methodology. While this argument intrinsically has merit, it is essential to evaluate it pragmatically. I will use the example of nuclear weapons to illustrate my argument that AI-powered disinformation campaigns are impossible to stop.

Under the direction of theoretical physicist J. Robert Oppenheimer, July 16, 1945, marks the detonation of the first-ever atomic bomb. With mushroom clouds 40,000 feet high and toxic levels of radiation that linger for years, the world quickly learned how deadly they are. Paradoxically, that mere fact encouraged mass development of these weapons. 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 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 in the development of deepfakes. Many lawmakers like Senator Rubio share the sentiment that disinformation can soon become a national security threat to our government and democracy and want 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 are exorbitant, massive, and require sophisticated machinery to operate and detonate. They also require a cohort of talented engineers and technicians to manage the infrastructure. Disinformation generators, on the other hand, do not tangibly exist. They are digital algorithms that can be emailed 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. 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 do not believe the development of AI-powered disinformation campaigns can be stopped, slowed, or even regulated.

Facebook, in partnership with Microsoft and others, is hosting the Deepfake Detection Challenge,[47] where they make deepfakes and award prizes for anyone who can algorithmically detect their falsity. The company is so committed that they plan to dedicate $10 million in resources to make it happen.[48] The vision is to add more financial incentive for people all over the world to invest their resources toward the collective goal of building a platform that can detect manufactured content.

Ideological Challenges

Democractic societies should be extremely worried about deepfakes. 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 you. However, a rational agent cannot act rationally when equipped with false information. While many agree that democracies are ethical because they give everyone a vote, it becomes unethical when the victims of disinformation are

[47] https://ai.facebook.com/blog/deepfake-detection-challenge

[48] Coldewey, Devin. 2019. *Facebook is making its own deepfakes and offering prizes for detecting them.* September 5.

exercising their rights to vote and are directly (negatively) influencing the government. In this situation, Kant would argue, the person did not act rationally because they were influenced by bias or an otherwise misrepresentation of ideas. 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.[49] To what degree does disinformation need to proliferate through society before we deem it unethical for humans to vote due to our inability to trust their biases and polarized views? I believe that we are already past that point and must do everything we can to fight disinformation to protect the nation's integrity.

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.

Conclusion

Given that the development of automated disinformation campaigns is unstoppable and getting better by the day, the only global solution I have determined is being aware that disinformation exists. I believe increasing our awareness for disinformation is a step in the right direction. Sandlin believes that all forms of disinformation take advantage of the one flaw in our

[49] Stankiewicz, Kevin. 2019. *'Perfectly real' deepfakes will arrive in 6 months to a year, technology pioneer Hao Li says.* CNBC LLC. September 20.

biology that gives us the desire to fight with our neighbors: “people literally make us hate each other, and then we turn around and give them our money.”[50] I do not believe any amount of platform-specific moderation, censorship, or deepfake detection will stop people from trying to spread disinformation. The only tools we have to fight disinformation are awareness and unity. Our propensities to separate people and ideas only fuel those with malintent. If we could come together and embrace all political ideas with grace and an open mind, we might be better off against this fight. Everything starts with a single person, and the more people that are actively aware of and calling out disinformation, the better our global humanity will be.

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

Figures

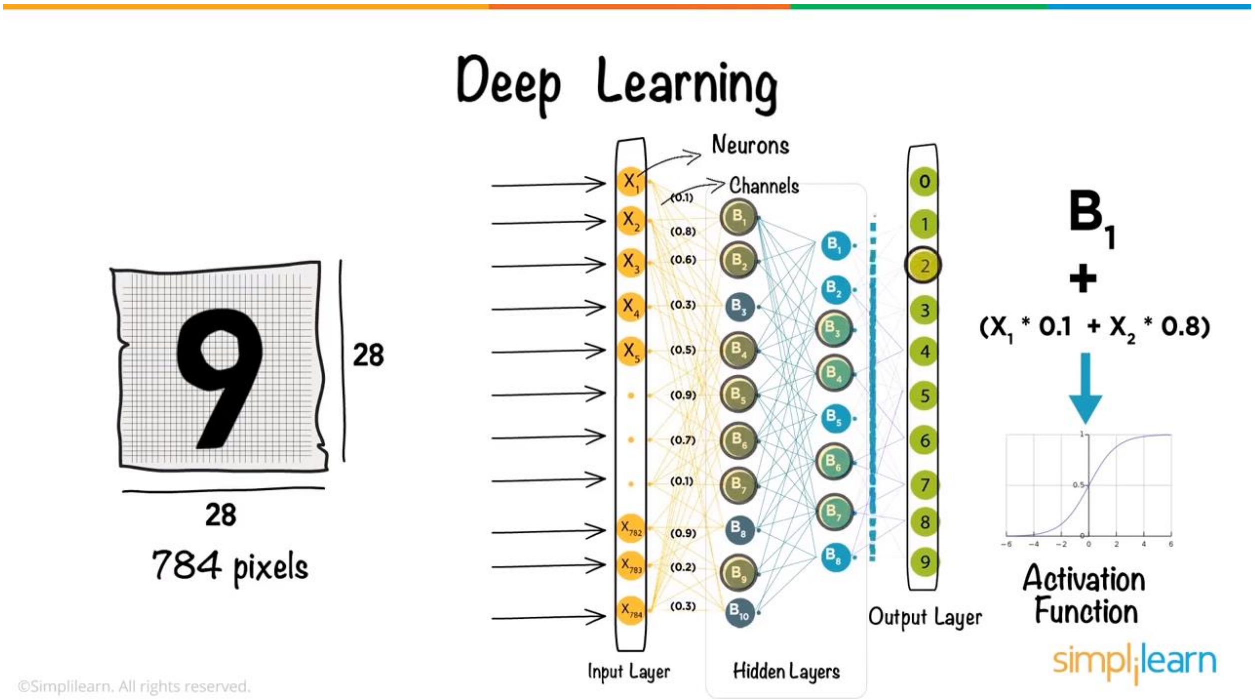

Figure 1: the topology of an artificial neural network (Simplilearn 2019)

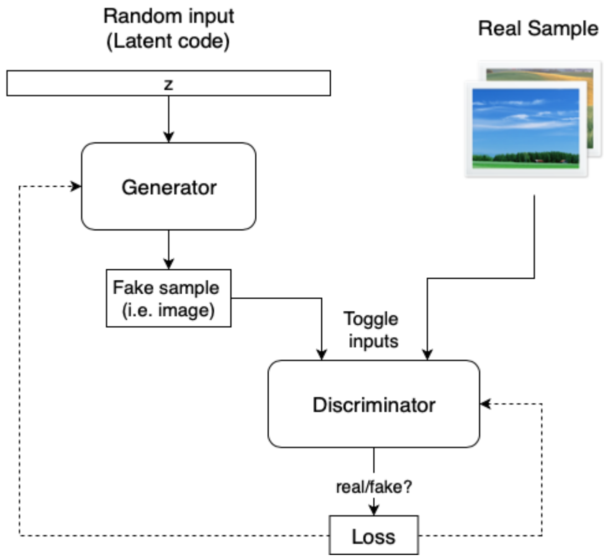

Figure 2: GAN cartoon overview (Horev 2019).

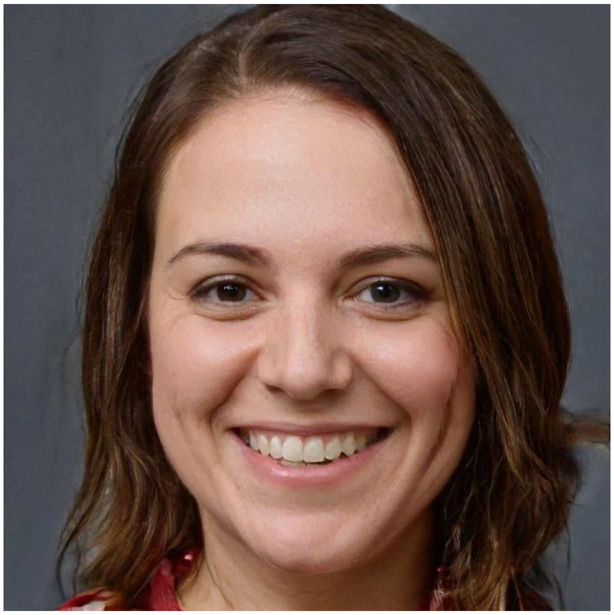

Figure 3: GAN-generated image by thispersondoesnotexist.com.

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

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.

Boyd, Danah. 2019. *Agnotology and Epistemological Fragmentation.* Data & Society Research Institute. April 26. Accessed May 8, 2019. https://points.datasociety.net/agnotology-and-epistemological-fragmentation-56aa3c509c6b.

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/.

—. 2019. *Facebook is making its own deepfakes and offering prizes for detecting them.* September 5. Accessed September 7, 2019. https://techcrunch.com/2019/09/05/facebook-is-making-its-own-deepfakes-and-offering-prizes-for-detecting-them/.

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.

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 AI 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 US 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.

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

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/.

Simplilearn. 2019. *Deep Learning In 5 Minutes | What Is Deep Learning? | Deep Learning Explained Simply | Simplilearn.* June 3. Accessed July 2, 2019. https://www.youtube.com/watch?v=6M5VXKLf4D4.

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. *US lawmakers say AI 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 AI 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.