Spring 4-12-2020

Living in a World Where Seeing Is No Longer Believing: Artificial Intelligence as a Disinformation Engine

Dylan Wheeler
University of New Hampshire

INQUIRY journal

www.unh.edu/inquiryjournal 2020

Research Article

Living in a World Where Seeing Is No Longer Believing: Artificial Intelligence as a Disinformation Engine

—Dylan Wheeler

I have been fascinated by technology since my childhood. In high school, I realized that I wanted to dedicate my life to the pursuit of technology and to unlock new ways it can help humanity grow and flourish. During my sophomore year at the University of New Hampshire (UNH), I became inspired to double major in philosophy in addition to my original coursework in information technology. Although seemingly contrasting majors, the two come together beautifully when studying how our society influences (and is influenced by) technological innovations. Recent groundbreaking advancements in artificial intelligence (AI) show no signs of slowing and get me excited about the future of computation, data processing, and automation. As the saying goes, however, with great power comes great responsibility. If we are not wise, these emerging automated technologies can disadvantage significant swaths of Earth's population through mass unemployment and systemic surveillance.

The author, Dylan Wheeler. *Photo by Jeremy Gasowski.*

My realizations that (1) technology has the capability to affect the lives of billions positively but that (2) this possibility is no guarantee and (3) we might not understand the can of worms we are opening led me to want to have a positive effect on the industry. After much deliberation with my mentor in the Department of Philosophy, Professor and Chair Nick Smith, I elected to use a Summer Undergraduate Research Fellowship (SURF) to study the nexus of my interests in emerging technologies and their social effects—more specifically, AI and its ability to wage ideological war. Although I might not know all the correct answers, my research will undoubtedly benefit high-tech CEOs, entrepreneurs, and engineers by encouraging them to consider the social effects of their investments today for the technologies of tomorrow.

Project Background

AI is intelligence demonstrated by machines. It is sometimes called machine intelligence, in contrast to the natural intelligence exhibited by humans and other animals. We consider the essence of intelligence as the ability to learn from and understand new situations in one's environment. People

colloquially use the term *AI* to describe machines or algorithms that mimic cognitive functions that we typically associate with people, such as learning, problem solving, and the skilled use of reason. AI is intrinsically nothing more than an algorithm (or a series of algorithms) that takes input data, performs a series of mathematical operations on it, and returns a result.

For the first time in history, automated processes (algorithms) can generate fake text, imagery, audio, and video so convincing that we humans are easily tricked into thinking it is real. Our society is on the verge of entering a post-truth future in which any form of digital media can be faked easily. With nothing more than a mere photo of a person, algorithms have been able to generate pornographic content featuring the person, false political statements from famous politicians, and anything else you could imagine. This is a reality that affects all of us, and, through my SURF and senior thesis in philosophy, I seek to understand how to solve (or, at least, mitigate) the problems it creates. My SURF work looked at the problems and questions that result from disinformation generated on massive scales using AI tools.

Artificial Intelligence and Deepfakes

One of the many AI techniques that enables the manufacture of synthetic media is the algorithm's ability to "learn" and automatically improve upon its calculations to yield a better outcome the next time. This learning component is the "secret sauce" that has enabled AI to become as important and pervasive as it is today. For example, we might want algorithms to generate convincing images of cats. To accomplish this goal, we would create one "generator" algorithm to try to generate images of cats from automatically generated random input data. This input data is abstract—it could be a timestamp, a random string of characters, or a series of numbers—and acts as a "seed" by which the generation takes place. Every output cat image corresponds to a specific input, and even changing that input slightly causes an entirely different cat to be generated.

The generator has not been told what a cat is supposed to look like, so its first few attempts to make an image of a cat are comically inaccurate. However, we create another algorithm, a "discriminator," that 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. We have these adversarial algorithms battle it out until the discriminator cannot distinguish the generator's synthetic images from real ones. Since these algorithms require no human intervention once started, they undergo unsupervised learning and can scale faster than any human-guided system could. In addition to images, this technology can generate text, audio, or even video, giving rise to deepfakes. The word *deepfake* is a portmanteau of *deep learning* and *fake* and refers to any AI-generated image, sound, or video of a subject.

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 false 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 images convincingly. With the advent of community-driven algorithms like those I've described, the process of creating fake media is becoming automated. Developers all over the world are researching, refining, and optimizing these algorithms and making them available for 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.

Moreover, companies like Amazon, Microsoft, and Google have built tools that enable developers to use their powerful processors and data warehouses for little cost (i.e., cloud computing). We are observing a shift in resources away from human creativity and talent toward processing power and data to train the discriminator—resources that are becoming ubiquitous 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.

Risks of Artificial Intelligence

New technologies tend to be neutral in the way they affect people in that they are tools. A hammer, for example, is not objectively beneficial or harmful; rather, it depends on how the user chooses to use the hammer. Although AI-generated news and media could be helpful to society, its invention also opens the door for misuse.

The ability of AI to automatically generate convincing imagery opens many new creative avenues for artists and videographers. As computer-generated imagery companies begin to 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 in the content production industry. With more players on the field, we could see digital art industries be radically transformed. The $135 billion video game industry would likely also see unprecedented levels of innovation in graphics.

The darker side of these algorithms involves involuntary pornography and ideological manipulation. Social discussion website Reddit had a user named u/deepfakes who posted several pornographic videos generated with deepfaking software to the subreddit "r/deepfakes." Without the consent of their subjects, these videos depicted celebrities engaging in lewd acts and led to a series of community-sourced deepfaked pornographic images of celebrities, including Daisy Ridley, Gal Gadot, Emma Watson, Katy Perry, Taylor Swift, and Scarlett Johansson. Celebrities are a frequent target of deepfaked media because of their popularity and the copious amounts of their image and video data available for free online. Reddit responded to the posts in this thread by banning the "r/deepfakes" subreddit altogether—including nonpornographic deepfakes—because involuntary pornography is against their terms-of-use agreement. Since Reddit rarely bans entire subreddits, the move sparked a massive controversy. Many people wondered what makes deepfaked pornography different from traditional look-alike nude pictures and videos of people. People, including myself, wonder if censorship is the most ethical defense against deepfakes.

In the same way that celebrities are common targets for deepfakes, politicians are also vulnerable. 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 influences an election before authorities can identify the media as fake (United States Senate, 2018). Senator Rubio's worries are not unfounded. The United States government can launch a nuclear weapon in mere minutes.

Further, there are currently no tools capable of detecting a deepfake on the time scale of minutes. If a deepfake of US President Donald Trump declaring war on North Korea went viral, would we be able to react in time?

Research Methods

For the duration of my SURF research, I worked from my home in Bow, New Hampshire, cotaught an advanced high school class on artificial intelligence in the Advanced Studies Program (ASP) at St. Paul's School in Concord, New Hampshire, and traveled to the University of New Hampshire's Durham campus to work closely with my mentor. For the duration of the summer, I combed through hundreds of current-event articles, dozens of code repositories, and any academic papers published on deepfakes and unsupervised AI that I could find. I wanted to understand not only the current state of these issues but also the public's knowledge of them. From there, I studied various open-source projects to learn the extent of these deepfaking software projects and ultimately uncovered several crucial details regarding deepfake software. Along the way, I have held dialogues with people young and old—from high school students at the ASP to university professors at UNH. Insights gained from talking with diverse groups of people have helped me frame recommendations that can work for everyone.

Key Findings and Observations

Every software company has the same objective: get users onto its platform and keep them there—usually 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. As of five years ago, these algorithms needed only as few as 300 of your "likes" to understand your personality traits better than anyone—even your spouse (Quenqua 2015). These technologies have made social media binges frequent and have caused a general addiction to social media by people of all ages. The students I worked with at the ASP confirmed my suspicions about this phenomenon. Further, much of the curriculum in the ASP's Mass Media class has evolved in recent years from conversations regarding traditional news sources to how social media platforms present curated newsfeeds to their users. These platforms have become critical to how the public consumes and shares current events. This reality becomes problematic when these algorithms serve up disinformation unbeknownst to their creators, since the goal is to serve captivating content—not necessarily truthful content. Facebook notably has no rules saying that content posted to the platform must be true (Harwell 2019). Social media platforms like Facebook, Twitter, and YouTube were never built to host an informed debate about the news; instead, they reward virality.

Everyday people like you and I can create disinformation with any number of the free tools available on the Internet. However, it takes a substantial coordinated effort to trick Facebook, Twitter, or YouTube's algorithms into thinking that real people are interacting with the manufactured content. A massive number of automated accounts is needed to spread the disinformation until enough of these fake interactions artificially lift the content to the eyes of real, unsuspecting people. These organizations are usually nefarious moneymaking schemes that accomplish this feat by programming thousands of hacked accounts to automatically engage with content through "click farms." These malicious groups of people typically acquire accounts through a myriad of social engineering strategies that get account owners (real people) to reveal their passwords. The hacked accounts are sometimes subsidized by other accounts that are autonomously created using fake email addresses. Programmers orchestrate their "farms" to like, comment, and share any piece of targeted media. Although this process can sometimes occur on a single machine, hundreds of recycled, discarded, or stolen phones can also be used to run the macros. A simple Google search of "buy YouTube subscribers" or "get Facebook likes" reveals merely the tip of the iceberg for the services that these illegal operations provide.

Click farms like these are hired to create fake likes for online content. *(Sources: Carr 2019, Equedia 2017).*

After the fake accounts have provided enough likes, comments, and shares for the content, the platform's algorithms start showing it to real people, thinking that it is engaging. From an engineering perspective, it is now 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 inauthentic automated accounts. Moreover, for every countermeasure that a social media platform invents to fight people gaming their system, it is not long before purveyors of disinformation develop counter-countermeasures, which prompts the platform engineers to develop counter-counter-countermeasures, 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 seems suspicious. To get around this, click farms use location-spoofing technologies to make their devices appear to come from all over the world. In this arms race, there is no winning. If there is a way to produce viral content authentically, there will always be a way to game that same reward system artificially.

Any attempt to build a deepfake detection algorithm is as ill-fated as any attempt to detect fake virality. Recall that since deepfakes are generated with competing algorithms, the generator algorithm stops getting better when the discriminator can no longer identify the synthetic creation as fake. If you can make the discriminator better at detection, the generator also will get better. Therefore, it will be impossible to solve the disinformation problem by examining the content directly. Zooming out on the issue, it may be possible to combat disinformation by targeting the sources themselves or evaluating the context of the disinformation. For example, a camera could cryptographically "sign" a photo it took with proof that the photo came from the camera's lens directly without alteration. However, any slight alteration to the photo (color corrections, flipping orientation, or cropping) would invalidate the digital signature, making this solution infeasible. Evaluating the context of disinformation is likely our best shot. This evaluation could include the quality of the media, the source, any conflict-of-interest disclosures, where the viewpoint of the media falls on the spectrum of views, whether it was peer reviewed, and other metrics. Unfortunately, as of now, that process requires human reason, is expensive, and is unsustainable at scale.

The *Washington Post* has reported on researchers who have designed algorithms that analyze videos for "telltale indicators of a fake" such as light, shadows, and movement patterns. Despite all their progress, they say that they remain overwhelmed by the technical challenge of detection (Harwell 2019). Hany Farid, a computer-science professor and digital-forensics expert, reports that "the number of people working on the video-synthesis side, as opposed to the detector side, is 100 to 1" (Harwell 2019). The researchers note that, although high-definition photos and videos are easiest to spot because there are more opportunities for flaws to reveal themselves, most social media platforms compress photos and videos into smaller formats to make them faster to share. The desired "portability" of smaller image files means that they are easier to manipulate and that it's harder to detect such manipulations.

Recommendations for the Future

Given that the development of automated disinformation campaigns is unstoppable and getting better by the day, we must consider the ways we can be smarter when evaluating news. I do not believe that any amount of platform-specific moderation, censorship, or deepfake detection software will stop people from trying to spread disinformation. The only immediate global solution for fighting disinformation that I have determined is an awareness of the problem. I believe that increasing our awareness of disinformation is perhaps the only universal step in the right direction until these platforms enact further technological investment in transparency.

Merely understanding that the content you are consuming could be false is the first step. Following this understanding, you will notice some telltale signs that what you are seeing is either fake or heavily biased. Most—if not all—forms of disinformation try to pit humans against each other by spreading false or biased ideologies for profit or power. Before forming an opinion, consider doing a quick Google search that could shed light on other perspectives. Think twice before sharing an article with an emotional headline. Notice when content is trying to pit people against each other. Help spread the word that disinformation is pervasive in our society.

The only tools that we humans possess to fight disinformation are our awareness and our unity. Our propensities to separate people and ideas only fuel those with mal-intent. If we come together and accept divergent political ideas with grace and an open mind, we might be better off in this fight

against ideological manipulation. Everything starts with a single person, and the more people who are actively aware of and calling out disinformation, the better our global humanity will be.

My research shows that our time to act is now. Our future will bring many challenges with informational integrity that may be insurmountable. The best thing we can do as citizens of the world is to have the courage to think for ourselves and use our rationality to reason through opposing sides of various issues. I am fortunate to have worked with brilliant high schoolers from all over New Hampshire this past summer to help spread my words of awareness and caution, and I sincerely hope that this article will help to accomplish the same goal by sharing my insight with you. Further, I am in the process of preparing in-depth technical recommendations for high-tech CEOs, entrepreneurs, and engineers, which I intend to defend in my senior thesis. Let us take full responsibility as rational agents to employ healthy skepticism, understand one another, and always pursue truth.

Thank you to my mentor, Dr. Nick Smith, for inspiring me to write about this topic. Had I not taken his classes in Science, Technology, and Society and Social and Political Philosophy, I would not have even considered these pressing problems. I am incredibly appreciative of the support I received from completing my foundational Summer Undergraduate Research Fellowship (SURF) project from Mr. Dana Hamel, Mr. and Mrs. Irving E. Rogers III, and Ms. Deborah Rogers Pratt. Finally, thank you to Mr. Terence Wardrop from the Advanced Studies Program at St. Paul's School for helping build my technical knowledge about artificial intelligence as well as to all the students from our class this past summer for your curiosity, willingness to grapple with these colossal ideas, and unique perspectives on these emerging technologies.

References

Carr, Sam. 2019. "What Is a Click Farm? The Quick Way to Thousands of Likes." PPC Protect. https://ppcprotect.com/what-is-a-click-farm/.

Equedia. 2017. "The Shady World of Click Farms." Equedia Investment Research. https://www.equedia.com/shady-world-click-farms/.

Harwell, Drew. 2019. "Top AI Researchers Race to Detect 'Deepfake' videos: 'We Are Outgunned." *Washington Post*. June 12. Accessed June 18, 2019. https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-....

Quenqua, Douglas. 2015. "Facebook Knows You Better Than Anyone Else". *New York Times*. January 19. Accessed December 8, 2019. https://www.nytimes.com/2015/01/20/science/facebook-knows-you-better-tha....

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....

Author and Mentor Bios

Dylan Wheeler, from Bow, New Hampshire, will graduate from the University of New Hampshire (UNH) in May 2020 with a bachelor of science degree in information technology and a bachelor of arts in philosophy. He is a Hamel Scholar in the Honors Program and will be graduating with honors in

both majors. Wheeler completed his research on artificial intelligence with a Summer Undergraduate Research Fellowship (SURF) grant, which laid the foundation for his philosophy senior thesis. “I have always been fascinated with the ways society interfaces with technology,” Wheeler says. “When I learned of these astonishing new deepfake technologies, I saw how serious the implications could be and wanted to learn everything I could about them.” He heard about *Inquiry* from his advisers at UNH and decided to submit his research to encourage a broader audience to engage with his work. “One of the most frustrating parts of this research was that new and sometimes conflicting information was literally popping up every day,” he says. “By the time I finished the summer, I realized that I had only scratched the surface of these nuanced and complex issues.” Wheeler plans to continue his work with his EdTech startups after graduation, noting, “I’d like to leave a positive mark on the technologies of tomorrow. Artificial intelligence is an unbelievably powerful tool but must be developed responsibly and thoughtfully.”

Nick Smith is professor and chairperson of the University of New Hampshire (UNH) Department of Philosophy. He has been at UNH since 2002. Before coming to UNH, Smith, who holds a JD and PhD, worked as a litigator for a private law firm and as a judicial clerk for the Honorable R. L. Nygaard of the United States Court of Appeals for the Third Circuit. He has published two books on his studies and contributes regularly to interviews in national media outlets. Smith says that working with Dylan was “a unique experience for me because his IT background and interest in artificial intelligence actually led me to create a new course, The Future of Humanity, so that we could continue discussing these issues. Dylan took the class and helped me design the class the first time I taught it. It was especially gratifying to have Andrew Ware (UNH ’18 and former *Inquiry* author) also involved in this process, because Andrew recently worked on a SURF with me on AI. He served as an example and mentor for Dylan on bridging philosophical research and emerging technology. Andrew and Dylan were part of the Department of Philosophy’s inspiration for a new major, Philosophy of Business, Innovation, and Technology.” Smith has mentored many undergraduate researchers and considers it “one of the genuine highlights of teaching at UNH.”