Bots & Deepfakes

AUGUST 2021

STRATEGIC MULTILAYER ASSESSMENT
Student paper series in collaboration with the TRADOC e-intern program

Author: Abdul-Rahman Kabbara
The College of William & Mary

This white paper presents the views and opinions of the contributing authors.
This white paper does not represent official USG policy or position.

Abdul-Rahman Kabbara, TRADOC e-Intern

The College of William & Mary

Abdul-Rahman Kabbara graduated from the College of William and Mary in 2017. He majored in Kinesiology and minored in Public Health. While at The College of William & Mary, Abdul-Rahman participated in the Middle Eastern Student Association, Amnesty International, and the Crew Club, and was placed on the Dean’s List. He was awarded the Chick-fil-a Scholarship award in 2013. He is proficient in Arabic and is trying to learn Turkish. In his free time, Abdul-Rahman likes to program code, brew coffee, and explore new hiking trails. He is currently pursuing studies on the interactions between the US and foreign nations and is interested in the impact and role of future technology on society. He served as part of the TRADOC Virtual Intern Program in 2021.

Bots & Deepfakes

Abdur-Rahman Kabbara[1]

This paper will explore the impact of two applications that *artificial intelligence* (AI) will have on information operations, bots, and deepfakes. Artificially created bots are automated accounts, while deepfakes are cases of manipulated visual and auditory content; both are powered by AI algorithms. Due to increasingly integrated and capable technology, the reach of AI applications is broadening. Bots and deepfakes are preferred tools of non-state and state actors due to their accessibility, wide reach, difficulty of detection, and difficulty of attribution (potential for blackmail). Digitally manipulated content, which is propagated from the likes of bots and deepfakes, are highly effective at generating false perceptions and diminishing trust in existing figures and institutions. These technologies have public facing consequences, and an examination of the ramifications of these applications will be explored. Possible solutions will be suggested to coordinate methods of restraint on these developing platforms and technologies whose influences and potential are just now beginning to be realized and thought out.

Introduction

In the summer of 1955, the *Dartmouth Research Project* explored whether intelligent human behavior could be reproduced by a machine which would later develop into the definition of AI (McCarthy et al., 1955). Machine learning is a field that came out of AI and is defined as the ability of machines to learn without being explicitly programmed to do so (Das et al., 2015). Today, the definition of AI has expanded to include creating automated programs that can perform well in complex domains through any means necessary, not just human-like means (Floridi, 2016). Bots and deepfakes are a part of this expansion. Bots are automated software programs powered by AI that operate on the internet and perform repetitive tasks. They can be used to imitate human users and infiltrate and influence communities online. Their ability to pose as users and quickly multiply allows for them to spread disinformation and polarizing ideology. Deepfakes are audio and visual content that have been either distorted or generated by AI and machine learning powered technology. Their effectiveness at manipulating content makes it indiscernible for casual viewers to identify as forged, and their believability can distort viewers' understanding of a specific situation or event. Targeted blackmail is

[1] *Contact Information*: ajkabbara@gmail.com

another use of deepfakes. Together, bots and deepfakes can undermine trust in “facts” and “reality” by attacking existing structures and figures while also bolstering controversial material to prominence, thereby endangering viewers into being manipulated later.

The difference between misinformation and disinformation is important to define for this paper. Misinformation is a broad category, in which false information is spread without the intention to mislead. However, disinformation is the spread of misleading information which is done intentionally to lead toward a purpose-driven outcome (Kluver et al., 2020). As global societies become progressively interconnected across technology platforms and these platforms become further integrated, the danger of disinformation is increasing (Metaxas, 2020).

Ubiquitous integration of technology allows for bots and deepfakes to have a wide reach and to be more influential in the day-to-day lives of people. For example, news traditionally was spread and received in the form of printed media, like newspapers, where editorial oversights were done to ensure that the content was factually accurate. In the present day, most news is received digitally through social media networks, which lack editorial oversight and are instead pushed by AI-powered algorithms. These online social networks and the content they contain can be accessed from integrated technological devices like smartphones, ensuring a wide reach. Another aspect of increased integration of technology is that an increasing amount of human life is being spent online. Social media platforms are the dominant methods of digital social interaction. In 2015, the Pew Research Center found that 65% of adults were using social networking sites, a tenfold jump from the previous decade (Perrin). As integrated technology increases in its ubiquitous reach, that number will continue to rise. Increased exposure to bots and deepfakes will be a result of this widened reach.

Bots

Bots can self-propagate, generate content on their own, collect information, and interact with human users. Machine learning can be used to automatically generate bot content and emulate the temporal patterns of humans via deep neural networks (DNNs). AI uses machine learning to train DNNs in a manner that is analogous to how the human brain functions. The more data a DNN is fed, the more efficient and capable it becomes (thereby making the bots themselves more efficient and capable as well). Many bots exist on social media platforms posing as credible accounts and are effective at imitating human users, but are faster, cheaper, and scalable. As a result, they are effective at tricking humans and the user-engagement algorithms that social media platforms run on. This enables bots to push out an impression or “trend” that some person or opinion is seemingly popular online. Potentially, bots can produce wide-scale manipulation of public opinion.

There are a wide variety of bot types. Some bots are assigned to do one task, such as posting content automatically. Others try to impersonate human users. Bots can infiltrate topically centered communities by recognizing and creating content centered around specific topics that enable the bots to gather the attention and trust of communities and be able to exert some form of influence over them (Yang et al., 2018). A botnet is a group of bots acting together. In coordination, this coordinated bot effort makes it difficult for anti-bot systems to detect bot behavior. A type of botnet is *fake-followers*, which is used to give off an impression that a certain user is popular online. This user can then use this impression to sway public opinion. An analysis done by Kai-Cheng Yang and their team found that bots on social media distort online conversations by bombarding the platform with an enormous amount of generated content (2018). Their analysis "suggests that bots were very effective at getting messages reshared in the human communication channels" (Yang et al, 2018). The ability for DNN bots to manipulate information and human behavior heightens the risk of misinformation and disinformation gaining prominence in society.

Bots can be used to spread disinformation and extreme ideology. Advances in AI make it feasible to design bots that sense, think and act cooperatively in social settings just like human beings. In the wrong hands, these bots can be used to infiltrate online communities, build up trust over time and then send personalized messages to elicit information, sway opinions, and call to action (Boshmaf et al., 2012). Boshmaf et al. in 2012, also characterized a form of social botnet attack called a Sybil attack where a set of fake identities (known as a Sybil) are used to join a targeted system multiple times to disrupt it. With the ability of bots to cause such disruption, it is even more important that we identify strategies and tools to discern and combat them.

The use of bots can already be seen in the last few years. In 2016, Russian-sponsored disinformation operations were able to successfully infiltrate and deepen existing social cleavages in the United States. An example of their operations was using bots posing as fake social media accounts sharing inflammatory content designed to stoke racial tensions (Chesney & Citron, 2019). Berger and Morgan (2015) found that a terrorist organization was utilizing bots on social media to spread its extreme ideology and promote terrorist propaganda. Botnets have been found that were misdirecting line discussion about the Syrian Civil War (Abokhodair et al., 2015). The accessibility and effectiveness of bots to expose the online public to such content will continue to make it a lucrative tool.

Deepfakes

The believability and accessibility of deepfakes are two reasons they are incredibly successful and powerful. The believability of deepfakes is growing increasingly worrisome as the technology behind it is getting better and fake content is becoming more believable. Also, advances in technology are making deepfakes easier to generate. The ability of social media platforms to provide individuals with the tools and technology to create and post such content on their platforms, makes it extraordinarily easy for deepfake distribution to reach a wide audience.

The danger in digitally manipulated content is in how susceptible humans are to being influenced by it. Fake content that is believable distorts trust and understanding of reality, making viewers vulnerable to further manipulation. Deepfakes' power lies in its ability to target the brain's visual system for misperception (Kietzmann et al., T. 2019). This "realism heuristic" effect allows deceptive videos to generate false perceptions more readily than deceptive verbal content. This is because audiovisual content resembles the everyday real-life experience more easily than text (Frenda et al., 2013; Sundar, 2018). Also, images and audiovisual content are easier to process and understand also because "metacognitive experience," where anything which elicits out some sort of feeling about our thinking can shape how we react and the manner of our responses (Schwarz et al., 2007). One example of such experience is "fluency", i.e., the concept that people are more likely to accept things as being true if they discern it as being familiar (Berinsky, 2017). Familiarity makes material easier to absorb and be perceived as being more credible. As Cristian Vaccari and Andrew Chadwick note in their paper on deepfakes and disinformation, when depicting well known public figures deepfake videos "intensify the already serious problem that fluency can be generated through familiarity, irrespective of the veracity of the video's content" (2020).

Deepfake content can be used in a variety of ways. For example, video footage from real protests or violent incidents can be doctored via the use of deepfake technology to misattribute the video content and use it to suggest that it happened elsewhere. Reuters found 30 instances of such use during an increase in foreign tension between India and Pakistan in 2019 (Westerlund, 2019). Mika Westerlund suggests that a foreign intelligence agency. . .

> . . . could produce a deepfake video of a politician using a racial epithet or taking a bribe, a presidential candidate confessing complicity in a crime, or warning another country of an upcoming war, a government official in a seemingly compromising situation, or admitting a secret plan to carry out a conspiracy, or US soldiers committing war crimes such as killing civilians overseas (2019).

A foreign intelligence agency (or other non-affiliated actors) could thus use deepfake technology to commit algorithmic blackmail – where victims are given a choice of either to pay to stop the deepfake, give into specific demands, or suffer the public consequences (Kietzmann et al., 2020). Adversarial actors can employ such algorithmic blackmail to target high profile individuals, such as people with power, influence, or popularity. Not only that, deepfakes can also have the potential to be used to target anyone for online harassment, defamation, revenge porn, bullying, and identity theft.

Implications

Bots and deepfakes pose a major threat to our social, political, and economic systems through their ability to spread disinformation with speed and efficiency. In the foreseeable future, deepfake technology will “be irresistible for nation-states to use in disinformation campaigns to manipulate public opinion, deceive populations, and undermine confidence in […] institutions” (Riechmann, 2018). With bot technology, even if there is no malicious intent behind bots, bots can still unintentionally spread misinformation by not verifying credibility of information before spreading it through social media (Gupta et al., 2013). Taken as a whole, the potential for bots and deepfakes to enable disinformation and misinformation to be planted gradually in the minds of a population threatens the health for all democratic governance. Mika Westerlund lists reasons why in her paper, *The Emergence of Deepfake Technology: A Review* (2019), deepfakes [and bots]:

> Put pressure on journalists struggling to filter real from fake news,
>
> - threaten national security by disseminating propaganda and interfering in elections,
> - hamper citizen trust toward information by authorities, and,
> - raise cybersecurity issues for people and organizations.

Deepfake and botnet technologies powered by AI will be highly valuable to non-state actors, who now have additional resources plus methods to create and propagate fraudulent yet seemingly credible media content. These technologies can be found both commercially and even free on the internet, reducing their barriers to entry, and making them accessible to use. Non-state actors can use these tools to target and depict their adversaries (like the US) using inflammatory words or engaging in offensive actions to demonize and weaken support around them. They can also be used by non-state actors to intensify the disinformation wars that are increasingly disrupting domestic politics in the US and elsewhere.

Increased usage of integrated technology and the resulting increased exposure to deepfake and botnet technology that may come from it, will result in citizens’ digital literacy and trust toward government provided information to become hampered. This makes it especially lucrative to state and non-state actors who wish to plant distrust toward a certain policy proposal or politician. In 2019, the US Democratic Party deepfaked Tom Perez, its chairman, to call attention to the threat that deepfakes could have during election cycles (Westerlund). Disinformation campaigns can now be efficiently deployed using deepfake and botnet technology and the pervasive reach of integrated technology. Political actors can engage in *political microtargeting techniques* (PMT), which is a relatively new technique used by political campaigns (Dobber et al. 2021). PMT is “a type of personalized communication that involves collecting information about people and using that information to show them targeted political advertisements” (Borgesius et al. 2018). Deepfakes can also be used in conjunction with PMT by doctoring visual and auditory content like political advertisements to target and influence subgroups of the electorate. Botnet technology is used by PMT to collect information from these subgroups and then can push deepfake content to them.

Solutions to Consider

To prevent a stronghold of disinformation from taking root into reality, it is of paramount importance that methods to combat deepfakes and bot technology are developed so that society does not fall into a trap of no longer being able to discern what is real and what is fake. The rise of these emerging technologies will benefit both non-state and state actors alike, who will try to grasp these new tools of power to weaken societal trust in important structures, institutions, and figures. It is important that some effort is done in ensuring a healthy environment of the merger between the digital and analog life-worlds.

When it comes to how to tackle deepfake risks, Keitzmann et al. in their paper on Deepfakes propose a R. E. A. L. Framework:

- Record original content to assure deniability
- Expose Deepfakes early
- Advocate for legal protection
- Leverage trust (2020)

This framework may be able to treat some symptoms of the malicious use of deepfake technology being used for nefarious intents. If one were to fall victim to a deepfake effort, the hope is that this framework may be enough to limit the extent of the deepfake's damage. However, it is not foolproof. Sometimes original content may be lost or destroyed, exposure of a deepfake may take longer than needed, legal protection may not be affordable or suitable enough, and trust may already be hampered to begin with. The result is that the damage done from exposure to a disinformation campaign using deepfakes and bots may be irrecoverable. As the threat of this technology is still relatively new, this framework is still something to work with and tweak.

Social media platform owners also employ machine learning tools to counter malicious and disingenuous content. Current implementations of such detection models have found some success in identifying and removing harmful content but with limited results. These detection models in use by platform owners often fall short of their role as the training datasets used for the algorithms to learn are immediately outdated the moment they are deployed. As a result, preventing harmful deepfake content and disinformation from spreading is a never-ending chasing game with bots.

Another possible suggestion to encourage tech firms to better handle malicious use of disingenuous content (manipulated by AI) on their platforms is to threaten regulation. The threat of regulation on tech firms may be enough to encourage some of these firms to put forth effort in their handling of misinformation campaigns from further spreading. As creators of said platform and tools, they have the most control on the spread of the malicious use of AI and machine learning. Such regulation would prompt platform owners to monitor and maintain their platforms continuously from the disingenuous use of bots and deepfakes. Already, the threat of some form of regulatory oversight has prompted Facebook to implement an independent oversight committee over its platform to appease legislators from enacting upon its legislative punishment. Further scrutinization may prompt the brainstorming of even more possible solutions. Additionally, regulation could also be put into place that explores the ethical ramifications of using someone's likeness without consent. Enforcing a "Code of Ethics" could also be a steppingstone into slowing the spread of disinformation or manipulated content. Regulation requiring transparency will also be important.

Conclusion

AI powered technologies like deepfakes and bots are constantly improving. The spread and introduction of new forms of integrated technology can widen their reach pervasively. They have enormous potential to be used maliciously and are incredibly valuable to both state and non-state actors alike. Their abilities to target the minds' cognitive abilities, trick audiences, and encourage false perceptions allow for them to be utilized in attacks targeting social, political, and economic figures and institutions. The introduction of these technologies is shaping a new reality that we must face and tackle head on. We are now living in an era in which these technological applications can be used to falsely portray or accuse prominent moments or individuals and spread disinformation. There are methods to detect deepfakes, but work is still needed to combat their use. Proper regulation, increased transparency, and a code of ethics are just some suggestions of what can be done to deter the illicit intended negative effects they have on society.

References

Abokhodair, N., Yoo, D., & McDonald, D. W. (2015). Dissecting a social bot- net: Growth, content and influence in Twitter. *Paper presented at the Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW),* 839–851.

Berger, J. M., & Morgan, J. (2015). The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter (The Brookings Project on US Relations with the Islamic World, Analysis Paper No. 20). Washington, DC: The Brookings Institution.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, *47*(2), 241–262.

Borgesius, F., Zuiderveen, J., Möller, J. Kruikemeier, S., Fathaigh, R., Irion, K., Dobber, T., Bodo, B., & Vreese, C. (2018). Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review, 14* (1): 82–96.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2012). Key challenges in defending against malicious socialbots. *University of British Columbia.*

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2012). Design and analysis of a social botnet. *University of British Columbia*.

Chen, Z., & Subramanian, D. (2018). An unsupervised approach to detect spam campaigns that use botnets on Twitter.

Chesney, R. & Citron, D. (2019). Deepfakes and the New Disinformation War.

Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect*. International Journal of Computer Applications.*

Dick, S. (2019). Artificial Intelligence. *Harvard Data Science Review*, *1*(1). DOI: 10.1162/99608f92.92fe150c

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & Vreese, C. (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?. *The International Journal of Press/Politics.* DOI: 10.1177/1940161220944364

Floridi, L. (2016). The 4th Revolution how the info sphere is reshaping human reality*. Oxford: Oxford University Press.*

Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2013). False memories of fabricated political events. *Journal of Experimental Social Psychology, 49*(2), 280–286.

Gupta, A., Lamba, H., & Kumaraguru, P. (2013). $1.00 per rt# bostonmara- thon#prayforboston: Analyzing fake content on Twitter [Paper presentation]. eCrime Researchers Summit (eCRS), San Francisco, CA, 1–12.

Kietzmann, J., Lee, L., McCarthy, I., & Kietzmann, T. (2019). *Deepfakes: Trick or treat?*. Kelley School of Business, Indiana University.

Kluver, Z., Cooley, S., Hinck, R., & Cooley, A. (2020). Propaganda: Indexing and Framing and the Tools of Disinformation. *The Media Ecology and Strategic Analysis Group.*

Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: exposing ai created fake videos by detecting eye blinking. 2018 IEEE International Workshop on Information Forensics and Security (WIFS).

Li, Y., & Lyu, S. (2018). Exposing Deepfake Videos by Detecting Face Warping Artifacts.

McCarthy, J., Minksy, M., Shannon, C. E., Rochester, N., & Dartmouth College. (1955). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine, 27*(4), 12-14.

Metaxas, P. T. (2020). Technology, propaganda, and the limits of the human intellect. In M. Zimdars & K.McLeod (Eds.), Fake News: Understanding Media and Misinformation in the Digital Age, (pp.245-252). The MIT Press.

Nguyen, X., Tran, T., Le, V., Nguyen, K., Truong, D. (2021). Learning Spatio-temporal features to detect manipulated facial videos created by the Deepfake techniques. *Elsevier.* DOI: https://doi.org/10.1016/j.fsidi.2021.301108

Perrin, A. (2015). Social Networking Usage: 2005-2015. *Pew Research Center*.

Riechmann, D. (2018). *I never said that! High-tech deception of 'deepfake' videos*. AP News. https://apnews.com/article/21fa207a1254401197fd1e0d7ecd14cb

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, *39*, 127–161. DOI: https://doi.org/10.1016/S0065-2601(06)39003-X

Siau, K. & Wang,W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*(2).

Sundar, S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. Metzger & A. Flanagin (Eds.), Digital media, youth, and credibility (pp. 73–100). MIT Press.

Susskind, J. (2019). *Future Politics.* Oxford University Press.

Vaccari, C. & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*. DOI: 10.1177/2056305120903408.

Weber, M. (2010). The Profession and Vocation of Politics. In Peter Lassman and Ronald Spiers (Eds.), *Political Writings* (pp. 356). Cambridge: Cambridge University Press.

Westerlund, M. (2019). The emergence of deepfake technology: a review. *Technology Innovation Management Review.*

Yang, K., Varol, O., Davis, C., Ferrara, E., Flammini, A., & Menczer, F. (2018). Arming the public with artificial intelligence to counter social bots. *Wiley*. DOI: 10.1002/hbe2.115