Stanford University
Human-Centered
Artificial Intelligence

NOVEMBER
2020

Preparing for the Age of Deepfakes and Disinformation

Dan Boneh, Andrew J. Grotto,
Patrick McDaniel and Nicolas Papernot

POPULAR CULTURE HAS ENVISIONED SOCIETIES of intelligent machines for generations, with Alan Turing notably foreseeing the need for a test to distinguish machines from humans in 1950. Now, advances in artificial intelligence that promise to make creating convincing fake multimedia content like video, images, or audio relatively easy for many. Unfortunately, this will include sophisticated bots with supercharged self-improvement abilities that are capable of generating more dynamic fakes than anything seen before.

In our paper "How Relevant is the Turing Test in the Age of Sophisbots," we argue that society is on the brink of an AI-driven technology that can simulate many of the most important hallmarks of human behavior. As the variety and scale of these so called "deepfakes" expands, they will likely be able to simulate human behavior so effectively and they will operate in such a dynamic manner that they will increasingly pass Turing's test.

The issue for policymakers is how to identify the right tools to reveal the use of such generative technology and how to develop the right regulatory framework to mitigate their negative impact. Regulators should be conversant in the latest technical developments but they must also take steps to address the threat of malicious actors by fitting technologies in question into broader regulatory structures, adopting legislative incentives for platforms to responsibly develop these powerful algorithms, and hold malicious actors accountable for harmful behavior.

KEY TAKEAWAYS

- Generative Adversarial Networks (GANs) produce synthetic content by training algorithms against each other. They have beneficial applications in sectors ranging from fashion and entertainment to healthcare and transportation, but they can also produce media capable of fooling the best digital forensic tools.
- We argue that creators of fake content are likely to maintain the upper hand over those investigating it, so new policy interventions will be needed to distinguish real human behavior from malicious synthetic content.
- Policymakers need to think comprehensively about the actors involved and establish robust norms, regulations, and laws to meet the challenge of deepfakes and AI-enhanced disinformation.

Introduction

The rapid development of Generative Adversarial Networks, sometimes referred to as GANs, has opened up a broad array of applications in fields as diverse as medical imaging, the arts, information technology, media, and even the design dental fixtures just to name a few. The underlying technology essentially teaches itself how to create progressively more and more believable content by taking two separate algorithms and training them against one another to improve their performance in tandem. The first algorithm, known as the generator, learns how to produce synthetic content from a set of training data while the second algorithm, known as the discriminator, learns how to determine if a given piece of content produced by the generator is real or fake.

If successfully implemented, this adversarial design paradigm can be applied to a vast array of use cases nearly indistinguishable from the reality of human communication. While responsible use of these programs shows they can provide clear benefits to society, malicious actors have increasingly shown a willingness to use them to expand the scale of disinformation campaigns and use their self-improving nature to make content forgery more difficult to spot.

At a time when faith in traditional institutions like the news media and the government are at a nadir, this new threat deserves critical attention. In the wake of repeated leaks of weaponized private information and efforts to undermine confidence in democratic political systems, the relatively open source nature of adversarial algorithm design makes creating these programs easy to do, so these systems will only proliferate more as time goes on.

Implications and Defenses

Forging an effective response to the threat of synthetic content falls into two broad categories. The first group of technical defenses ranges from digital forensics tools to better detect synthetic content (like DARPA's MediFor program and Eulerian Video Magnification), to ensuring the authenticity of digital content (potentially through blockchain-based verification systems), to even more extreme measures like voluntary surveillance systems (designed to ensure accountability for an individual's whereabouts for all twenty-four hours every day). The second broad group of defenses lies in closer regulatory scrutiny on the wide array of stakeholders responsible for the design and management of content creation systems.

Technological innovation is unlikely to be able to address the challenges of fake content emulating human behavior by itself.

Techniques for manipulating videos often introduce specific imperfections that can be detected. Digital forensic tools like DARPA's MediFor provide automated assessments of an image or video's integrity while

providing detailed information about how these manipulations were performed. Taking careful note of physiological inconsistencies such as irregular eye blinking patterns, they are also building a centralized repository of markers of content known to be synthetic. But since GANs are trained to evade detection by design, a major issue is that servers hosting these types of media likely cannot rely on detectors to successfully identify fake videos in the long term.

The Federal Trade Commission could hold the platforms accountable using its unfair trade practices authority.

A second technical approach involves building a secure record of all entities and systems that manipulate a particular piece of content. The goal of such a "data provenance" approach would be to identify deepfakes as content that was digitally synthesized instead of being captured using a camera. One proposed solution in this vein is to equip every digital camera with a tamper proof cryptographic content signing key. Such a system could ensure every piece of media is accompanied by a digital signature, potentially based on blockchain technology. However key creation, distribution, and authentication might make implementing this logistically difficult in practice.

The third technological defense frequently touted is "total accountability," in which individuals record every minute of their lives on a tamper proof camera that signs and timestamps all of its captured video. While such a system would be useful for a relatively small sample of the population who are likely to be the target of sustained disinformation campaigns, the potential loss of privacy from this kind of 24/7 self-surveillance may cause more harm than good. Taken in the aggregate, technological innovation by itself is unlikely to be able to address the challenges of fake content emulating human behavior.

Jaw correction as an example of steps GANs can take to seamlessly blend synthetic content together.

Policy Discussion

From the perspective of technologists engaged in the details of developing GANs in a responsible manner, significant policy interventions will be required to ensure the proliferation of the technology does not lead to negative outcomes. Regulators need to reorient their focus away from the immediate output of individual programs and towards the constellation of different actors with a stake in the technology from the designers of the applications used to create fake content and the authors of fake content, to the owners of the platforms

How we distinguish reality from the synthetic in our evolving world of thinking machines presents one of the most pressing questions of our time.

that host fake content and the manufacturers who create hardware like cameras for capturing content. Identifying all of the equities at play in the realm of synthetic content will allow for the specification of a more precise threat model and enable regulators to more creatively design policy tools aimed at nudging, shaping, or informing their behavior in a holistic manner.

Some potential authors of deepfakes—politicians, for example—could commit to not depicting their rivals in deepfakes. Federal or state governments could establish criminal penalties for the use of GANs in any context and campaign committees could withhold funding from candidates suspected of using them. The authors of software capable of producing deepfakes could be incentivized to include cryptographic signatures to aid detection of deepfakes, perhaps by holding developers who do not include a signature liable for works created using their software. Likewise, app stores could refuse to carry software that lacks this capability and the law could be updated to establish that depicting a third person in a deepfake without their consent constitutes defamation.

It may also be possible for forensics experts to rely on attribution mechanisms to identify models which generated known deepfakes. For instance, one of us is part of a research team that has shown attribution may be possible by inspection of a generative model's random seed space. Conversely, model owners who wish to obtain guarantees of plausible deniability can keep a record of all seeds used with their generative model, for instance in a ledger. This would allow them to provide evidence indicating they have not generated a specific deepfake.

Platforms that host fake content could be required to not only establish a procedure for receiving complaints about deepfakes—as some have already done voluntarily—but to also provide a concise overview of the principles behind such standards. The Federal Trade Commission could then hold platforms accountable using its unfair trade practices authority. Platforms could also label content known or suspected to be machine generated, and the educators who train aspiring engineers could elevate policy and ethical literacy as important facets of technical education.

While none of these interventions will likely provide a quick fix to eroding trust in the information ecosystem, they offer a starting point for valuable discussions and provide a critical opportunity to affirm the values we hold most dear. Some considerations will undoubtedly lead to tradeoffs (both foreseen and unforeseen), but user research will be useful in finding best practices on implementation. How we distinguish reality from the synthetic in our evolving world of thinking machines presents one of the most pressing questions of our time.

Policymakers and the technical community are urged to embrace and address these challenges as readily as they're exploring the fascinating and exciting new uses of artificially intelligent systems.

The working paper, **"How Relevant is the Turing Test in the Age of Sophisbots?"** can be found here: https://ieeexplore.ieee.org/abstract/document/8886907

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Dan Boneh heads the applied cryptography group at the Computer Science department at Stanford University.

Andrew J. Grotto leads the Program on Geopolitics, Technology and Governance at Stanford University's Cyber Policy Center.

Patrick McDaniel is the William L. Weiss Professor of Information and Communications Technology and Director of the Institute for Network and Security Research at the Pennsylvania State University.

Nicolas Papernot is an assistant professor in computer engineering and computer science at the University of Toronto and a Canada CIFAR AI Chair at the Vector Institute.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**