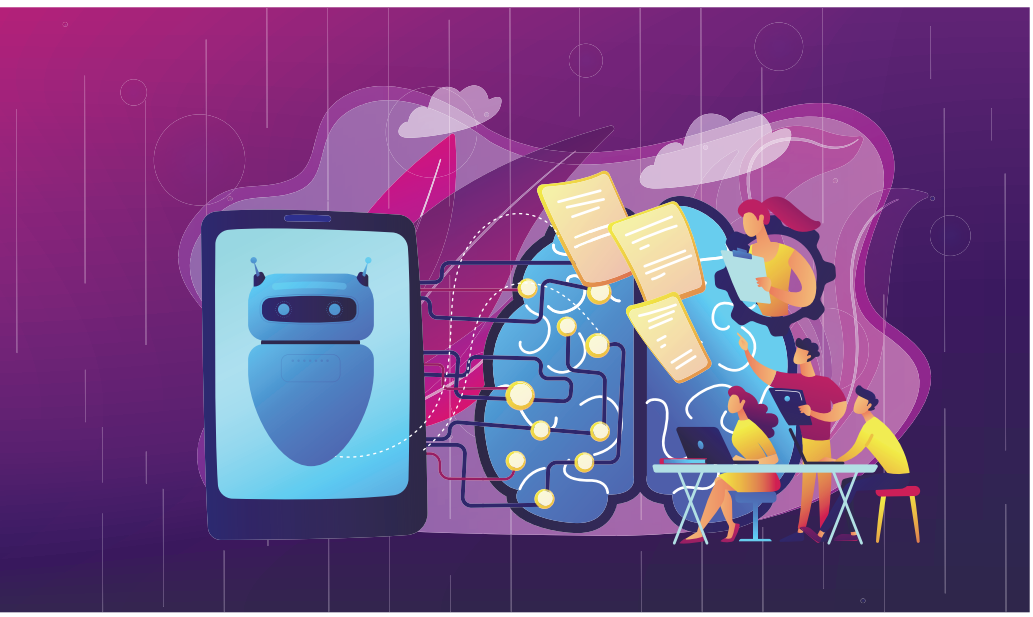
How Relevant Is the Turing Test in the Age of Sophisbots?

Dan Boneh and Andrew J. Grotto | Stanford University
Patrick McDaniel | The Pennsylvania State University
Nicolas Papernot | The University of Toronto

Popular culture has contemplated societies of intelligent machines for generations. Today, we find ourselves at the doorstep of technology that can at least simulate thinking, feeling, and other behaviors. The question is: Now what?



Books, movies, and TV shows have for decades featured a world of cognitive, autonomous machines and imagined future scenarios that ranged from utopian to dystopian. Visions of HAL and R2-D2 have excited our imagination and, at the same time, prompted concerns about how such entities would impact us individually and our society at large. This future is arguably here. We are on the brink of technology that can think, act for itself, and feel, and we face the dilemma of how to handle it.

Was this the future that Alan Turing famously envisioned in 1950 when he created the Turing test to distinguish machines from humans, based on whether a machine could fool a human into thinking it was a person? Maybe and maybe not. Advanced applications of machine learning (ML) and artificial intelligence enable us to engineer programs that are capable of outputting hyper-realistic communication, images, and potentially fully human personas. There are numerous positive uses for that software, but does the technology have darker implications associated with the intent to cause harm?

We take the example of sophisticated online bots that are able to emulate human behavior and interact with us seamlessly. Because they make the distinction between authentic and fake communication harder than it is already, these sophisbots could have a profound impact on society by gradually manipulating our trust in content such as images and videos. We argue that technical solutions, while important, should be complemented by efforts to inform policy and international norms so that they evolve beside technological developments. Indeed, we believe it is essential to foster an increased public understanding of the nature of our interactions with systems deploying ML to ensure that the advances they enable produce a positive impact.

Can Machines Pass the Turing Test Today?

ML can emulate human behavior, thought processes, and strategies to the point of indistinguishability

Digital Object Identifier 10.1109/MSEC.2019.2934193
Date of current version: 29 October 2019

between people and machines in certain contexts. Google's Duplex makes reservations by conversing with humans over the phone. There, learning algorithms capture subtle artifacts of spoken English to replicate human characteristics, such as hesitations and pauses, thereby generating speech that is very conversational and lifelike. In another domain, Christie's announced in 2018 that it was the first auction house to sell art generated by a neural network.[21] Such a development questions the necessity of involving human forms of creativity as a prerequisite to producing enjoyable art.

It is not hyperbole to suggest that those examples—which were, until now, exclusively associated with humans—are only the first of many more to come. ML techniques are no longer limited to responding to inputs created and curated by people, such as requests to place a reservation. They are capable of producing original synthetic content completely from scratch by using a class of generative algorithms. This area of research has, in many ways, been revolutionized by a particular technique, Generative Adversarial Networks (GANs).[8] They enable many creative applications that are beneficial to society, ranging from automatically designing dental crowns that fit individual patients to augmenting the creativity of artists by synthesizing musical notes.

The GAN framework involves two ML models trained in a competing fashion. The first, called the generator, learns how to produce synthetic inputs, for example, images and videos. The second, referred to as the discriminator, learns how to classify inputs, such as data from the generator, as real or fake. The process may remind readers of the Turing test. Indeed, the discriminator is orchestrating some form of weaker but automated imitation game.[19] Assuming that neither model wins the game early in the training process, the discriminator gets better at distinguishing synthetic and real inputs, while the generator gradually improves at producing synthetic inputs that the discriminator finds harder to detect. We revisit this relationship between the GAN framework and the imitation game later since it has implications for the (lack of) existing automated techniques for detecting machine-generated content.

It is exciting that scientific progress enables us to get closer to building machines that may pass the Turing test. The ability to generate synthetic media, including videos that appear natural to humans, has many positive applications; for instance, *Star Wars: Episode IX* is set to include the late actress Carrie Fisher. However, as with any technology, scientific progress made by researchers interested in generative ML could enable potentially less honorable applications by individuals with malicious intentions to bully, blackmail, extort, defame, and mislead.

Take the example of research conducted by Suwajanakorn et al. in which the team applied another type of generative ML model, a recurrent neural network, to synthesize the face and voice of Barack Obama.[17] The researchers were able to produce realistic video footage of the former U.S. president giving an address from a text transcript of their choice. Although a careful human observer would notice some minor inconsistencies in the videos, it is clear that such a technique could be leveraged by malicious individuals to increase the sophistication of the fake news spread by social media bots.

That was demonstrated, most likely without malicious intentions, when a Belgian party created a May 2018 video of U.S. President Donald Trump urging Belgium to follow the United States in withdrawing from the Paris climate agreement.[22] By circulating the video on social networks, the party's supposed intent was to generate attention to and a debate of the issue of climate change. While the motive likely was not to deceive, releasing the video contributed to the public's eroding confidence in the media. In another example, Yao et al. showed that ML is capable of generating fake Yelp reviews that people not only confused with real reviews but found just as useful.[20] In some ways, this technology is close to passing a Turing test, albeit a passive one where the human consumes the machine's outputs but is unable to query it with specific inputs.

Implications of ML Emulating Human Behavior

All of these discoveries and observations lead to a single conclusion: we are rapidly reaching a point where computational algorithms can create nearly any form of human communication that is, for all intents and purposes, indistinguishable from reality. Apart from the many obvious positive uses, such as telepresence and human-computer interaction, what does this mean to us as a society, and what does it do to online (and offline) public discourse, politics, due process, commerce, and society at large?

In the hands of individuals with malicious intent but low technical sophistication, the issue is first one of forgery. There is a history of written words and reported discourse being faked. Counterfeit letters claimed, in 1777, that George Washington thought the Revolutionary War against Great Britain was a mistake.[23] After decades of improvements to the Adobe software, the expression "to Photoshop" became synonymous with

digitally altering images. In short, the barrier to entry for manipulating content has been lowering for centuries. Progress in ML is simply accelerating the process. Items that were less forgeable in the past, such as voice and video when compared to written words and photos, can be counterfeited now or soon will be.

We already have a substantial problem with forged and out-of-context content: the Internet is plagued by fake reviews, posts, and people fueling misinformation. We now know that carefully crafted social media messages were created to disrupt public discourse and influence opinion prior to the 2016 U.S. presidential election. Suppose that adversaries could automate the process and create such content algorithmically, which is a hypothesis that was reinforced by the controversial way that researchers presented their recent results[16] on natural language processing to the public. How does that capability change the Internet and affect our society?

As the sophistication of our adversaries increases, consider the hypothetical but logical conclusion of content forgery: a fake human that we will refer to as a *sophisbot*. Here, the sophisbot is a program running in the ether of the social media and other infrastructure. It never sleeps or ages and is not bound by geography, culture, and conscience. It can have opinions, agendas, and biases and consume enormous amounts of information while maintaining a nearly infinite number of simultaneous conversations. Sophisbots can have real visages and personalities that will draw people to them.

Take the same bot and give it a task. It can be political, such as promoting one candidate over others, sowing mistrust among the citizens, spreading racist and misogynistic propaganda, and so forth. Other tasks can be more on the human scale, such as harassing an at-risk teenager into suicide, ruining personal relationships, and provoking a victim into doing something financially or personally risky. Perhaps less damaging but still ethically questionable, the bots can simply promote products or services. Sophisbots would tirelessly use all the tools of forgery and social engineering to achieve their goals. Older technologies transformed communication in similar ways: email reduced the resources needed to craft sophisticated junk mail that could reach millions of individuals. Consider a scenario where you could use ML to create billions of sophisbots simulating fake humans, in a matter of seconds.

Such bots are only some distance from those we see today. Twitter bots already generate fake but widely read content and retweet it. A 2017 study by the Pew Research Institute showed that more than half of the links on Twitter were posted by automated accounts.[24] That reality was also observed in earlier academic studies[4] and during the Defense Advanced Research Projects Agency (DARPA) Twitter bot challenge.

In short, while sophisbots would not introduce a new problem, they would reinforce the existing issue of content forgery by potentially providing scalability to the existing techniques for manipulating content. Adversarial parties have already exploited online services that were driven by automated reasoning for such malicious ends. In one well-publicized example, Microsoft's Tay was a chatbot that learned to speak (in part) from user queries. Malicious people on the Internet quickly learned of the service and trained it to post racist and inflammatory tweets.[11] The service was shut down within 24 h of its launch. The question is, how do we identify and possibly eliminate malicious bots and content from public discourse?

Do Technological Defenses Exist?

The central issue concerns how we defend against this form of malicious activity. There are really two answers to the question: science and policy. We first discuss the former.

We argue that, as a matter of science, ML must evolve to make its systems and models accountable and its inputs and outputs reliably identifiable.[1] Digital forensic techniques are being developed to detect manipulated content. For instance, specialists commonly use the lack of camera-induced imperfections in synthetic images and videos to identify fakes. Unfortunately, the approach does not scale given the shortage of human experts, unless their work can largely be automated by technological solutions that identify content produced and edited by machines. We explore three potential approaches but stress that, despite the partial progress they afford, they are quite limited. Our analysis suggests the likelihood that there is no robust technological defense against this problem.

Detecting Artifacts of Synthetic Content

A natural first approach is to automate the process of digital forensics and attempt to identify machine-manipulated content by detecting its imperfections. For instance, techniques for manipulating videos often introduce specific imperfections that can be distinguished. Generators that rely on deep learning to produce phony videos that are known as deepfakes often operate by face swapping; body movements and proportions are typically unchanged from the stand-in actors. Techniques such as Eulerian video magnification could help to identify human pulses in videos.[5] In principle, a detector could spot a deepfake by perceiving those and other imperfections.

DARPA operates a media forensics program called MediFor that funds research to develop robust detectors that operate in such a way. Engineers are developing tools to identify deepfakes through physiological inconsistencies, such as detecting irregular eye-blinking patterns or the lack of blinking altogether.[9] To facilitate learning detectors for synthetic content, researchers have collected datasets of content known to be the output of generative models.[15]

While it may be possible to design an effective detector against the current generation of deepfake generators, the effort to combat counterfeit content is likely to be a losing battle or, at best, a stalemate.[6] Indeed, ML detection is more likely to result in an arms race than traditional system and network anomaly detection was. That is because ML algorithms developed to create synthetic content, such as GANs, involve by design a generator trained to evade detection, meaning that as the detector gets better at distinguishing synthetic from natural content the generator that creates synthetic content improves, too.

Every time a new detector is deployed, the generator can be retrained to evade it. Improving the generator is efficient as long as enough information about the discriminator can be quickly gathered. At every iteration the generator gets better, as does the detector. The process may never produce a setting that steadily favors the defender as long as retraining the generator continues to be less costly than coming up with an improved discriminator. For example, a detector that used blinking to identify deepfakes[9] could be effective during the short term, but eventually it would likely drive the development of generators that correctly emulated human eye blinking. In the end, sites hosting videos would not be able to rely on the blinking detector indefinitely, at least not solely by analyzing the video itself.

To summarize, progress in generative model research will probably continue to give an edge to those creating fake content during the long term. In the short term, that scenario would very much resemble the status quo in signature-based malware detection where defenders are constantly defining signatures for new forms of malware.

Content Provenance

The second approach seeks to improve the provenance of the human forms of digital communication. By provenance we mean building a secure record of all entities and systems that manipulate a particular piece of content.[10] Consider our deepfake example. The goal of a data provenance approach would be to identify a deepfake as content that was digitally synthesized instead of captured by a camera. A fairly obvious solution would be to equip every digital camera with a tamper-proof cryptographic content-signing key. The camera would use the key to sign all of the video clips that it exported. That way, every clip would be accompanied by a digital signature that identified the camera on which it was shot. Such functionality is available through applications like the Guardian Project's ProofMode.[25]

Presumably, a deepfake generator would be unable to sign a fake video because it would not have the signing key that was embedded in the hardware of a real camera. However, due to key creation, distribution, authentication, and other issues, achieving that level of security would be logistically difficult. It would be further complicated if the content were postprocessed by users (for example, by cropping and applying filters) because the edits would have to preserve the image or video signature.[12] Alternatively, one might be able to defeat the provenance system by using the "analog hole" attack: simply play an unsigned deepfake video on a screen, and record the screen using an approved camera that properly signed videos. In the absence of other identifying factors, such as the physical location of the signature in verified content, it would most likely be difficult to detect such an attack through artifacts added by the screen as discussed previously.

Total Accountability

The third defense involves a regime of total accountability. Consider a fictional public figure, Bernie, who is concerned about fake videos, such as deepfakes, and willing to take extreme measures to protect himself. In principle, Bernie could record every minute of his life on a tamper-proof camera that signed and timestamped all of its captured videos. If a deepfake of Bernie were published, he could prove that at the purported time of the deepfake he was engaged in a slightly different activity. That after-the-fact defense would not mitigate the potential damage to his public image but would enable him to prove that the deepfake was a forgery. Of course, the cure could be worse than the disease: the potential loss of privacy from that 24/7 self-surveillance may cause more harm than the concern about deepfakes. In truth, we are fairly certain that a different instantiation of total accountability is required to avoid having every person create his or her own version of *The Truman Show*.

Such an approach is, in fact, being explored in industry: a product called Amber Authenticate proposes to have cameras periodically compute video signatures and record them publicly on a blockchain.[26] It would not require sharing the content of the videos. Nevertheless, it would be possible for anyone to access the hashes recorded on the blockchain and verify that the hash

for a segment of video footage corresponded appropriately. That would enable one to authenticate the video as having been recorded by the camera at the date and time claimed by its author.

Why Are Technological Solutions Limited?

It is clear from this discussion that technology alone cannot address the challenges posed by fake content that emulates human behavior through ML. Furthermore, the very involvement of humans highlights the limitations of any solution that addresses the problem purely through technology. Take the example of fake news. Research has shown that humans actively seek to reinforce their opinions.[3] People want to hear what they like to hear. As a consequence, each of us reinforces our bubble of opinions through the selectivity and bias of our online connections. This effect is more prevalent in certain demographics, age being the characteristic observed, in a recent study, to have the most significant effect on sharing fake news with online connections.[7]

Even if individuals attempt to fact check their opinions and break out of their bubbles, unbiased information can be difficult to find[18] and have a limited effect on their misperceptions.[13] This problem is not new, but it is aggravated by the level of personalization afforded by generative ML. Combined with large intentional and unintentional leaks of private data that describe the traits of our personal preferences and personalities, those techniques are instrumental in generating media that can effectively manipulate populations.

Does this mean we need to give up some forms of anonymity? Social networks have started to provide mechanisms to distinguish anonymous users from those linked to real-world individuals. For instance, Twitter adds a blue star next to users whose identity has been verified. The verification is optional but recommended by Twitter for populations that are often targeted by bots, such as journalists and celebrities. Obviously, it will not solve the problem of opinion bubbles, but it could begin to address the crisis of content authenticity that will result from Internet users' increasingly high exposure to material generated by machines. It is not a silver bullet. The loss of anonymity through approaches that improve content provenance may have unintended negative consequences, for example, by enabling the tracking and identification of dissidents, minorities, and other vulnerable groups that might face repercussions for the content they create.

The key is to develop public policy, legal, and normative frameworks to refine technology and manage its malicious applications.

The key is to develop public policy, legal, and normative frameworks to refine technology and manage its malicious applications. Law and policy typically lag innovation because the implications of new technologies and how to address them can take time to come into focus and/or emerge as politically salient enough for decision makers to pay attention to them. Asimov summarized it well: "Science gathers knowledge faster than society gathers wisdom." For the most part, the lag is a feature of a vibrant innovation ecosystem because it enables experimentation, risk taking, and freer exchanges of ideas and capital. It can act as a bug, however, when innovation results in rapid paradigm shifts in the relative symmetry between malevolent uses of technology and the efforts to defend against them, as the case was with information technology and potentially is, now, for ML. New technologies are not always "penetration tested" from a policy perspective, since the forces behind innovation often focus on positive applications and are not always incentivized to think proactively about malicious applications, which are easily cast aside as "somebody else's problem, not mine."

Today, we see ML creating numerous policy challenges, some of which mirror many past (and current) concerns and others that are more novel and may create opportunities. For example, what recourse does a victim have when a fake video is created of him or her? Individual companies have complementary processes in place to help users report and remove content from their platforms, but how would those procedures be applied to content created via ML? What rights or options, if any, do individuals have, as a practical matter, for protecting their content from being consumed for the purposes of future faked photos and videos? What is the appropriate response to governments that deploy the technology to interfere with liberal democratic institutions?

Even the basic definition of content ownership becomes murky in this new reality: Are ML data and algorithms the property of the individuals who designed and built the machines or a collective holding

of the people who contributed the information that the software learned from? Answers to those questions and many, many more are going to shape our society and future. We simply cannot wait to see the harm emerge before dealing with it.

We argue for thinking comprehensively about the tool kit for dealing with these and other potential harms and for disaggregating a given problem into as many smaller pieces as possible. For example, in the case of manipulated videos such as deepfakes, one approach to disaggregating the problem is to organize our thinking around the different actors that have a stake in the matter and identifying policy tools aimed at nudging, shaping, and informing their behavior.

From this perspective, there are a number of different parties whose behavior and decisions are relevant to fake content. They include (among others) the authors of fake content; authors of applications used to create fake content; owners of platforms that host fake-content software; educators who train engineers in sensitive technologies; manufacturers and authors that create platforms and applications for capturing content (for example, cameras); owners of data repositories used to train generators; unwitting people depicted in fake content, such as images and deepfakes; platforms that host and/or distribute fake content; audiences who encounter fake content; journalists who report on fake content; and so on.

Breaking down the problem in that way enables us to think more creatively about the range of policy tools that are relevant to the task at hand. It puts us in a stronger position to identify the right policy tool(s) for the job of shaping the behavior of a given actor and, if and when necessary, develop new tools. As is the case for research on the security of computer systems, creating a precise threat model that captures the intentions and capabilities of the parties that are relevant to the system being analyzed is the first step toward building principled defenses. Ultimately, that approach has the potential to evolve into a more comprehensive strategy that aligns incentives across different actors. No single tool may prove to be decisive, but a comprehensive approach that draws on multiple methods that affect different parties could materially move the needle.

For instance, certain groups of would-be authors of deepfakes—politicians, for example—could commit to not depict their rivals. In a democracy such as the United States, we submit that many (and perhaps even most) politicians would likely find a norm along those lines to be attractive through a mutually assured destruction sort of logic: all things being equal, most democratic politicians would prefer to operate and campaign in a world where they and their opponents do not resort to outright fabrications as opposed to one where such behavior is accepted.

The U.S. Congress or a state legislature could endorse a similar standard in a joint resolution, and the legislative campaign committees could do so as well; they could even withhold funding for candidates that violated the standard. Obviously, some politicians and other actors might reject or violate such a system. That would be even more concerning in less democratic societies where totalitarian governments may themselves use ML to enable propaganda at scale. Platforms could help by rejecting fake political content when they discover it or, at least, downgrading it in their promotion algorithms. Each of those measures, on its own, is incomplete, but together they could have impact.

Of course, politics isn't the only domain where fake content could gain traction. Bullies and extortionists will also find uses for the technology. Here, as well, an actor-centric approach yields many possibilities. For example, legislatures (and courts) could clarify that depicting a third person in a deepfake without consent is defamation; victims would then have a cause of action for recovering monetary damages from authors. Legislatures could establish criminal penalties such as the legislation pending before the California legislature. Some malicious authors will hide their identities or may not have deep pockets, so holding them liable is only a partial solution.

Technical measures may be useful in this context despite their limitations. Indeed, the authors of software capable of producing deepfakes could be incentivized to include cryptographic signatures to aid detection of counterfeits, perhaps by holding developers who do not include a signature liable for work created using their products. App stores and other fora for acquiring software could refuse to carry programs that lacked the detection capability, and they could be motivated to do so through legal mandates and civil liability. Obviously, software that lacked such a capability would still be available elsewhere, but these and other barriers could deter casual users while limiting the options available to power users that have malicious intentions.

Meanwhile, platforms that host content could be required not only to establish a procedure for receiving complaints about deepfakes, as some have done voluntarily for content that violates their terms of service and community standards, but to provide a concise overview of the principles behind those guidelines. The Federal Trade Commission could hold the platforms accountable to those published commitments by using its

unfair-trade-practices authority. Platforms could also label content known or suspected to be machine generated. Obviously, they could not label all machine-generated content for the previously described technical reasons that concern false negatives and positives.[2] A balance will be more difficult to achieve as many photo and video editing tools are likely to start including some amount of ML.

Nevertheless, labels could be useful in situations where the authenticity of the content is of paramount importance to its authors and viewers. User research would be valuable, here, to find an implementation that best ensured effective long-term interaction with the labels and avoided pitfalls such as the implied truth effect on unlabeled content.[14] Educators who train the next generation of engineers could elevate policy and ethical literacy to important facets of technical education. Bolstering digital media literacy, especially for the demographics at the highest risk of being deceived, is also essential. Indeed, research has found that correcting misperceptions through the presentation of factual evidence has a limited effect and can sometimes be counterproductive by strengthening misperceptions.[13]

We present these governance interventions to illustrate how breaking the problem down can yield insights into the possibilities for shaping behavior. None of them is a silver bullet, in the same way that the technical possibilities described previously are not. Some of them also raise other challenges and concerns and indicate difficult tradeoffs across other important values and equities. In addition, determined bad actors will often find ways around them. But for less determined bad actors, the interventions that we describe could prove decisive if put in place jointly. Operating in a governed environment would make it costlier for even the determined bad actors to create and spread malicious content.

As Turing envisioned in 1950, machines are on track to become capable of producing any form of human communication. It is likely that they will eventually simulate human behavior effectively and allow for the creation of sophisbots. Perhaps one of the most pressing technical questions for the first half of this century is how we can distinguish reality from the synthetic in our evolving world of thinking machines. Long into the future, the answers will shape how we as a broader society communicate and live. In that regard, the Turing test is more relevant than ever: Will humans continue to be able to identify sophisbots, albeit using increasingly higher levels of knowledge and logic abstractions, until we are able to create an artificial intelligence? The call is clear. Let us as a technical community commit to embracing and addressing these challenges as readily as we do the fascinating and exciting new uses of intelligent systems. ■

References

1. M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media Soc.*, vol. 20, no. 3, pp. 973–989, Dec. 2016. doi: 10.1177/1461444816676645.
2. K. Clayton et al., "Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media," *Polit. Behav.*, vol. 41, pp. 1–23, Feb. 11, 2019. doi: 10.1007/s11109-019-09533-0.
3. E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, June 2015. doi: 10.1126/science.aaa1160.
4. Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on Twitter: Human, bot, or cyborg?" in *Proc. 26th Annu. Computer Security Applications Conf.*, 2010, pp. 21–30. doi: 10.1145/1920261.1920265.
5. M. Elgharib, M. Hefeeda, F. Durand, and W. T. Freeman, "Video magnification in presence of large motions," in *Proc. 2015 IEEE Conf. Computer Vision Pattern Recognition*, pp. 4119–4127. doi: 10.1109/CVPR.2015.7299039.
6. E. Gibney, "The scientist who spots fake videos," *Nature*, Oct. 6, 2017. doi: 10.1038/nature.2017.22784.
7. A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Sci. Adv.*, vol. 5, no. 1, Jan. 2019. doi: 10.1126/sciadv.aau4586.
8. I. Goodfellow et al., "Generative adversarial nets," in *Proc. Conf. Advances Neural Information Processing Systems*, 2014, pp. 2672–2680.
9. Y. Li, M. C. Chang, H. Farid, and S. Lyu, In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. 2018. [Online]. Available: arXiv preprint arXiv:1806.02877
10. K. K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. Seltzer, "Provenance-aware storage systems," in *Proc. 2006 USENIX Annu. Technical Conf.*, pp. 43–56.
11. G. Neff and P. Nagy, "Talking to bots: Symbiotic agency and the case of Tay," *Int. J. Commun.*, vol. 10, pp. 4915–4931, Oct. 2016.
12. A. Naveh and E. Tromer, "PhotoProof: Cryptographic image authentication for any set of permissible transformations," in *Proc. 2016 IEEE Symp. Security Privacy (SP)*, pp. 255–271. doi: 10.1109/SP.2016.23.
13. B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Polit. Behav.*, vol. 32, no. 2, pp. 303–330, June 2010. doi: 10.1007/s11109-010-9112-2.

14. G. Pennycook, A. Bear, E. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings," *Manag. Sci.*, Aug. 9, 2019. doi: 10.2139/ssrn.3035384.
15. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, FaceForensics: A large-scale video dataset for forgery detection in human faces. 2018. [Online]. Available: arXiv:1803.09179
16. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019, preprint. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
17. S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph. (TOG)*, vol. 36, no. 4, July 2017. doi: 10.1145/3072959.3073640.
18. F. Tripodi. (2018). Searching for alternative facts: Analyzing scriptural inference in conservative news practices. Data & Society Res. Inst. New York. [Online]. Available: https://datasociety.net/output/searching-for-alternative-facts/
19. A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, Oct. 1950. doi: 10.1093/mind/LIX.236.433.
20. Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *Proc. 2017 ACM Conf. Computer Communications Security*, pp. 1143–1158. doi: 10.1145/3133956.3133990.
21. "Is artificial intelligence set to become art's next medium?" Christie's, Dec. 12, 2018, [Online]. Available: https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx
22. "A faked video of Donald Trump points to a worrying future," *The Economist*, May 24, 2018. [Online]. Available: https://www.economist.com/leaders/2018/05/24/a-faked-video-of-donald-trump-points-to-a-worrying-future
23. G. S. Schneider, "The fake news that haunted George Washington," *The Washington Post*, Apr. 10, 2017. [Online]. Available: https://www.washingtonpost.com/news/retropolis/wp/2017/04/10/the-fake-news-that-haunted-george-washington/?noredirect=on
24. S. Wojcik, S. Messing, A. Smith, L. Rainie, and P. Hitlin, "Bots in the Twittersphere," Pew Research Center, Apr. 9, 2018. [Online]. Available: https://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/
25. "ProofMode: Verified Visuals (A CameraV-inspired micro app)." [Online]. Available: https://github.com/guardianproject/proofmode
26. Amber Authenticate, San Francisco, CA. [Online]. Available: https://ambervideo.co/

Dan Boneh is a professor of computer science at Stanford University, California. His research focuses on applications of cryptography to computer security. Boneh received a Ph.D. in computer science from Princeton University, New Jersey, in 1996. He is a recipient of the Gödel Prize and the Association for Computing Machinery (ACM) Prize in Computing. He is a member of the U.S. National Academy of Engineering and a Fellow of ACM. Contact him at dabo@cs.stanford.edu.

Andrew J. Grotto is a William J. Perry International Fellow at the Cyber Policy Center and a research fellow at the Hoover Institution, both at Stanford University, California. His research interests center on the national security and international economic dimensions of America's global leadership in information technology. Grotto received a J.D. from the University of California at Berkeley in 2003. Contact him at grotto@stanford.edu.

Patrick McDaniel is the William L. Weiss Professor of Information and Communications Technology in the School of Electrical Engineering and Computer Science at The Pennsylvania State University, State College. His research focuses on networking and security in computing environment. McDaniel received a Ph.D. in computer science and engineering from the University of Michigan, Ann Arbor, in 2001. He is the director of the National Science Foundation Frontier Center for Trustworthy Machine Learning. He is a Fellow of the IEEE and Association for Computing Machinery. Contact him at mcdaniel@cse.psu.edu.

Nicolas Papernot is an assistant professor of electrical and computer engineering at the University of Toronto, Canada. His research focuses on the security and privacy of machine learning. He received a Ph.D. in computer science and engineering from The Pennsylvania State University, State College, in 2018. He is a Canadian Institute for Advanced Research Artificial Intelligence chair at the Vector Institute. He is on the Program Committee of IEEE Symposium on Security and Privacy, USENIX Security Symposium, and Association for Computing Machinery Conference on Computer and Communications Security. Contact him at nicolas.papernot@utoronto.ca.