The New York Times

11 THINGS WE'D REALLY LIKE TO KNOW

How Will We Outsmart A.I. Liars?

For better and worse, humans are only improving their ability to deceive themselves with technology.

By Cade Metz

Nov. 19, 2018

During the summer before the 2016 presidential election, John Seymour and Philip Tully, two researchers with ZeroFOX, a security company in Baltimore, unveiled a new kind of Twitter bot. By analyzing patterns of activity on the social network, the bot learned to fool users into clicking on links in tweets that led to potentially hazardous sites.

The bot, called SNAP_R, was an automated "phishing" system, capable of homing in on the whims of specific individuals and coaxing them toward that moment when they would inadvertently download spyware onto their machines. "Archaeologists believe they've found the tomb of Alexander the Great is in the U.S. for the first time: goo.gl/KjdQYT," the bot tweeted at one unsuspecting user.

Even with the odd grammatical misstep, SNAP_R succeeded in eliciting a click as often as 66 percent of the time, on par with human hackers who craft phishing messages by hand.

[Like the Science Times page on Facebook.** | Sign up for the* ***Science Times newsletter.]

The bot was unarmed, merely a proof of concept. But in the wake of the election and the wave of concern over political hacking, fake news and the dark side of social networking, it illustrated why the landscape of fakery will only darken further.

The two researchers built what is called a neural network, a complex mathematical system that can learn tasks by analyzing vast amounts of data.

A neural network can learn to recognize a dog by gleaning patterns from thousands of dog photos. It can learn to identify spoken words by sifting through old tech-support calls.

And, as the two researchers showed, a neural network can learn to write phishing messages by inspecting tweets, Reddit posts, and previous online hacks.

Today, the same mathematical technique is infusing machines with a wide range of humanlike powers, from speech recognition to language translation. In many cases, this new breed of artificial intelligence is also an ideal means of deceiving large numbers of people over the internet. Mass manipulation is about to get a whole lot easier.

“It would be very surprising if things don’t go this way,” said Shahar Avin, a researcher at the Center for the Study of Existential Risk at the University of Cambridge. “All the trends point in that direction.”

Many technology observers have expressed concerns at the rise of A.I. that generates Deepfakes — fake images that look like the real thing. What began as a way of putting anyone’s head onto the shoulders of a porn star has evolved into a tool for seamlessly putting any image or audio into any video.

In April, BuzzFeed and comedian Jordan Peele released a video that put words, including “we need to be more vigilant with what we trust from the internet,” into the mouth of Barack Obama.

The threat will only expand as researchers develop systems that can metabolize and learn from increasingly large collections of data. Neural networks can generate believable sounds as well as images. This is what enables digital assistants such as Apple Siri to sound more human than they did in years past.

Google has built a system called Duplex that can phone a local restaurant, make reservations, and fool the person on the other end of the line into thinking the caller is a real person. The service is expected to reach smartphones before the end of the year.

Experts have long had the power to doctor audio and video. But as these A.I. systems improve, it will become easier and cheaper for anyone to generate items of digital content — images, videos, social interactions — that look and sound like the real thing.

Inspired by the culture of academia, the top A.I. labs and even giant public companies such as Google openly publish their research and, in many cases, their software code.

With these techniques, machines are also learning to read and write. For years, experts questioned whether neural networks could crack the code of natural language. But the tide has shifted in recent months.

Organizations such as Google and OpenAI, an independent lab in San Francisco, have built systems that learn the vagaries of language at the broadest scales — analyzing everything from Wikipedia articles to self-published romance novels — before applying the knowledge to specific tasks. The systems can read a paragraph and answer questions about it. They can judge whether a movie review is positive or negative.

This technology could improve phishing bots such as SNAP_R. Today, most Twitter bots seem like bots, especially when you start replying to them. In the future, they will respond in kind.

The technology also could lead to the creation of voice bots that can carry on a decent conversation — and, no doubt one day, will call and persuade you to divulge your credit-card information.

These new language systems are driven by a new wave of computing power. Google engineers have designed computer chips specifically for training neural networks. Other companies are building similar chips, and as these arrive, they will accelerate A.I. research even further.

Jack Clark, head of policy at OpenAI, can see a not-too-distant future in which governments create machine-learning systems that attempt to radicalize populations in other countries, or force views onto their own people.

“This is a new kind of societal control or propaganda,” he said. “Governments can start to create campaigns that target individuals, but at the same time operate across many people in parallel, with a larger objective.”

Ideally, artificial intelligence could also provide ways of identifying and stopping this kind of mass manipulation. Mark Zuckerberg likes to talk about the possibilities. But for the foreseeable future, we face a machine-learning arms race.

Consider generative adversarial networks, or GANs. These are a pair of neural network systems that can automatically generate convincing images or manipulate existing ones.

They do this by playing a kind of cat-and-mouse game: the first network makes millions of tiny changes to an image — snow gets added to summery street scenes, grizzlies transform into pandas, fake faces look so convincing that viewers mistake them for celebrities — in an effort to fool the second network.

The second network does its best not to be fooled. As the pair battle, the image only gets more convincing — the A.I. trying to detect fakery always loses.

Detecting fake news is even harder. Humans can barely agree on what counts as fake news; how can we expect a machine to do so? And if it could, would we want it to?

Perhaps the only way to stop misinformation is to somehow teach people to view what they see online with extreme distrust. But that may be the hardest fix of them all.

“We can deploy technology that patches our computer systems,” Mr. Avin said. “But we cannot deploy patches to peoples’ heads.”

More things we'd really like to know

Does the Universe Still Need Einstein? Nov. 19, 2018

How Did We Get to Be Human? Nov. 19, 2018

Where's Our Warp Drive to the Stars? Nov. 19, 2018

Will We Survive Climate Change? Nov. 19, 2018

Honestly, Some Questions Are Better Left Unanswered Nov. 19, 2018

Cade Metz is a technology correspondent, covering artificial intelligence, driverless cars, robotics, virtual reality, and other emerging areas. He previously wrote for Wired magazine. @cademetz

A version of this article appears in print on Nov. 20, 2018, on Page D6 of the New York edition with the headline: 9. Can We Outsmart Fake News?