WWW.ECONOMIST.COM - 1 LUGLIO 2017

Fake news: you ain’t seen nothing yet

Generating convincing audio and video of fake events

Earlier this year Françoise Hardy, a French musician, appeared in a YouTube video (see link). She is asked, by a presenter off-screen, why President Donald Trump sent his press secretary, Sean Spicer, to lie about the size of the inauguration crowd. First, Ms Hardy argues. Then she says Mr Spicer “gave alternative facts to that”. It’s all a little odd, not least because Françoise Hardy (pictured), who is now 73, looks only 20, and the voice coming out of her mouth belongs to Kellyanne Conway, an adviser to Mr Trump.

The video, called “Alternative Face v1.1”, is the work of Mario Klingemann, a German artist. It plays audio from an NBC interview with Ms Conway through the mouth of Ms Hardy’s digital ghost. The video is wobbly and pixelated; a competent visual-effects shop could do much better. But Mr Klingemann did not fiddle with editing software to make it. Instead, he took only a few days to create the clip on a desktop computer using a generative adversarial network (GAN), a type of machine-learning algorithm. His computer spat it out automatically after being force fed old music videos of Ms Hardy. It is a recording of something that never happened.

Mr Klingemann’s experiment foreshadows a new battlefield between falsehood and veracity. Faith in written information is under attack in some quarters by the spread of what is loosely known as “fake news”. But images and sound recordings retain for many an inherent trustworthiness. GANs are part of a technological wave that threatens this credibility.

Audio is easier to fake. Normally, computers generate speech by linking lots of short recorded speech fragments to create a sentence. That is how the voice of Siri, Apple’s digital assistant, is generated. But digital voices like this are limited by the range of fragments they have memorised. They only sound truly realistic when speaking a specific batch of phrases.

Generative audio works differently, using neural networks to learn the statistical properties of the audio source in question, then reproducing those properties directly in any context, modelling how speech changes not just second-by-second, but millisecond-by-millisecond. Putting words into the mouth of Mr Trump, say, or of any other public figure, is a matter of feeding recordings of his speeches into the algorithmic hopper and then telling the trained software what you want that person to say. Alphabet’s DeepMind in Britain, Baidu’s Institute of Deep Learning in Silicon Valley and the Montreal Institute for Learning Algorithms (MILA) have all published highly realistic text-to-speech algorithms along these lines in the past year. Currently, these algorithms

require levels of computing power only available to large technology companies, but that will change.

Generating images is harder. GANs were introduced in 2014 by Ian Goodfellow, then a student at MILA under Yoshua Bengio, one of the founding fathers of the machine-learning technique known as deep learning. Mr Goodfellow observed that, although deep learning allowed machines to discriminate marvellously well between different sorts of data (a picture of a cat v one of a dog, say), software that tried to generate pictures of dogs or cats was nothing like as good. It was hard for a computer to work through a large number of training images in a database and then create a meaningful picture from them.

Mr Goodfellow turned to a familiar concept: competition. Instead of asking the software to generate something useful in a vacuum, he gave it another piece of software-an adversary-to push against. The adversary would look at the generated images and judge whether they were “real”, meaning similar to those that already existed in the generative software’s training database. By trying to fool the adversary, the generative software would learn to create images that look real, but are not. The adversarial software, knowing what the real world looked like, provides meaning and boundaries for its generative kin.

Today, GANs can produce small, postage-stamp-sized images of birds from a sentence of instruction. Tell the GAN that “this bird is white with some black on its head and wings, and has a long orange beak”, and it will draw that for you. It is not perfect, but at a glance the machine’s imaginings pass as real.

Although images of birds the size of postage stamps are not going to rattle society, things are moving fast. In the past five years, software powered by similar algorithms has reduced error rates in classifying photos from 25% to just a few percent. Image generation is expected to make similar progress. Mike Tyka, a machine-learning artist at Google, has already generated images of imagined faces with a resolution of 768 pixels a side, more than twice as big as anything previously achieved.

Mr Goodfellow now works for Google Brain, the search giant’s in-house AI research division (he spoke to The Economist while at OpenAI, a non-profit research organisation). When pressed for an estimate, he suggests that the generation of YouTube fakes that are very plausible may be possible within three years. Others think it might take longer. But all agree that it is a question of when, not if. “We think that AI is going to change the kinds of evidence that we can trust,” says Mr Goodfellow.

Yet even as technology drives new forms of artifice, it also offers new ways to combat it. One form of verification is to demand that recordings come with their metadata, which show when, where and

how they were captured. Knowing such things makes it possible to eliminate a photograph as a fake on the basis, for example, of a mismatch with known local conditions at the time. A rather recherché example comes from work done in 2014 by NVIDIA, a chip-making company whose devices power a lot of AI. It used its chips to analyse photos from the Apollo 11 Moon landing. By simulating the way light rays bounce around, NVIDIA showed that the odd-looking lighting of Buzz Aldrin's space suit—taken by some nitwits as evidence of fakery—really is reflected lunar sunlight and not the lights of a Hollywood film rig.

Amnesty International is already grappling with some of these issues. Its Citizen Evidence Lab verifies videos and images of alleged human-rights abuses. It uses Google Earth to examine background landscapes and to test whether a video or image was captured when and where it claims. It uses Wolfram Alpha, a search engine, to cross-reference historical weather conditions against those claimed in the video. Amnesty's work mostly catches old videos that are being labelled as a new atrocity, but it will have to watch out for generated video, too. Cryptography could also help to verify that content has come from a trusted organisation. Media could be signed with a unique key that only the signing organisation—or the originating device—possesses.

Some have always understood the fragility of recorded media as evidence. "Despite the presumption of veracity that gives all photographs authority, interest, seductiveness, the work that photographers do is no generic exception to the usually shady commerce between art and truth," Susan Sontag wrote in "On Photography". Generated media go much further, however. They bypass the tedious business of pointing cameras and microphones at the real world altogether.

This article appeared in the Science and technology section of the print edition under the headline "Creation stories"