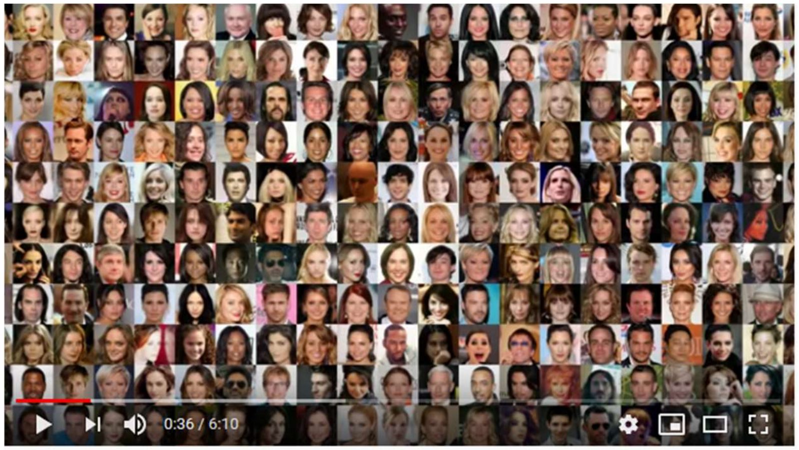
DEEP LEARNING

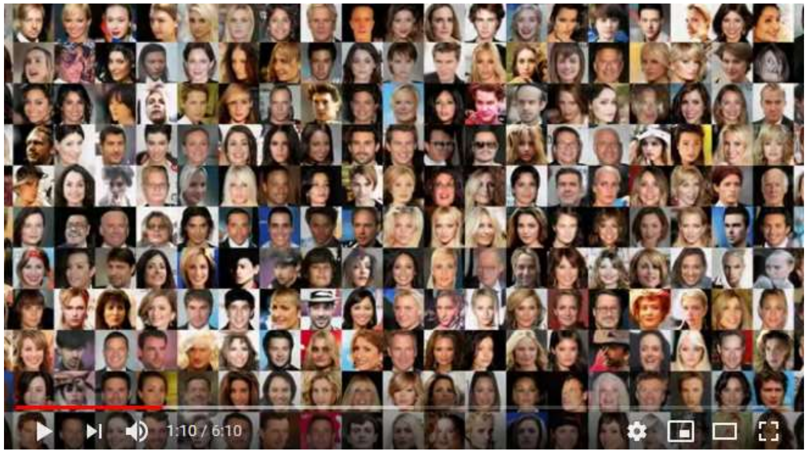
Generative Adversarial Networks

THE DEEP FAKES INDUSTRY

WE KNOW HOW TO GENERATE VERY REALISTIC FAKES

Deep fakes – Images





https://www.youtube.com/watch?v=XOxxPcy5Gr4&feature=youtu.be

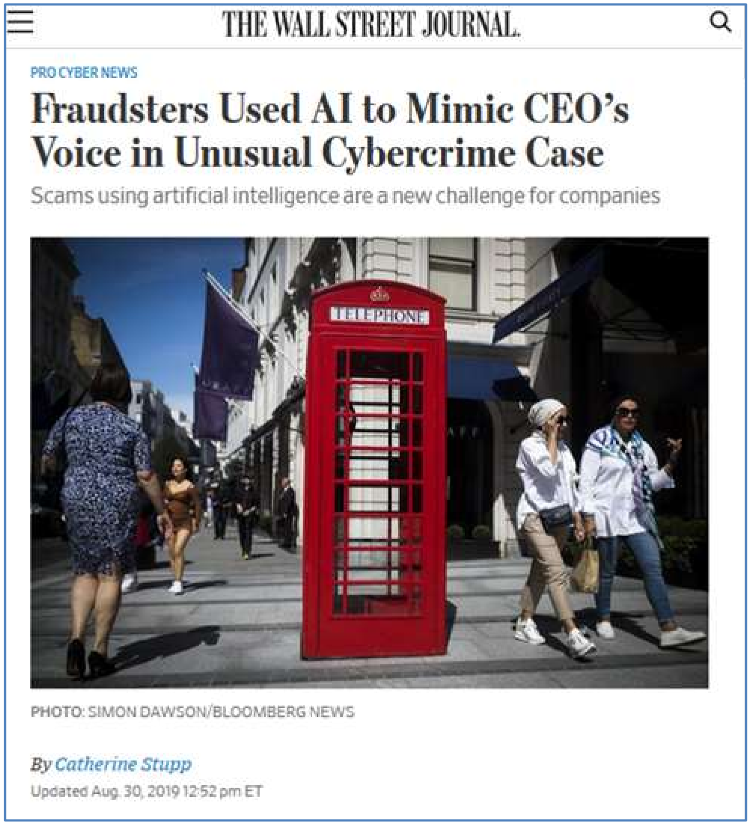
(Real : left. Fake : right)

https://www.lefigaro.fr/secteur/high-tech/2019/02/15/32001-20190215ARTFIG00218-une-intelligence-artificielle-cree-des-visages-ultrarealistes-de-personnes-qui-n-existent-pas.php

https://www.thispersondoesnotexist.com/

Fakes

Deep fakes – Audio



THE WALL STREET JOURNAL.

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

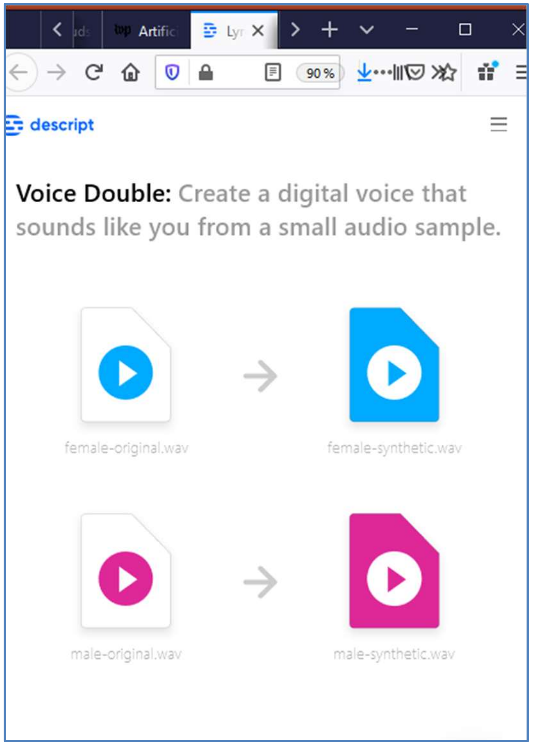
Scams using artificial intelligence are a new challenge for companies

PHOTO: SIMON DAWSON/BLOOMBERG NEWS

By *Catherine Stupp*

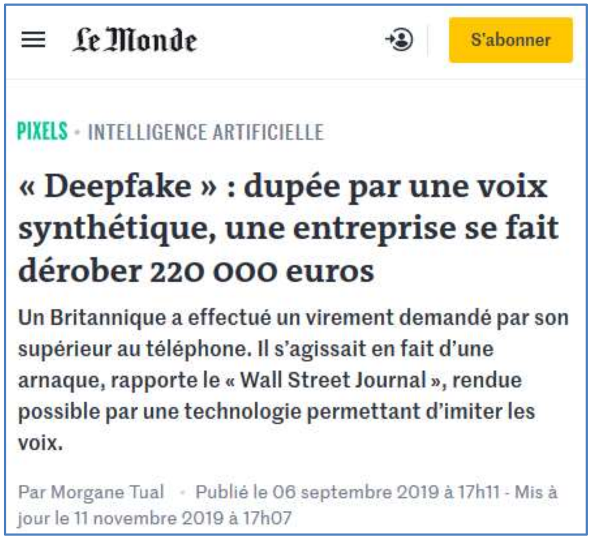
Updated Aug. 30, 2019 12:52 pm ET

https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402



Voice Double: Create a digital voice that sounds like you from a small audio sample.

https://www.descript.com/lyrebird-ai?source=lyrebird



PIXELS • INTELLIGENCE ARTIFICIELLE

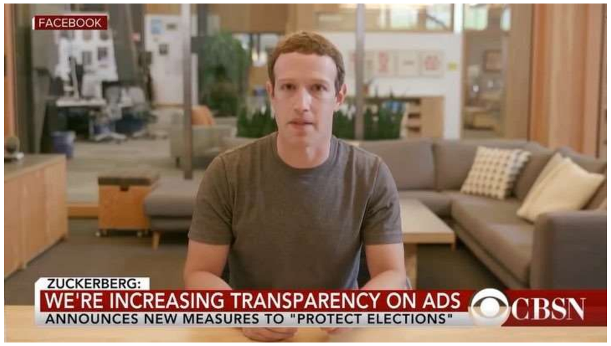
« Deepfake » : dupée par une voix synthétique, une entreprise se fait dérober 220 000 euros

Un Britannique a effectué un virement demandé par son supérieur au téléphone. Il s'agissait en fait d'une arnaque, rapporte le « Wall Street Journal », rendue possible par une technologie permettant d'imiter les voix.

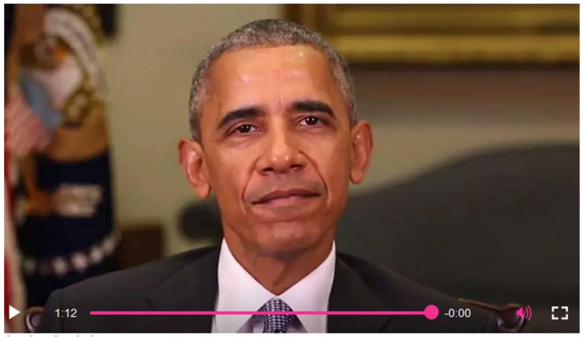
Par Morgane Tual • Publié le 06 septembre 2019 à 17h11 - Mis à jour le 11 novembre 2019 à 17h07

https://www.lemonde.fr/pixels/article/2019/09/06/deepfake-dupee-par-une-voix-synthetique-une-entreprise-se-fait-derober-220-000-euros_5507365_4408996.html

Deep fakes – Videos


FACEBOOK
ZUCKERBERG:
WE'RE INCREASING TRANSPARENCY ON ADS
ANNOUNCES NEW MEASURES TO "PROTECT ELECTIONS"
CBSN

https://www.instagram.com/p/ByaVigGFP2U/?utm_source=ig_embed&utm_campaign=embed_video_watch_again



https://youtu.be/F2mwz_cnAIk
https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed

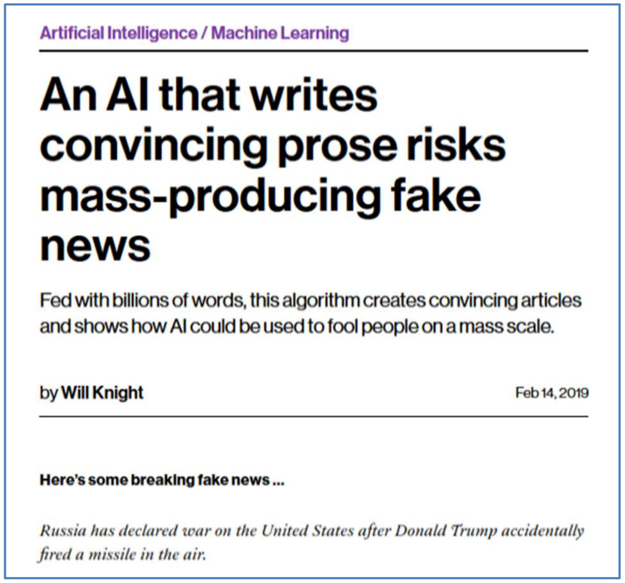
http://svencharleer.com/2018/02/02/family-fun-with-deepfakes-or-how-i-got-my-wife-onto-the-tonight-show/

Deep fakes – Text

Russia has declared war on the United States after Donald Trump accidentally fired a missile in the air. Russia said it had "identified the missile's trajectory and will take necessary measures to ensure the security of the Russian population and the country's strategic nuclear forces." The White House said it was "extremely concerned by the Russian violation" of a treaty banning intermediate-range ballistic missiles. The US and Russia have had an uneasy relationship since 2014, when Moscow annexed Ukraine's Crimea region and backed separatists in eastern Ukraine.

Once upon a time there was a fairy king. His name was Luigi, and for five years he ruled his kingdom together with Princess Toadstool. All of a sudden, all of a sudden, a terrible flood came and all of the fairy land was submerged. Luigi decided that the only way for his land to be rebuilt was for a clever but invisible woman to come and show the way. After a lot of thought, the princess found this hard task to be an absolute pain and so decided that she wanted to take this task on herself and go rescue the lost inhabitants. After a short period of training, Mario and Luigi set out to find Princess Toadstool, in the hopes that this journey will grant them their desire to travel the world, then finally come back and live happily

Try it on https://talktotransformer.com/



Artificial Intelligence / Machine Learning

An AI that writes convincing prose risks mass-producing fake news

Fed with billions of words, this algorithm creates convincing articles and shows how AI could be used to fool people on a mass scale.

by **Will Knight** Feb 14, 2019

Here's some breaking fake news ...

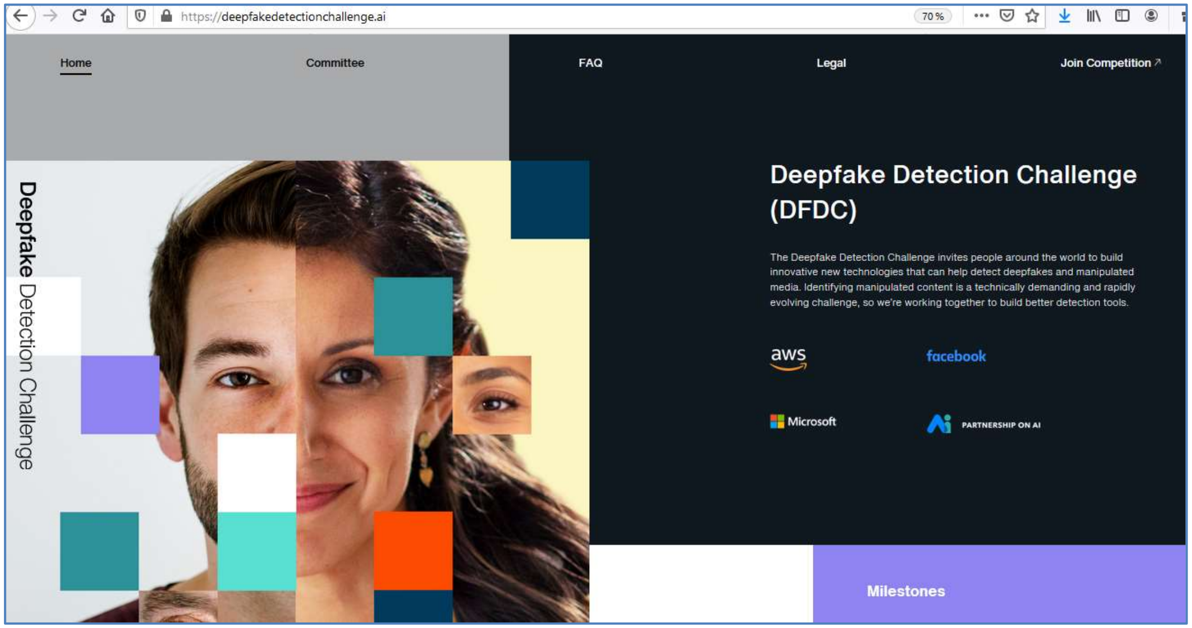
Russia has declared war on the United States after Donald Trump accidentally fired a missile in the air.

https://www.technologyreview.com/s/612960/an-ai-tool-auto-generates-fake-news-bogus-tweets-and-plenty-of-gibberish/?utm_campaign=the_download.unpaid.engagement&utm_source=hs_email&utm_medium=email&utm_content=69948867&_hsenc=p2ANqtz-89F1HFExtrMDp0rgK4mJ_c1LGRRdNZkiooePnsYoq-yMqLLnoP34SV73DsOKHgSGQw3hGk_MuDxK3otJK4AZe_JjThAI3ycB4MaflDdRdSmmAyl4g&_hsmi=69948867

https://www.theverge.com/2019/11/7/20953040/openai-text-generation-ai-gpt-2-full-model-release-1-5b-parameters

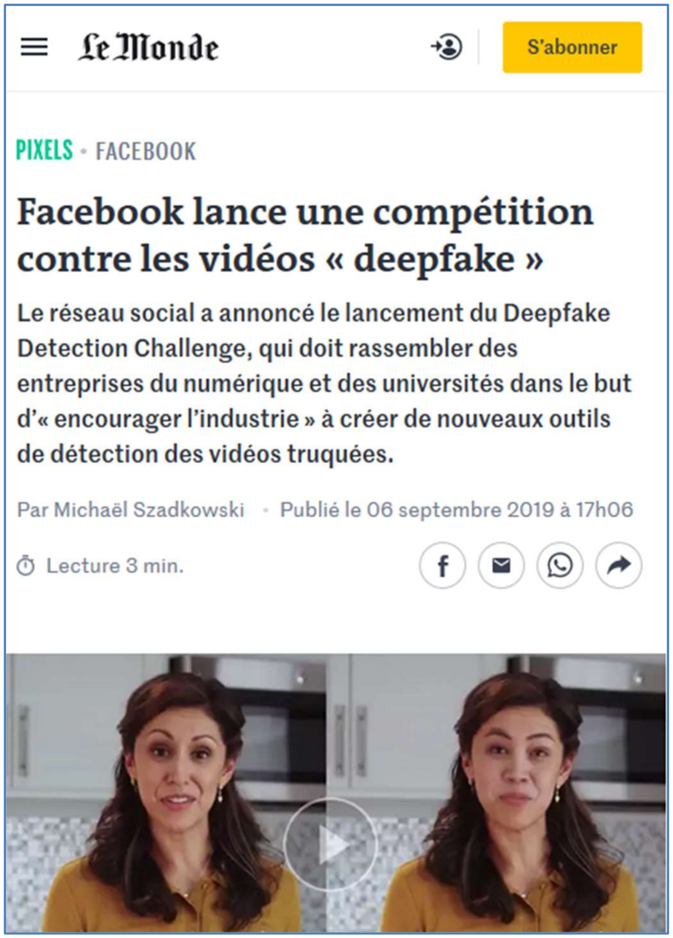
Risks posed by deep fakes

- For individuals, companies & countries
- So the fight against fakes is starting



Deepfake Detection Challenge (DFDC)

The Deepfake Detection Challenge invites people around the world to build innovative new technologies that can help detect deepfakes and manipulated media. Identifying manipulated content is a technically demanding and rapidly evolving challenge, so we're working together to build better detection tools.



Facebook lance une compétition contre les vidéos « deepfake »

Le réseau social a annoncé le lancement du Deepfake Detection Challenge, qui doit rassembler des entreprises du numérique et des universités dans le but d'« encourager l'industrie » à créer de nouveaux outils de détection des vidéos truquées.

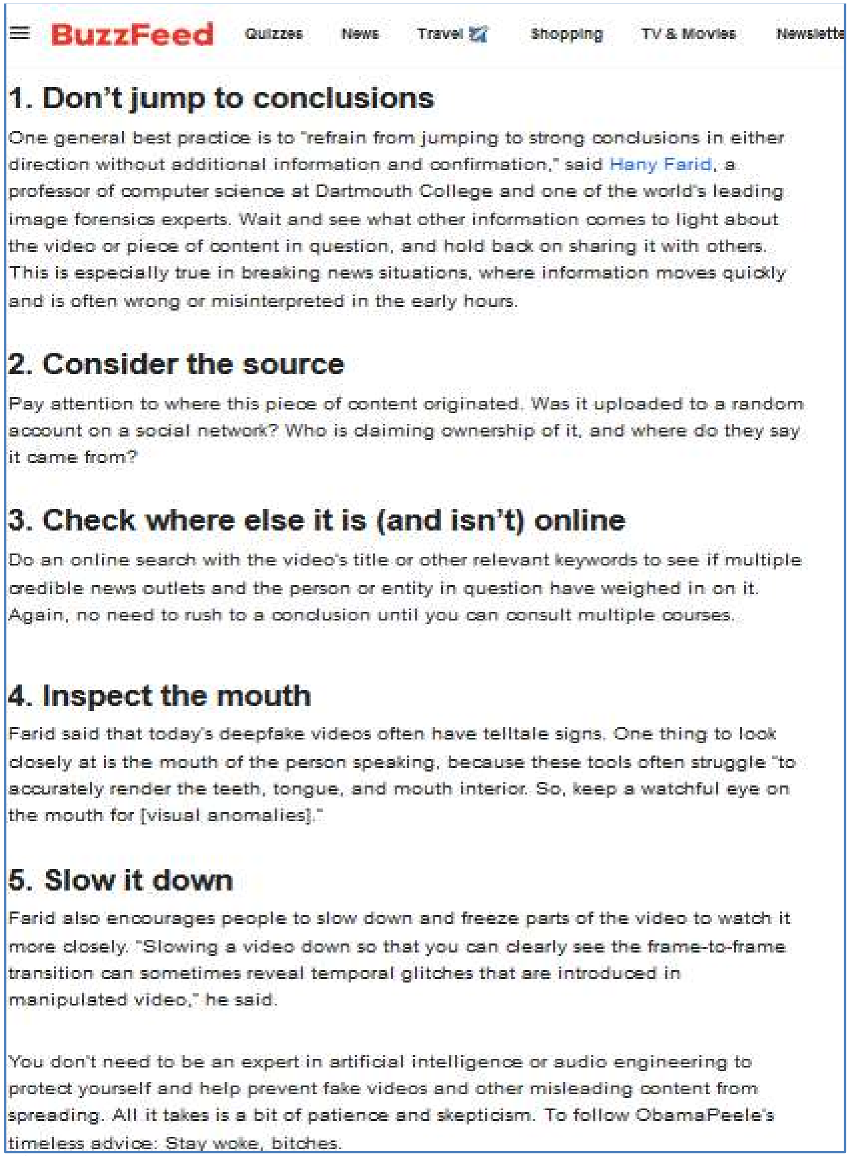
Par Michaël Szadkowski · Publié le 06 septembre 2019 à 17h06

https://deepfakedetectionchallenge.ai/
https://www.lemonde.fr/pixels/article/2019/09/06/facebook-lance-une-competition-contre-les-videos-deepfake_5507362_4408996.html#xtor=AL-32280270

How can you tell

- It is hard but there are signs
- Automated detection is needed
  - But research is still in its infancy (→ Facebook challenge)
- A never-ending fight for the future

https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed?utm_term=.dvGOqXPd6#.qqPxNEwlV



1. Don't jump to conclusions

One general best practice is to "refrain from jumping to strong conclusions in either direction without additional information and confirmation," said Hany Farid, a professor of computer science at Dartmouth College and one of the world's leading image forensics experts. Wait and see what other information comes to light about the video or piece of content in question, and hold back on sharing it with others. This is especially true in breaking news situations, where information moves quickly and is often wrong or misinterpreted in the early hours.

2. Consider the source

Pay attention to where this piece of content originated. Was it uploaded to a random account on a social network? Who is claiming ownership of it, and where do they say it came from?

3. Check where else it is (and isn't) online

Do an online search with the video's title or other relevant keywords to see if multiple credible news outlets and the person or entity in question have weighed in on it. Again, no need to rush to a conclusion until you can consult multiple courses.

4. Inspect the mouth

Farid said that today's deepfake videos often have telltale signs. One thing to look closely at is the mouth of the person speaking, because these tools often struggle "to accurately render the teeth, tongue, and mouth interior. So, keep a watchful eye on the mouth for [visual anomalies]."

5. Slow it down

Farid also encourages people to slow down and freeze parts of the video to watch it more closely. "Slowing a video down so that you can clearly see the frame-to-frame transition can sometimes reveal temporal glitches that are introduced in manipulated video," he said.

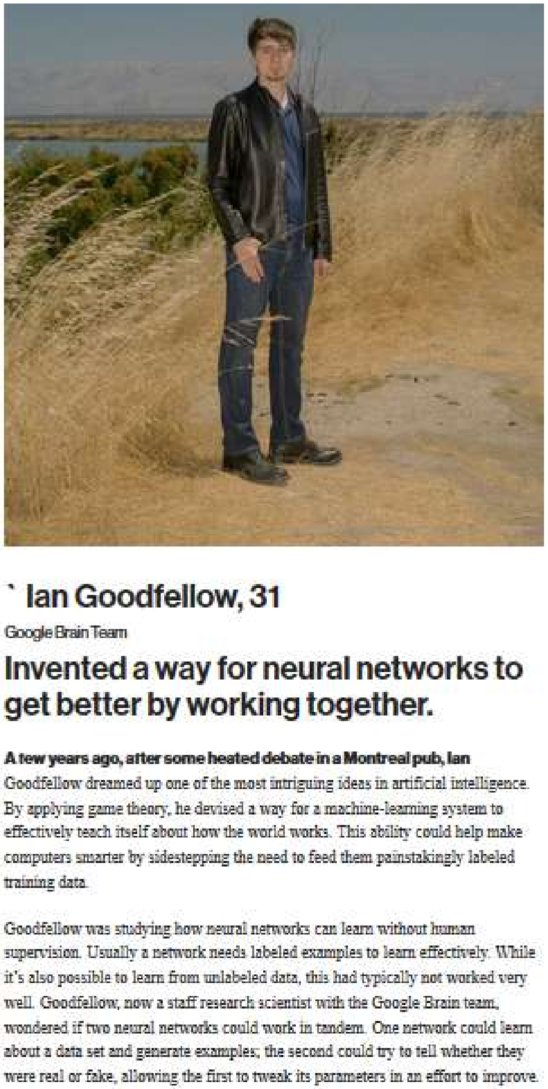
You don't need to be an expert in artificial intelligence or audio engineering to protect yourself and help prevent fake videos and other misleading content from spreading. All it takes is a bit of patience and skepticism. To follow ObamaPeele's timeless advice: Stay woke, bitches.

HOW IS IT DONE

GAN

- In 2014, Ian Goodfellow invented a “way for a machine-learning system to effectively teach itself about how the world works”
- He called this system a Generative Adversarial Network
  - Actually 2 networks !

https://www.technologyreview.com/lists/innovators-under-35/2017/inventor/ian-goodfellow/



` Ian Goodfellow, 31

Google Brain Team

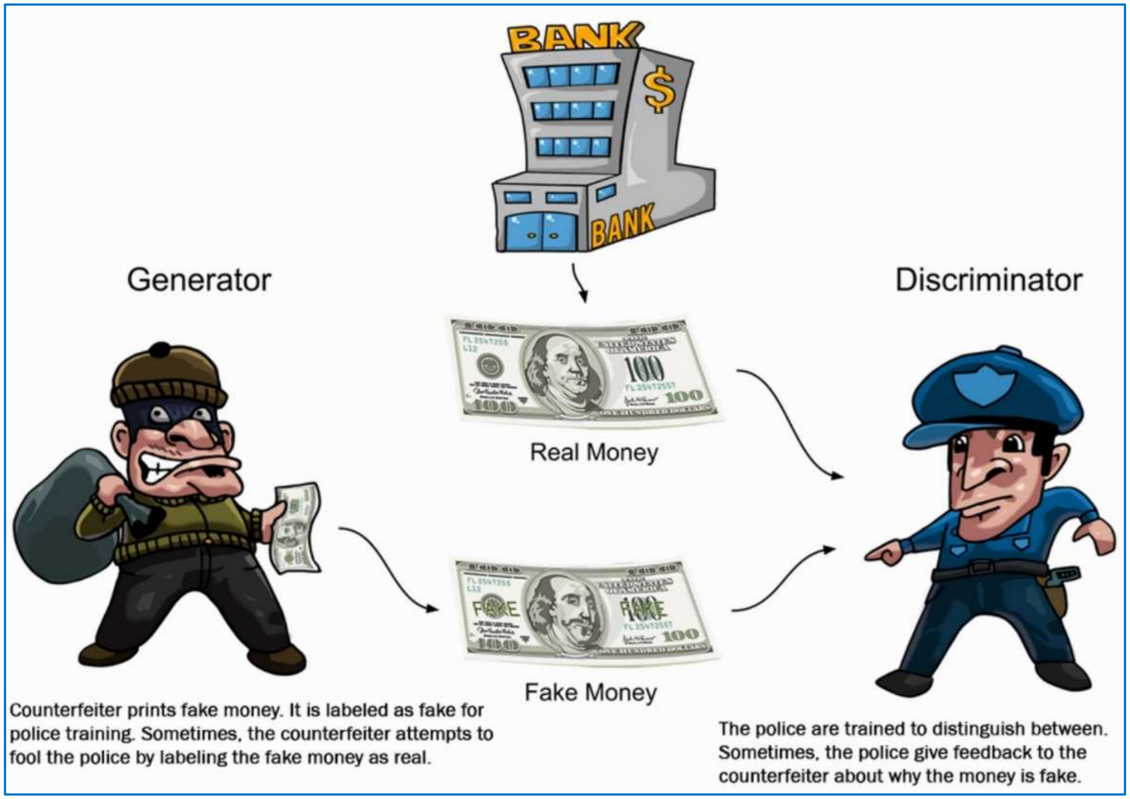
Invented a way for neural networks to get better by working together.

A few years ago, after some heated debate in a Montreal pub, Ian Goodfellow dreamed up one of the most intriguing ideas in artificial intelligence. By applying game theory, he devised a way for a machine-learning system to effectively teach itself about how the world works. This ability could help make computers smarter by sidestepping the need to feed them painstakingly labeled training data.

Goodfellow was studying how neural networks can learn without human supervision. Usually a network needs labeled examples to learn effectively. While it's also possible to learn from unlabeled data, this had typically not worked very well. Goodfellow, now a staff research scientist with the Google Brain team, wondered if two neural networks could work in tandem. One network could learn about a data set and generate examples; the second could try to tell whether they were real or fake, allowing the first to tweak its parameters in an effort to improve.

The idea

- 2 networks fight against each other
  - Discriminator ("Police") tries to perfectly distinguish real from fake
  - Generator ("Rogue") tries to produce perfect fakes & fool Discriminator



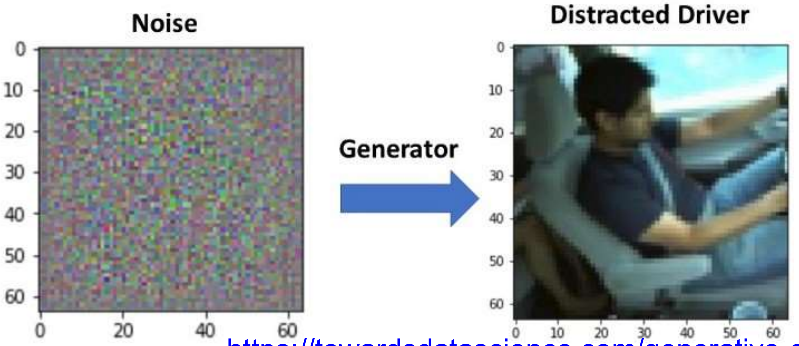
Counterfeiter prints fake money. It is labeled as fake for police training. Sometimes, the counterfeiter attempts to fool the police by labeling the fake money as real.

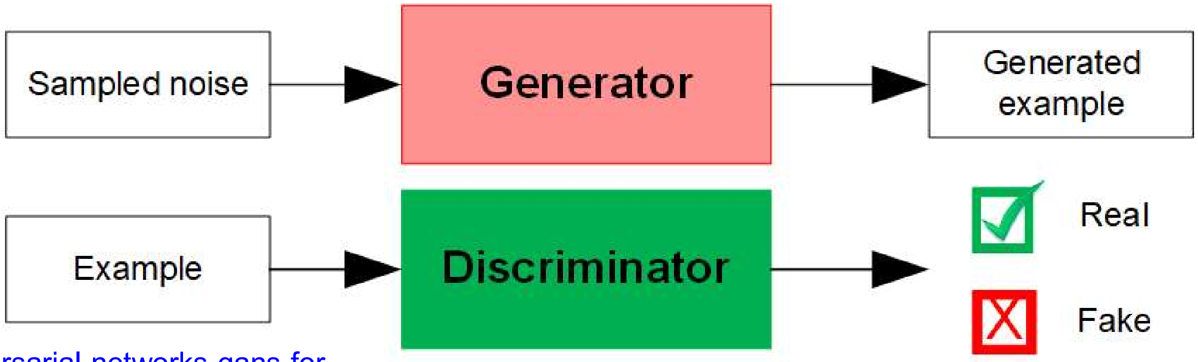
The police are trained to distinguish between. Sometimes, the police give feedback to the counterfeiter about why the money is fake.

https://towardsdatascience.com/the-math-behind-gans-generative-adversarial-networks-3828f3469d9c

The 2 networks

- Discriminator (“the Police”) receives an example
  - It classifies it as real or fake
- Generator (“the Rogue”) generates examples
  - In same format as training examples (from noise sampled from some distribution) : the fakes





https://towardsdatascience.com/generative-adversarial-networks-gans-for-beginners-82f26753335e

The Discriminator

- Discriminator is a deep network receiving as inputs
  - Examples from training set (the "real") and examples generated by Generator (the "fakes") (50/50)
- It learns in supervised mode, maximizing the cross-entropy loss ("log-loss" measures how different distributions of real & fake are)

$$L[D(x), y] = y.\,log[D(x)] + (1 - y).\,log[1 - D(x)]$$

Where $\hat{y} = D(x)$ is the output produced for input x

Training Discriminator

- Since we have $y = 1$ (real) and $y = 0$ (fake)

$$L[D(x), 1] = log[D(x)]$$

$$L\left[D\left(G(z)\right), 0\right] = log\left[1 - D\left(G(z)\right)\right]$$

where $x = G(z)$ is used as input to the Discriminator & $G(z)$ is the output produced by Generator for input z

- So Discriminator learns to solve

$$L^{(Disc)} = \max_{D}\left[log\left(D(x)\right) + log\left(1 - D\left(G(z)\right)\right)\right]$$

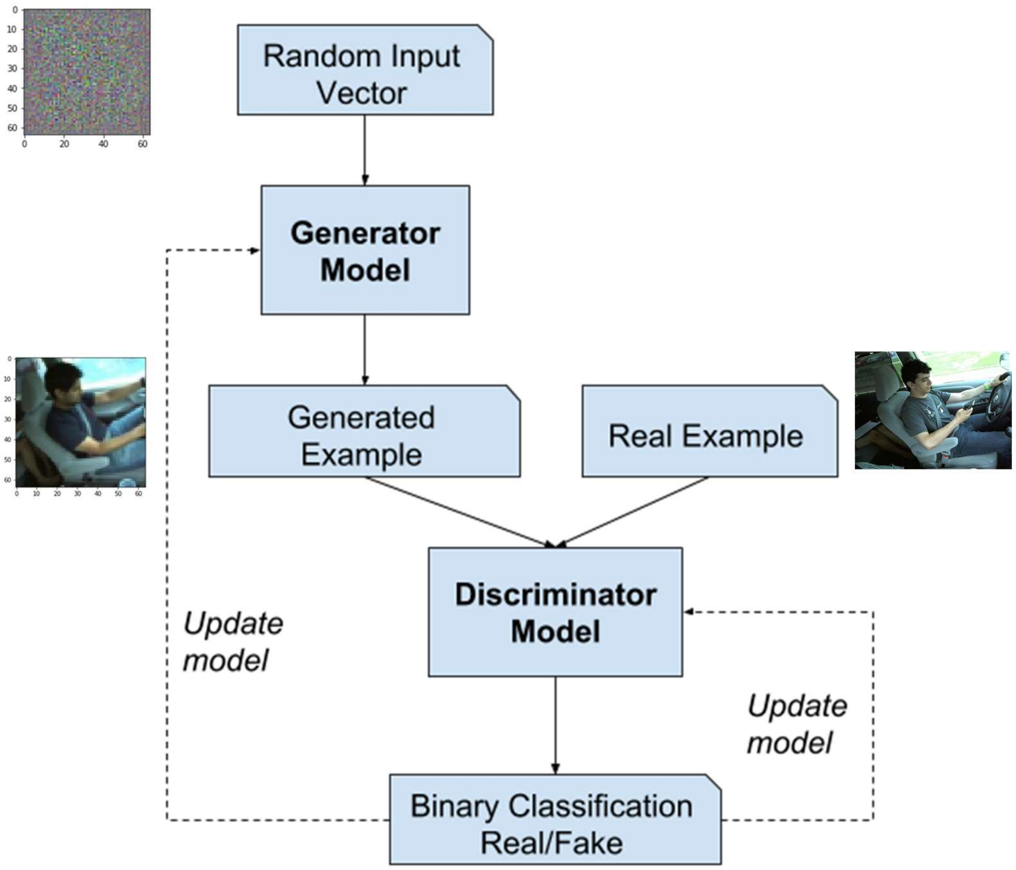
The Generator

- Generator is a deep network receiving as inputs
  - Examples sampled from noise (with some prior)
- It learns in unsupervised mode, to <u>minimize</u> the log-loss, which the Discriminator tries to maximize:

$$L^{(Gen)} = \min_{G}\left[log\big(D(x)\big) + log\left(1 - D\big(G(z)\big)\right)\right]$$

Training the GAN

- Use Generator to generate examples
- Train Discriminator on real & generated data
  - Update Discriminator
  - Update Generator



https://towardsdatascience.com/generative-adversarial-networks-gans-for-beginners-82f26753335e
https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Training the GAN

- The global loss is thus (for one point)

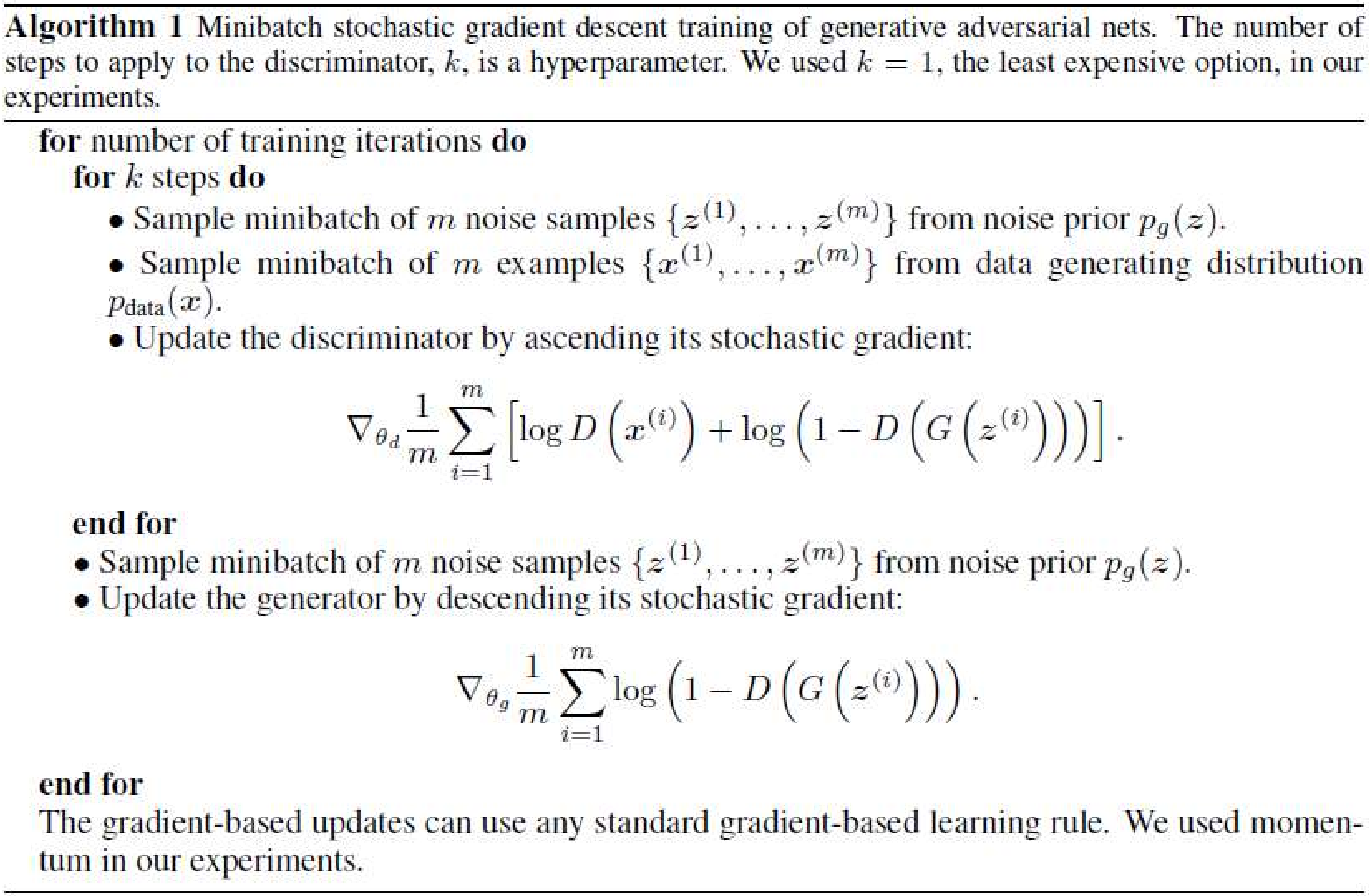
$$l(x) = \min_G \max_D \left[log\left(D(x)\right) + log\left(1 - D\left(G(z)\right)\right)\right]$$

And for the entire dataset:

$$L = \min_G \max_D \frac{1}{m} \sum_i \left[\left(log\left(D\left(x^i\right)\right)\right) + \left(log\left(1 - D\left(G\left(z^i\right)\right)\right)\right)\right]$$

- Alternate
  - k steps of optimizing for D (up-gradient of L)
  - One step of optimizing for G (down-gradient of L)

Training the GAN



Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

```
for number of training iterations do
    for $k$ steps do
        • Sample minibatch of $m$ noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$.
        • Sample minibatch of $m$ examples $\{x^{(1)}, \dots, x^{(m)}\}$ from data generating distribution $p_{\text{data}}(x)$.
        • Update the discriminator by ascending its stochastic gradient:
              $$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right) \right].$$
    end for
    • Sample minibatch of $m$ noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$.
    • Update the generator by descending its stochastic gradient:
          $$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right).$$
end for
The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.
```

k iterations up-gradient for D

1 iteration down-gradient for G

https://arxiv.org/abs/1406.2661

Training the GAN

- Training GANs is notoriously difficult
  - The loss function has some limitations (vanishing gradient)
  - Generator may get stuck into mode collapse: Discriminator and Generator need to be synchronized (G must not be trained too much without updating D)
- Look for tricks to train GANs https://github.com/soumith/ganhacks, https://arxiv.org/abs/1701.00160

Conclusion

- GAN is a powerful technique to generate fakes : i.e. data very similar to training data
- Today fakes are generated for images, audio, videos and text with deep networks
  - They might be put to dangerous use
  - It is still hard to automatically tell fakes from reals
- Training a GAN entails solving a min max problem
  - It can be hard and still requires many tricks

References

1. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672-2680. 2014. https://arxiv.org/abs/1406.2661
2. Mayank Vadsola. The math behind GANs (Generative Adversarial Networks). https://towardsdatascience.com/the-math-behind-gans-generative-adversarial-networks-3828f3469d9c
3. Soumith Chintala, Emily Denton, Martin Arjovsky, Michael Mathieu. How to Train a GAN? Tips and tricks to make GANs work. https://github.com/soumith/ganhacks
4. Ian Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. https://arxiv.org/abs/1701.00160

On line courses

1. Ian Goodfellow. Tutorial on Generative adversarial networks – Introduction. ICCV17. November 2017. https://www.youtube.com/watch?v=sgHdUYHGvtA