Deepfakes & Algorithms

Threat or Opportunity?

Octobre 2021

(club) prAxis

Contents

The authors

Supervision

Emilie Rannou is a graduate of ENSAE and holds a master's degree in applied mathematics from Paris Dauphine University. As a Senior Manager at Ekimetrics, she advises various companies on their data science strategy and supervises teams of data science consultants to implement and optimize new models. Emilie has twelve years' experience in the tech and financial industry, notably as an Executive Director at JPMorgan Chase in New York and at Criteo as a Senior Product Manager.

Work group

Alexis Benichoux holds a doctorate in signal processing from the University of Rennes. His work focused on learning methods applied to sound signal processing. Alexis is also a graduate of the ENS Cachan MVA (Mathematics, Vision and Learning) master program. Now, as a research engineer at Yubo he develops algorithms for image recognition and content recommendation.

Rémi Forgeas leads the France Business Services Center of an international audit and consulting firm. Based in the United States for more than 20 years, he assists French groups in their development in the United States. He graduated from Sciences Po Paris and has a master's degree in econometrics.

Simon Gaillard is a data protection consultant. Based in New York City, he specializes in online privacy and assists businesses with their GDPR and CCPA compliance strategy. A graduate of Sciences Po Paris, he also advises international organizations on public policies, new technologies and data protection issues.

Jérémie Mary has a doctorate in computer science from the University of Paris-Sud. He is a lecturer, authorized to supervise research at the University of Lille, and is currently on secondment as a researcher at Criteo. He is an associate member of the Inria SequeL research group, which specializes in sequential decision-making. He won and organized three challenges for recommendation systems in major machine learning conferences using Yahoo !, Adobe and Twitter data (ICML 2011, ICML 2012, RecS ys 2014). His current research work is oriented around recommendation sequences and the use of generative models for recommendation systems.

Minh Trinh is a graduate of the Ecole Polytechnique and a Doctor in Economics from Harvard University. He is a consultant based in New York and specializes in quantitative and machine learning methods applied to economics and finance. He is the author of a book on the management of artificial intelligence projects. He is a member of the American Finance Association and the CFA Institute.

Gabriel Turinici is a graduate of ENS Ulm and a former member of the Institut Universitaire de France. He was a researcher in mathematics and computer science at INRIA until 2005 and was appointed since then as a professor at the University of Paris Dauphine-PSL. He specializes in simulation algorithms, with applications in quantum physics, finance and medicine. He is the author of more than 100 scientific publications and a member of the editorial board of several journals.

Emilie Waxin has been a lawyer at the Paris Bar since 2010. After practicing in France and Luxembourg for eight years, she created her own firm, WE Avocatsarly in social law, comm, in June 2017. She practices both as a litigation counsel in general commercial matter, with a special focus on intellectual property law. She also teaches litigation at the Paris Bar Training School (EFB) and is a member of several societies such as the AFFJ (Association Française des Femmes Juristes).

About Club Praxis

Club Praxis is a Franco-American think tank created in 2007, directed by several French leaders, including Henri de Castries, former chairman and CEO of AXA. They position themselves as an outsider in French public life and seek to promote new and innovative ideas for the regeneration of institutions and the economic system on the basis of the following principles: stability, equity, transparency, citizen-friendly decision-making, lifelong education and financial independence.

Pragmatic and non-partisan position

Independent of any school of thought, Club Praxis champions a renewed dialogue entered into with an open mind between citizens and political decision-makers, in particular through greater transparency of public data and decision-making processes. Through concrete and actionable reform recommendations, it aims to put the long-term general interest back at the heart of public action, and help define a new path for a France at the forefront of research and more open to the world.

Independent of any subsidy

Club Praxis operates on a completely voluntary basis; it does not receive any public or private grants. It consists mainly of researchers, economists, senior officials, senior executives and lawyers who want to put their skills and experience at the service of their country. This report reflects the views only of its authors.

Acknowledgments

The authors would like to thank everyone who made suggestions, in particular Yann Coatanlem, Guy Sorman, Xavier Dupré, Pablo Winant, Adélaïde Sorel, Eric Nahon, the law firm Bold and the data science consulting firm Ekimetrics for their contribution to the report. Thanks also to Vannina Malekzadeh who illustrated the report.

Our twelve recommendations to reposition Europe as an expert on content generation algorithms and a leader in the fight against deepfakes

We are suggesting twelve ways to make Europe an expert in content generation algorithms and a leader in the fight against deepfakes. How? By defining standards of governance and regulation around deepfakes and by protecting citizens through education.

A. Make Europe a leader in the fight against deepfakes

1. **Help improve and develop techniques for classifying content**, whether true, false or suspicious (fact-checkers, watermarks, algorithms, blockchain), while defining European standards, to support the population in identifying reliable content.
2. **Support scientific research in the field of content generation algorithms**, to be able to better combat them but also to use them for more positive applications for society.
3. **Foster European partnerships**, pool resources to acquire the means to catch up with the progress of China, Russia and the United States in this area.

B. Strengthen the accountability of platforms at European level

4. **Reinforce the responsibilities required of platforms** (for example within the Digital Services Act). Require each platform to set up an internal governance structure for the regulation of its content. Place each governance body under European supervision, drawing inspiration from the regulations implemented in the banking sector following the 2008 crisis.
5. **Encourage platforms to implement tools to reduce the impact of deepfakes**

   Sanctions such as:

- A penalty for authors publishing non-obvious false or malicious content.
- A clear statement warning the user that the publication could not be verified, and may even be suspicious.
- A limit on the number of recipients of a publication.
- A limit on transfers per user for each publication.

6. **Require the platforms to share with judicial authorities** any information they hold on the alleged perpetrators of deepfakes, as part of criminal investigations, without the platforms being able to refuse.

C. Build a regulatory environment adapted to an effective fight against deepfakes

7. **Promote access to their rights for victims of deepfakes**, identity theft or harassment (to legal remedies and information about their rights) through the development and awareness of support structures (whether through physical offices or via government platforms).
8. **Strengthen existing criminal sanctions** against authors of content, and reassess them taking into account the new technological and digital environment where the creation of deepfakes is made possible.
9. **Strengthen the responsibilities of platforms at state level, but above all at supra-state level (the transmission of information knowing no borders).** To this end:
    - Oblige platforms to verify and keep identification data of new accounts, and, in the absence of real identification of the author, close the account without delay
    - Limit and legally quantify the withdrawal period for illegal content reported to platforms under penalty of a heavy fine

- Require hosts located outside France to appoint a representative in France who assumes the responsibility of the host who has not withdrawn within this legally fixed period content reported as illegal
- Allow in some exceptional cases, where there is clear evidence of illegal content, to order the permanent removal of a site, a profile, an account, a page, a group, without giving the site in question the opportunity to defend itself
- Systematically combine the financial sanctions imposed on platforms with publicity measures
- Develop a supranational organization for cyber-policing and judicial cooperation focused on the matter in order to facilitate the exchange of information between countries

D. Protect citizens from the impact of deepfakes

10. **Make the fight against deepfakes a European objective for 2021.** Set up awareness campaigns to explain and educate about deepfakes, their uses and the associated risks in terms of the misappropriation of images, voices and videos, while highlighting the possible applications to manipulate public opinion

11. **Increase the authors' responsibilities** Require authors to report any deepfake content.

12. **Involve employers in the fight against deepfakes.** Encourage them to adopt ethical charters, to inform their employees, customers, suppliers and partners of their good ethical practices on the subject. Encourage training around deepfakes.

Objective of the report

In this report, we look specifically at **content generation algorithms**. What are they capable of today? How do they work? How well do we recognize content produced by an artificial intelligence algorithm? What are **GANs**, the Generative Adversarial Networks, the algorithms behind the most sophisticated deepfakes? Are they only dedicated to making deepfakes, or can we identify beneficial applications for society? How can we raise public awareness to better recognize a deepfake and thwart manipulation attempts? What legal framework is in place, and how should it evolve to adapt to new algorithms? We will address multiple ethical questions related to this subject, but also explore how it is possible to better evolve in an environment where content generation algorithms are a growing presence.

Our objective is, primarily, educational: a **desire to popularize the subject for public authorities and companies**, and to build awareness of the vertiginous issues raised by content generation algorithms. Secondly, it is also to **weigh in on the debate around the digital, regulatory and legal** framework for deepfakes as well as on the economic and social debate, in France and in Europe, in order to make the population more aware of these issues.

It is with this in mind that we, thinkers, researchers and experts in the scientific, legal and ethical fields, all passionate about the subject of content generation algorithms and deepfakes, have come together to present this report, at a worrying time in history, when it is becoming easier to manipulate content and where the power of digital leaders to influence opinions and elections continues to grow. We feel there is an urgency to realize more widely that content generation algorithms have opened a real Pandora's box. Developing more in-depth knowledge in this sector will be essential to position Europe as the leader in the fight against deepfakes.

Introduction

In May 2019, Facebook came under fire: the social network let circulate a fake video of the Speaker of the United States House of Representatives, Nancy Pelosi, in which she appeared to be drunk. A special effect made possible by slowing down the voice while keeping a credible visual. The video went viral and was seen by more than 2 million people in two days, causing a wave of reactions on social networks. The video was also shared by relatives of President Donald Trump, during a period of high tensions between Nancy Pelosi and the American president, each seeking to discredit their adversary. Faced with pressure to remove the video, **Facebook invoked freedom of expression**. Considering itself as neither guarantor nor judge of the veracity of contents shared online, the company announced its refusal to remove the video even though it was misleading and likely to manipulate public opinion. Facebook nevertheless offered to label the video "partially false".

The debate around manipulation and freedom of expression is not new. However, it is increasingly important. On the one hand, the growing power of social media networks make it possible to reach an increased number of people, and on the other hand technological progress that favors the creation of false content and in particular deepfakes: fake content created using technologies called deep learning. However, Facebook changed their position in early 2020 following the staging of its founder and CEO Mark Zuckerberg in a fake video that shows his likeness announcing that he is ready to use billions of people's stolen data to control the future and that he owes everything to a secret organization called "Spectre". Although the video, which is the work of two hacktivist artists, is of poor quality and the subterfuge is obvious, could it have weighed in the balance to convince Mark Zuckerberg of the harmful power of deepfakes? Be that as it may, in January 2020, Facebook issued a press release saying it will ban them.

Now deepfakes are deployed as a tool for **manipulation**, even **propaganda**, and their impact on society is not yet kept in check. Their existence and their use raise many legal and ethical questions; however, beyond the laws and regulations that may be implemented to

control them, a fundamental problem remains: that of the ability to detect a deepfake. As technology evolves, it's getting harder and harder to spot a fake: according to MIT researcher Hao Li, we're just a few months away from completely undetectable deepfake videos. Developing European knowledge of these technologies, already mastered by many experts, particularly in China and Russia, therefore appears urgent in order to better understand and control them.

The Villani report, published in 2018, details why **France must imperatively invest more in the field of artificial intelligence** (AI). For decades, artificial intelligence has fueled fears and fantasies, inspiring film directors and writers, but until recently its impact on society was tenuous. Indeed, artificial intelligence did not yet worry humanity, as it only repeated what humans had taught it. However, a first event made its mark in 1997: the computer *Deep Blue* (IBM) beat the world chess champion Gary Kasparov. The event shook up a hesitant society: should it be concerned with the imminent and nevertheless foreseeable revolution that was looming? Eventually, it was above all the advances in processing power of computers over the past ten years that have put artificial intelligence algorithms back in the spotlight. Until then, the major obstacle for AI were the long learning times involved.. Thanks to improvements in computing power, this has now been reduced. The second issue was the availability of data for training. This was alleviated by the advent of digital technology, at least in part, through the proliferation of applications and the ability to collect, store and share massive amounts of data, in particular using cloud-based storage.

Therefore, it seems essential to raise awareness of these technologies among the general public. Our **main objective is to explain how they work and to offer suggestions on their use and control**. Thus, we will start by showing to what extent their fields of application are rich and varied, in particular in the generation of texts, images, sounds and videos, by giving concrete examples of both harmful and beneficial application for society. We will then look at the ethical issues which could be particularly worrisome with imminent political events. Then in a third part we will study the regulations around deepfakes and how we envisage

their evolution. Finally, our desire being to protect the people, we will present existing approaches to limit the impact of deepfakes.

1. Introduction to *deepfakes* and content generation algorithms

It usually takes several years for research and development fields to find concrete and useful applications for humanity. This was not the case for content generation algorithms, as three years after their creation in 2017, GANs are already widely used by the general public, and their main application, deepfakes, is starting to gain attention. One simple approach to measuring interest in a topic is to look at Google statistics. Below is an excerpt from the statistics for deepfake research around the world.

Source: Google Trends

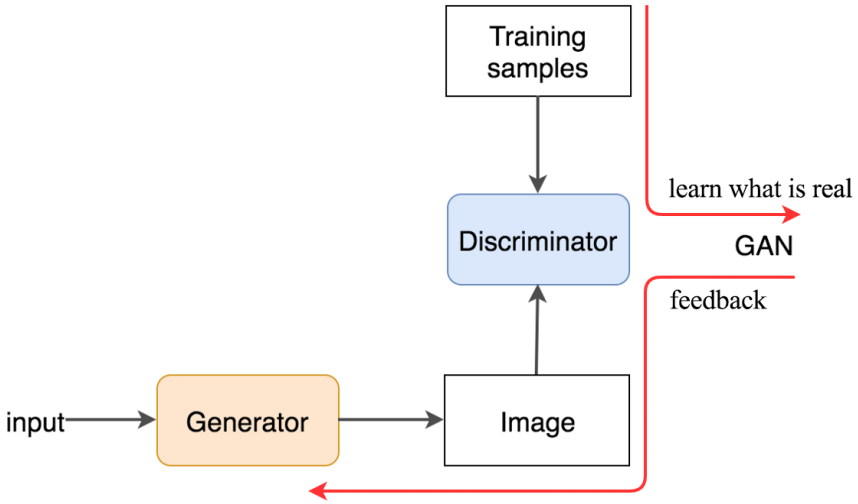
The first search spike detected by Google Trends appears in January 2018. It corresponds to the deepfake porn scandal (pornographic images and videos created by deep learning algorithms). Interest in deepfakes then drops rapidly, before gradually rising again in 2019. In fact, two deepfakes published in 2019 created the buzz. One featured Barack Obama insulting Donald Trump, the other showed Mark Zuckerberg admitting to stealing personal data thanks to a secret organization called Spectre. It is also in September 2019 that Facebook launched the Deepfake Detection Challenge, which we will talk about later.

But few are those who fully understand what a deepfake is.

What is a deepfake? It is fake content generated by deep learning algorithms, hence the name of deepfake. But what is deep learning? First of all, we have to introduce what is called

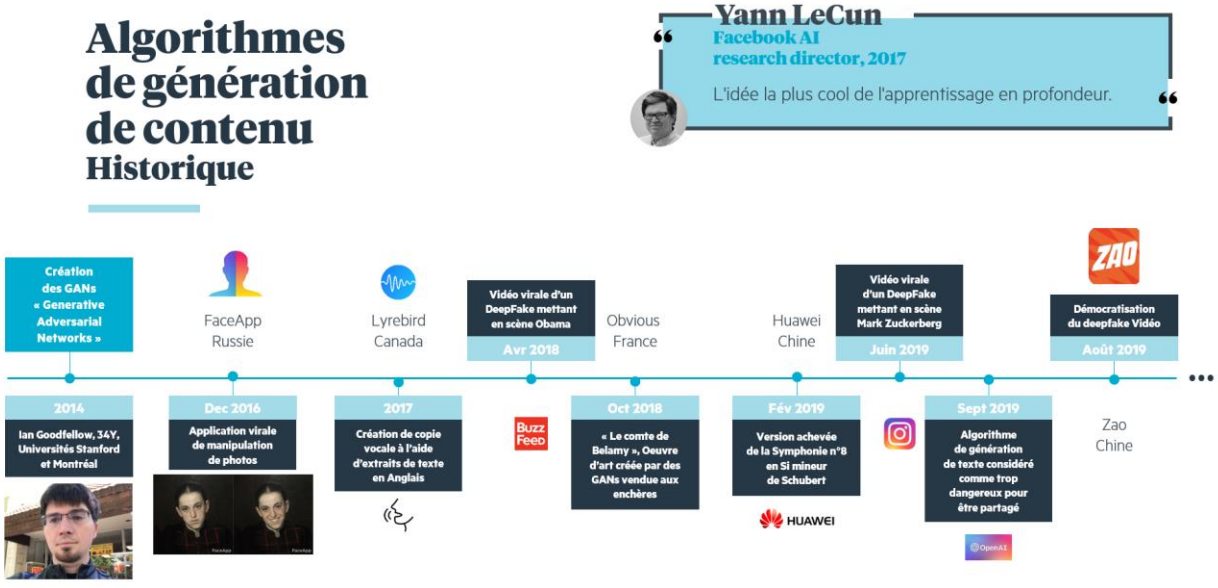
a neural network. It is an artificial intelligence algorithm that is inspired by the functioning of the brain, and in particular the transfer of information between neurons via synapses. A single layer of neurons allows the algorithm to learn simple rules and to classify, for instance, a word into feminine, masculine or neutral using training on a dictionary. A neural network with several layers will allow more complex learning, such as learning a concept, and will identify whether an image represents a cat, a dog or a monkey based on similar images. These networks with several layers are what we call deep networks, and deep learning brings together all the algorithms developed using deep neural networks.

In this report, we will take a closer look at the flagship algorithms used for the creation of deepfakes: GANs. The method of these content generation algorithms, for which we will give more details in the technical part, is to confront two deep networks: a generator network (generator) and a discriminator network (discriminator). To take a concrete example, let's imagine a case where we want to generate a photo of a car in the style of the painter Gauguin. The generator network transforms the initial photo of the car and generates a slightly modified image using what it has learned from Gauguin's paintings. This generated image, which is therefore not a real image, will then be confronted by the discriminator which, like an inspector, can precisely evaluate the image produced by the generator with the objective of concluding whether it is false or real. If it concludes that it is real, the discriminator then learns that it was wrong and improves its knowledge. If it concludes that it is false, the generator learns that its creation was not good enough and also improves its knowledge, which allows it to offer a new copy to the discriminator. And so on... This process allows the generator to produce more and more realistic data.



Content generation algorithms were introduced in 2014 by Ian Goodfellow, then a researcher at the University of Montreal (he also obtained a master's degree in computer science from Stanford University). After working for Google Brain and a collaboration with OpenAI, a non-profit organization founded by Elon Musk among others, he now works at Apple.

Here is a history of content generation algorithms:



We will detail these many applications later.

1.1. Image generation

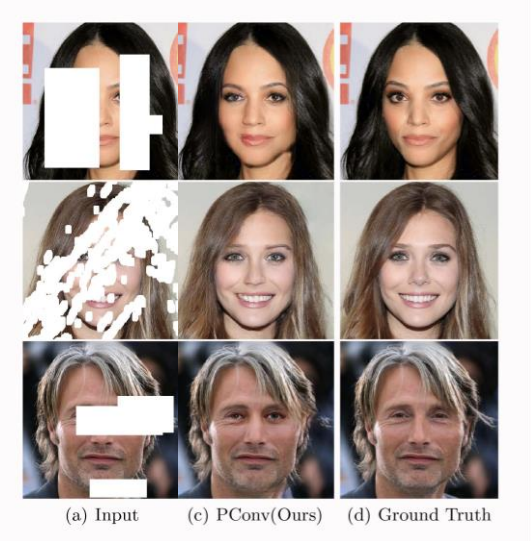
1.1.1. Fun and creative apps

Anyone who owns a smartphone can now have fun applying filters to style their photos. Indeed among others, and this is not the only one, the Deep Art Effects application allows you to generate images in the style of great painters such as Monet or Van Gogh. Few people suspect that behind these filters lies one of the latest machine learning technologies, and in particular a specific category of GANs (CycleGAN[1]). For instance, below we see a photograph stylized through learning the style of many existing paintings.

Source: Jun-Yan Zhu and al., 2017

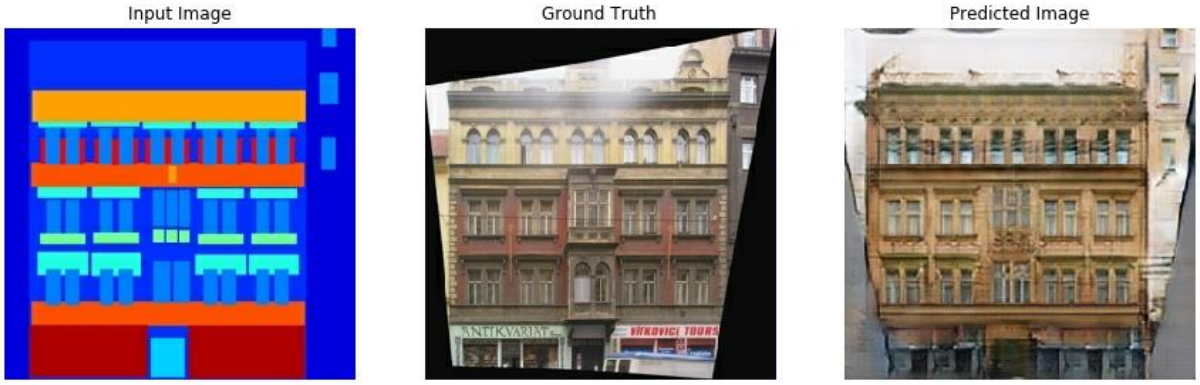
Image applications can go beyond mere entertainment and be extremely useful for those looking to complement a damaged photo or mural. Below are the results of an algorithm implemented in April 2018 by the American company NVIDIA to complete images which have some missing parts. Column (a) is the incomplete photo, (d) is reality and column (c) is the algorithm's prediction. Although the prediction is very realistic, we must nevertheless qualify these impressive results and note that the algorithm was able to learn from many similar examples.

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, A CycleGAN is a GAN that uses two generators and two discriminators. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, 2017, https://junyanz.github.io/CycleGAN



Source: NVIDIA, 2018

The algorithms allow to generate images from other images but are also able to create photos from drawings and even just texts. Below is an example of photographs using a drawing (with the Pix2Pix[2] library):



Source: TensorFlow Pix2Pix

Below is an example of generating images from texts (with StackGAN[3] models):

[2] Pix2Pixis a *Tensorflow* module, an open source library (developed by Google) widely used by the developers' community. *Tensor Flow*, *Pix2Pix*, https://www.tensorflow.org/tutorials/generative/pix2pix

[3] Rajat Garg developed an example of this. Rajat Garg, *Implementing StackGAN using Keras — Text to Photo-Realistic Image Synthesis,* 2019, https://medium.com/@mrgarg.rajat/implementing-stackgan-using-keras-a0a1b381125e

Source: Rajat Garg, 2019

Now let's move on to an example that made an impression in 2018: the first auction of a painting generated by an algorithm. The creative power of these algorithms and their potential to make people dream and create differently has been well understood by the French collective Obvious[4], which auctioned off the first painting generated by an algorithm at Christie's in New York in October 2018.

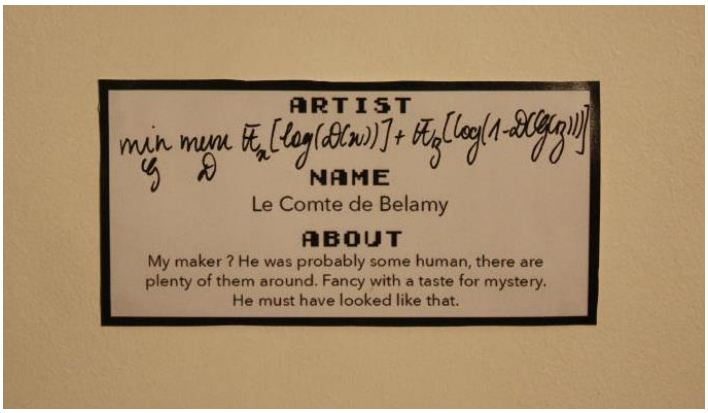
This painting, entitled “The Earl of Belamy” as a tribute to to Ian Goodfellow (“good fellow” can be translated as “bel ami” in French), creator of content generation algorithms, was designed using these algorithms, which had been trained on approximately 15,000 portraits painted between the 15th and the 19th centuries. The collective signed the painting with the mathematical formula corresponding to the content generation algorithms’ learning process. It was sold at Christie's for over 400,000 USD[5].

[4] Obvious, https://obvious-art.com

[5] CHRISTIE’S, *Is artificial intelligence set to become art’s next medium?,* 2018, https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

Source: CHRISTIE'S, 2018 © Obvious

We show below a zoomed-in image of the amusing signature, which does not show the name of the collective but the mathematical formula characteristic of content generation algorithms. We will come back to this example in the part of this report discussing the legal angle, and in particular the intellectual property of the content or works produced by these algorithms.



Source: CHRISTIE'S, 2018 © Obvious

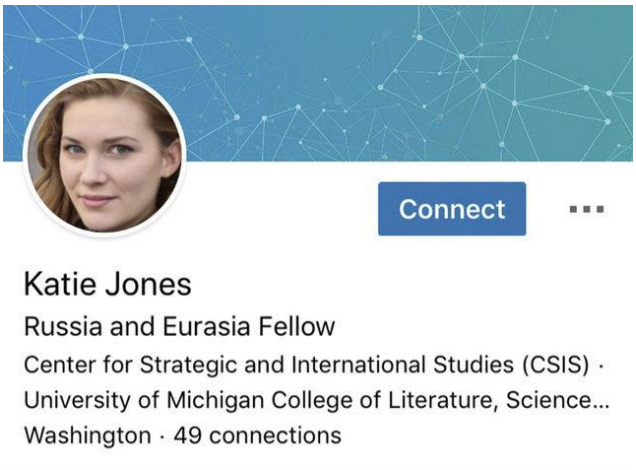
1.1.2. More sinister applications

To raise awareness of the impact of content generation algorithms, an engineer at Uber created the website thispersondoesnotexist.com[6] in February 2017, using an open source algorithm from NVIDIA. The aim was to show how realistic a face generated by this algorithm can be. These faces can unfortunately be used to create fake social media profiles. Indeed, LinkedIn, Facebook and Twitter face great difficulties to prevent the creation of fake profiles. These profiles can be created automatically and can spread enormous amounts of false information on the networks. In particular, they can also be created to influence a positive rating for a restaurant (by companies "buying" a certain number of likes or positive comments), or a negative rating to discredit a competitor, or for political purposes to advance or discredit a candidate.

> ***These profiles can be created automatically and can spread enormous amounts of false information on the networks.***

[6] This Person Does Not Exist, https://thispersondoesnotexist.com/

To show that image generation can be used for suspicious security or political purposes, the American news agency *Associated Press* conducted investigations to detect such cases. It spotted the fake LinkedIn profile below, whose profile photo was generated by an algorithm. According to a newspaper publication: *"In Katie Jones' 50 or so connections, there are a number of prominent figures in Washington, such as a senior adviser to a senator and assistant under-secretaries of state. According to her profile, the young woman had worked for many years at the Center for Strategic and International Studies in Washington,and that's probably why she was able to add so many politicians to her contacts. However, she would never have been able to work there, since she simply does not exist."*[7]



Source: Rémi Lou, 2019

Senior figures were fooled by a simple fake profile, which raises a number of security concerns. Another exemple is the recent deactivation of Facebook and Twitter accounts suspected by the FBI of having been created by the Russian state with the aim of recruiting American freelancers to destabilize the country. The profile photos of many accounts would

[7] Rémi Lou, *Des espions infiltreraient LinkedIn avec des photos de profils créées par l'IA*, 2019, https://www.journaldugeek.com/2019/06/14/des-espions-infiltreraient-linkedin-avec-des-photos-de-profils-creees-par-lia/

have been created synthetically by an algorithm[8]. Researcher Ben Nimmo has detailed this in the image below.

Source: Twitter

On another level, applying content generation to photos of real people complexifies the topic and raises numerous ethical debates.

According to the Huffington Post, the most downloaded app for iPhone and Android on July 17, 2019 was the FaceApp[9], an app developed in 2017 by the Russian company Wireless Lab. Based on a photo of someone's face, it shows what the person will look like when old. For instance, the chef Gordon Ramsey had a try at this "game":

[8]Le Monde, *Facebook et Twitter bloquent une opération d'influence reposant sur un « média indépendant » attribuée à la Russie*, 2020 https://www.lemonde.fr/pixels/article/2020/09/02/facebook-et-twitter-bloquent-une-operation-d-influence-reposant-sur-un-media-independant-attribuee-a-la-russie_6050716_4408996.html

[9]FaceApp uses more simple AI algorithms than GANs. Wikipedia, *FaceApp,* https://en.wikipedia.org/wiki/FaceApp

Source: Gordon Ramsey's account

In addition to aging, FaceApp offers other features, such as changing a man into a woman or adding a smile. In 2017, the app added a filter to make a photograph sexy with a "hot" filter, which prompted strong criticism, the notion of beauty being subjective. Asking an algorithm to define what is sexy is difficult. Here, the "hot" filter systematically whitened the skin of black men:

Source: FaceApp

Accused of racism and following several criticisms such as "*FaceApp is not just bad, it's racist. The hot filter has whitened my skin and changed my nose to make it more European,*" *"FaceApp tells us that white is beautiful. #racism*" or "*Why does beautiful mean white?*", FaceApp removed this filter, and its creator Yaroslav Goncharov has also apologized.

FaceApp was also controversial because of the use of personal data: users had to agree to give up their rights to their photos irrevocably and perpetually:

> You grant FaceApp a perpetual, irrevocable, nonexclusive, royalty-free, worldwide, fully-paid, transferable sub-licensable license to use, reproduce, modify, adapt, publish, translate, create derivative works from, distribute, publicly perform and display your User Content and any name, username or likeness provided in connection with your User Content in all media formats and channels now known or later developed, without compensation to you. When you post or otherwise share User Content on or through our Services, you understand that your User Content and any associated information (such as your [username], location or profile photo) will be visible to the public.
>
> You grant FaceApp consent to use the User Content, regardless of whether it includes an individual's name, likeness, voice or persona, sufficient to indicate the individual's identity. By using the Services, you agree that the User Content may be used for commercial purposes. You further acknowledge that FaceApp's use of the User Content for commercial purposes will not result in any injury to you or to any person you authorized to act on its behalf. You acknowledge that some of the Services are supported by advertising revenue and may display advertisements and promotions, and you hereby agree that FaceApp may place such advertising and promotions on the Services or on, about, or in conjunction with your User Content. The manner, mode and extent of such advertising and promotions are subject to change without specific notice to you. You acknowledge that we may not always identify paid services, sponsored content, or commercial communications as such.

This created a lot of uncertainty as to the use of the data collected, especially since the company is Russian and its headquarters are in St. Petersburg.

Other apps are more directly harmful, such as the DeepNude app, which no longer officially exists today.

Source: DeepNude App

This app, which was officially taken down a few weeks after its online launch on June 23, 2019, but is in fact still downloadable online, allows you to strip a woman naked thanks to AI from a simple photo. Created for the entertainment of users, the app harms women whose photos have been used, undermines their modesty and humiliates them, especially when the fake naked photos are published on the web. The app was officially taken down, claiming it was more successful than expected and not able to cope with the number of requests. DeepNude also raises ethical questions related to the high probability of misuse of the app. Github, a platform where developers share the code of their algorithms for free, observed multiple instances of replicating the DeepNude code and tried to ban them because of their obscene content. However, there are still many versions of the DeepNude code on the platform[10].

1.2. Audio generation

Many years after launching its Photoshop image editing software, the American company Adobe announced that it was working on a voice reconstruction software called VOCO (in reference to Voice Conversion). The speech synthesis tool generates speech using a text and a 20-minute voice sample, thanks to a text-to-speech algorithm. At a demo in November 2016[11], the presenter took a recording of a voice saying *"I kissed my wife and my dogs"* and typed different texts into the software. He was then able to generate different phrases like *"I kissed my wife and my wife"* and eventually *"I kissed Jordan three times"*. We note that the voice is indeed the same, but it seems like the tone of the voice cannot be modulated. This project is still in its testing phase and is not yet marketed.

[10] GihHub, *deepnude*, https://github.com/topics/deepnude

[11] #VOCO https://www.youtube.com/watch?v=I3l4XLZ59iw&feature=youtu.be

A little earlier, in September 2016, Google DeepMind announced the creation of Wavenet[12], its text-to-speech software. It is able to reproduce any voice and any sound, such as the voice of a person with a German accent, for instance, or the sound of a piano. In 2018, Google announced that it was able to replace one voice with another (voice swapping). In 2019, the company announced that it only needs a few minutes of a voice sample to generate speech.

In 2017, audio generation takes a leap forward: the Canadian start-up Lyrebird[13] developed an algorithm to synthesize speech from a one-minute voice sample. It is also able to synthesize voices with different emotions. The start-up foresees a variety of use cases such as virtual personal assistants, audio books, web-enabled devices, voice synthesis for people with disabilities, making films or video games, etc.

On its website, Lyrebird addresses ethical issues. Indeed, an audio recording can serve as evidence in a trial and influence a decision. Lyrebird believes its tool can educate listeners not to consider an audio recording as irrefutable proof[14]:

"By revealing the existence of this technology and making it available, we wish to avert the emergence of such risks. We hope that the public will realize that imitating a voice has become possible and that this should put into perspective in the future the evidential value given to audio recordings."

The fear of such applications materialized when in September 2019, the CEO of a British energy firm was robbed of $220,000 via an audio deepfake. An individual called the CEO using the voice of his chairman to ask for a transfer, the latter was not suspicious and made the requested transfer.

[12] WAVENET: A GENERATIVE MODEL FOR RAW AUDIO, Sep. 2016, https://arxiv.org/pdf/1609.03499.pdf

[13] Descript, *Overdub*, https://www.descript.com/lyrebird-ai?source=lyrebird

[14] TNW, *Eerie tech promises to copy anyone's voice from just 1 minute of audio*, https://thenextweb.com/news/eerie-tech-promises-promises-to-copy-anyones-voice-from-just-1-minute-of-audio#.tnw_XqPkm406

> ***The fear of such applications materialized when in September 2019, the CEO of a British energy firm was robbed of $ 220,000 via an audio* deepfake.**

The technologies used are accessible to everyone, which shows the urgency of setting up effective security systems to deal with voice imitations by algorithms, for example by setting up watermarks (tattoos). This last point will be explained in detail later.

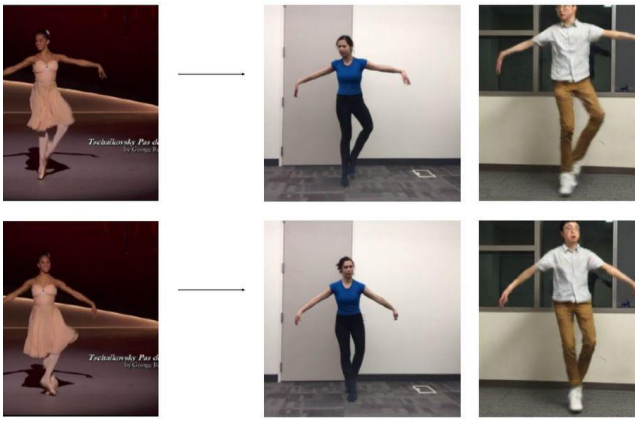
1.3. Video generation

Playful examples

Researchers at UC Berkeley published a research article called *Everybody Dance Now*[15] in August 2018 together with a video posted on Youtube[16]. Their research paper details how they can produce a video depicting a "target"person dancing the same motions as a "source" dancer in a video that was used as the input to their algorithm. Below are the images of the result with a professional prima ballerina as a source and a target of two individuals who cannot dance. The algorithm is capable of transferring the movements of the ballet dancer onto the body of a target person previously filmed performing basic movements.

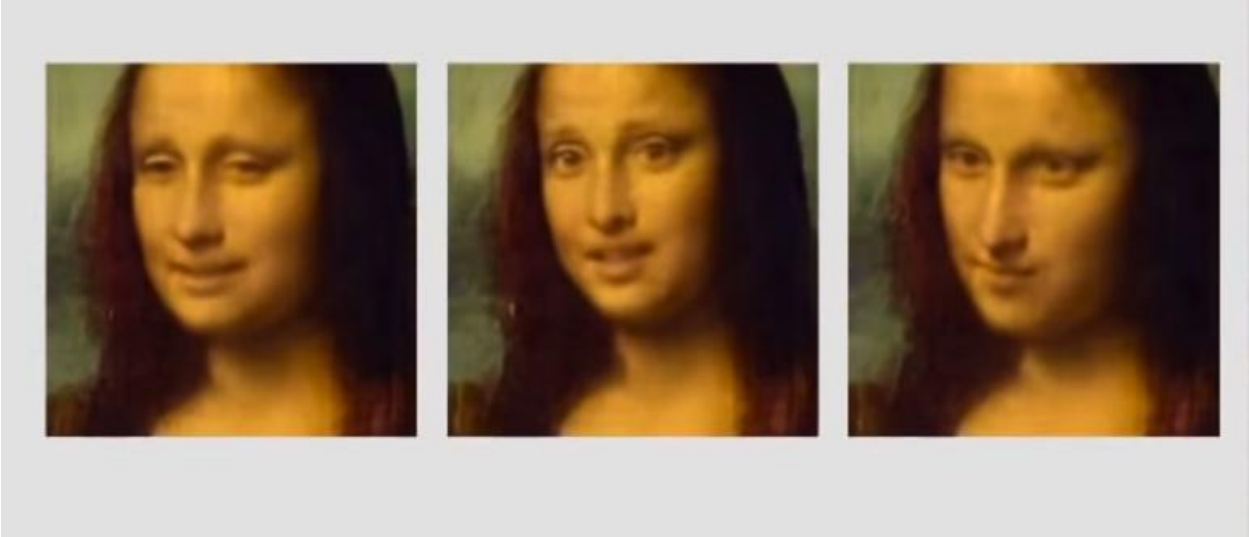
[15] Caroline Chan, Shiry Ginosar, Tinghui Zhou and Alexei A. Efros, *Everybody Dance Now*, 2019, https://arxiv.org/pdf/1808.07371.pdf

[16] Click here to watch the video: https://youtu.be/PCBTZh41Ris



Source: Caroline Chan et al., 2019

In May 2019, Samsung launched a new app[17] that can animate a face from a single photo. The algorithm, based on GANs, allows the animation of a painting like the Mona Lisa below:



Source: Bastien L., 2019

Worrisome uses

Launched in China on August 30, 2019, the ZAO app makes it possible to replace one face with another (*"face-swapping"*), and in particular to replace in a few seconds the face of an actor in a video with another by using a simple photo[18]. Created by the Chinese company

[17] Bastien L, *L'IA de Samsung donne vie à n'importe quelle photo ou portrait*, Le Big Data, 2019, https://www.lebigdata.fr/ia-samsung-photo-portrait

[18] The app is accessible from China exclusively. *Can you believe your eyes? How deepfakes are coming for politics,*

MOMO Inc., initially a dating app publisher, the app has been hugely successful. *"The ability to create deepfake videos has never been so easy as with Zao's interface. Uploading a photo (or a small video of their face, in order to get a better effect) is enough to replace an actor's face in seconds. And with one click to share the inevitably funny result on social networks."*[19]

This brings us to the scandal that affected many people, including many celebrities, whose faces were used in pornographic videos, and subsequently broadcast on platforms such as Reddit and PornHub. Despite the two platforms announcing in February 2018 that they condemn deepfakes, there are still multiple videos of deepfake porn on their websites. Google, for its part, bans *"unintentional synthetic pornographic imagery"* and allows victims of deepfake porn to request that videos in which they appear be blocked. From a legal point of view however, although it is possible to fight someone who steals a private photo and shares it on the Internet, few remedies exist today in the United States to fight those who share created content. At the end of 2018, actress Scarlett Johansson, victim of deepfake porn, including a video with more than 2 million views, believed that fighting against deepfake porn is a lost cause.

From a legal point of view, however (...) few remedies exist today in the United States to fight those who share created content.

https://www.google.com/url?q=https://www.ft.com/content/4bf4277c-f527-11e9-a79c-bc9acae3b654&sa=D&ust=1582556271698000&usg=AFQjCNFswch5I94ZUVyPrdEC2cVsvICDSg

[19]https://www.lemonde.fr/pixels/article/2019/09/03/a-peine-lancee-l-application-de-videos-deepfakes-zao-suscite-des-inquietudes_5505933_4408996.html

We can nevertheless note a breakthrough on the subject in the U.S. state of Virginia which banned deepfake in July 2019. This was followed by a law in California that allows all Californians to file a complaint if they are victim of a pornographic deepfake[20].

Beyond the primary goal of entertaining viewers, the YouTube channel *Facial Deception* published a video of Jim Carrey's face superimposed on the body of American actress Alison Brie. The aim of the video was also to raise awareness about deepfakes and to show that deepfakes are unfortunately increasingly easy to make for those who want to misuse them.

Source: YouTube channel - Facial Deception[21]

Deepfakes are unfortunately increasingly easy to make for those who want to misuse them.

The main purpose for creating the videos of Mark Zuckerberg or Jim Carrey was to raise public awareness about the issue of AI, and more specifically of content generation algorithms. Let's first come back to the video imitating Mark Zuckerberg.

---

[20] Kari Paul, *California makes 'deepfake' videos illegal, but law may be hard to enforce*, 2019, https://www.google.com/url?q=https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce&sa=D&ust=1582556272252000&usg=AFQjCNFDv5BvCoRyuhSxcE0VWZlPDfEm1A

[21] Watch the e video here: https://www.youtube.com/watch?v=b5AWhh6MYCg

It was created by Bill Posters and Daniel Howe in collaboration with CannyAI software[22], initially in the context of an art exhibition entitled Spectre[23]. Bill Posters and Daniel Howe[24] are two British "hacktivists" who describe themselves as research artists. CannyAI is an Israel-based company that produces software for dubbing videos into multiple languages, including the Zuckerberg video shown above. This video was posted by its two authors to catch Facebook in its own trap after the company refused to remove Nancy Pelosi's video. Zuckerberg's video had a big impact in the news, as a result of which Facebook announced, at the end of June 2019, the launch of an internal project to better control deepfakes. In January 2020, Facebook announced a ban on deepfakes.

Positive examples

The advances in the field of video generation are impressive, and fortunately we are seeing more and more examples of applications with more beneficial goals for society.

We note in particular the launch by Huawei of an app offering books with videos in sign language to help children learn sign language and introduce them to the world of books: *"So in 2018, using AI and augmented reality, we created StorySign - the world's first literacy platform for deaf children. Learning to read can be difficult for any child, but for small children who are*

[22] Startup offering the possibility to replace dialogues from videos, and thus to create deepfakes. CannyAI, https://www.cannyai.com/

[23] Bill Posters, *Spectre*, 2019, http://billposters.ch/projects/spectre/

[24] Daniel Howe, *Spectre*, 2019, http://rednoise.org/daniel/detail.html#spectre

severely or profoundly deaf, it can be an overwhelming challenge. We knew our technology could help to open the realm of books to 32 million deaf children around the globe and introduce them to a world of imagination, curiosity and creativity. StorySign is a free mobile app that aims to help deaf children read by translating the text from selected books into sign language. Through the power of AI and augmented reality, StorySign brings these books to life and helps deaf children enjoy storytime as every child should."[25]

1.4. Text generation

1.4.1. Open AI: GPT-2 and GPT-3

OpenAI is a "capped-profit" company (initially a non-profit, but the statutes changed in 2019) founded in 2015, among others, by the entrepreneur Elon Musk. Its objective is to develop and promote humane and positive artificial intelligence for society.

In February 2019, OpenAI announced that it had created a language generation algorithm that was far too powerful to be released to the general public: GPT-2 (Generative Pretrained Transformer 2). It succeeds GPT (June 2018). Indeed, OpenAI researchers believe that GPT-2 is too dangerous to be published in open source. The algorithm makes it possible to reconstruct a text using a few words. This decision to keep their code closed to the public was widely criticized by AI communities. After choosing to only partially release its code, OpenAI was accused of wanting to "buzz" even though the results were not very impressive. Nine months after the partial release, the organization finally released the entire code of its algorithm into open source.

See here two examples of text generated by GPT-2[26]:

[25] https://consumer.huawei.com/uk/campaign/storysign/

[26] Numerous others can be found here: https://openai.com/blog/better-language-models/#sample8

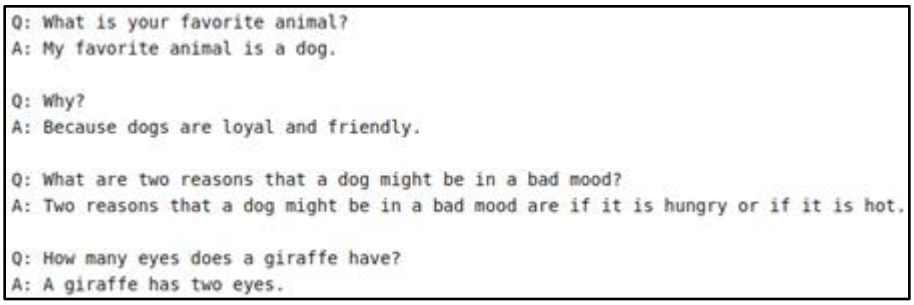
```
Questions Q: What is love?
 A: love
 Q: How do you know if someone has a disease?
 A: look at them
 Q: When does the disease begin?
 A: when they are born
```

```
Links https://www.nytimes.com/2020/01/14/us/
trump-arrested-impeach.html
 Trump Is Arrested and Impeached

 WASHINGTON — President Trump, who has been accused of sexual assault
 by more than a dozen women, was arrested early Friday morning at his
 golf club in Bedminster, N.J., on charges that he had abused the
 power granted to him under an obscure 1799 law.

 The White House said Mr. Trump's arrest stemmed from "multiple
 allegations made against him," but it provided no further details.

 A lawyer for Mr. Trump, Rudolph W. Giuliani, told reporters outside
 the courthouse after the president left court that "the indictment is
 not based upon anything related to campaign activity."

 But several people close to Mr. Trump described him as furious about
 what they called a politically motivated prosecution.

 "The president's base is very upset," one person said.
```

Source: OpenAI

As we see in the example above, the algorithm can be used to create realistic fake articles. It could be misused, in particular to automatically generate opinions or comments on social media networks, spam emails, false scientific or political articles, or even fake news.

About a year later, OpenAI produced a new version of its algorithm: GPT-3. With 175 billion parameters, this successor algorithm of the GPT-2 is able to generate not only text, but also lines of code, a web page, guitar tabs or even missing values in databases. With the aim of revolutionizing the world of AI, GPT-3 constitutes a new form of artificial intelligence.



Source: Twitter[27]

Presented with random questions, GPT-3 is able to answer in a fairly logical and precise manner. But how does this mechanism translate?

Based on deep learning, GPT-3 seeks to understand the structure of the data and to find recurring patterns (i.e. simplified models of a more complex structure). Without any human intervention, and using the volume and frequency of the occurrence , it is able to make predictions based on the text supplied as input. This algorithm was trained on a huge compilation of texts in order to exploit the patterns. The nodes of the neural network subsequently store these regularities as weighted connections[28].

To illustrate very simply how it works, let's take the example of the sentence “I wanted to make an omelet, so I went to the fridge and took out some ___.” The blank at the end can theoretically be filled with any word. But the word "eggs" probably scores high enough to fill that space in most normal texts. A higher probability than, say, the word "elephants". The probability of the presence of the word “eggs” in the text is therefore higher than the probability of the presence of the word “elephants”.

[27] watchable here: https://twitter.com/lacker/status/1280276200582025216

[28] James Vincent, "OpenAI’s latest breakthrough is astonishingly powerful, but still fighting its flaws”, The Verge, 2020, https://www.theverge.com/

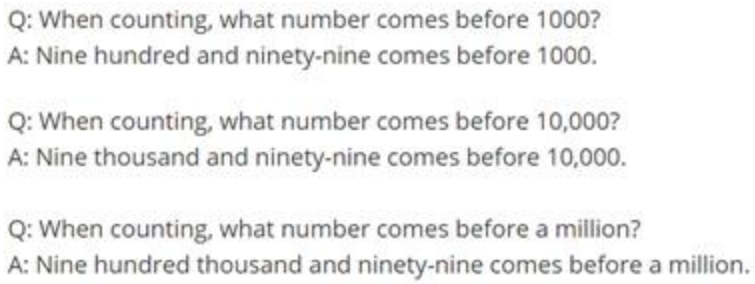
> ***The algorithm can be used to (...) automatically generate opinions or comments on social media networks, spam emails, false scientific or political articles, or even fake news.***

Thanks to its ability to automate tasks, GPT-3 will be able to become an important asset in the business world, and especially in high-tech or data companies. GPT-3 is notably capable of writing HTML code[29], as shown in the following interface:

Source: Zoltan Szogyenyi's blog in the ThemesBerg, 2020

Despite these various characteristics, this does not prevent the existence of several drawbacks. At the enterprise level, the essential complication is that GPT-3 requires a lot of data and computing power. Its compilation costs are significant. The generated text may look impressive, but the algorithm makes more mistakes when it comes to more complex requests.

[29]Zoltan Szogyenyi, "We built an OpenAI powered Tailwind CSS code generator using GPT-3", ThemesBerg, 2020, https://themesberg.com/



Source: Mind Matters, 2020[30]

The example above makes clear that above a certain number of digits, GPT-3 does not perform well in number sequences when relying only on learned data. Hence, it turns out that GPT-3 can make mistakes that the human brain wouldn't.

1.4.2. Generation of musical scores

In music, we can also observe applications in the generation of music. Recurrent networks with temporal memory have proven their worth: creating a jazz piece has become a classic exercise in coding with the LSTM algorithm, thanks to a lecture by Andrew Ng from Stanford University[31]. In 2017 the MidiNet[32] model, using GANs with CNNs (Convolutional Neural Networks),[33] was dedicated to the generation of music.

Similarly, in 2017 Huawei broadcasted a completed version of Schubert's Eighth Symphony, which the composer began in 1822 and never completed. *"Using the power of AI, Huawei has taught its Mate 20 Pro smartphone to compose the third and fourth movements of Schubert's famously 'Unfinished Symphony'. This was achieved by harnessing the power and intelligence of*

[30] https://mindmatters.ai/2020/07/gpt-3-is-mindblowing-if-you-dont-question-it-too-closely/

[31] deeplearning.a, *Recurrent Neural Networks - Why sequence models?*, https://cs230.stanford.edu/files/C5M1.pdf

[32] Li-Chia Yang, Szu-Yu Chou, Yi-Hsuan Yang, *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation*, 2017, https://arxiv.org/abs/1703.10847

[33] In machine learning, a convolutional neural network or convolutional neural network (CNN or ConvNet for Convolutional Neural Networks) is a type of acyclic (feed-forward) artificial neural network, in which the connection pattern between neurons is inspired by the visual cortex of animals.

the Mate 20 Pro, utilising its in-built advanced technology which has been specifically designed to complete AI-based tasks.

The Mate 20 Pro smartphone listened to the first two movements of Schubert's Symphony No. 8, analysed the key musical elements that make it so incredible, then generated the melody for the missing third and fourth movement from its analysis."[34] Even if the critics are not all convinced that they can hear Schubert's style in the notes generated, the fact remains that it was a successful marketing ploy on Huawei's part to highlight its advances in IA.

1.5. Data generation

Content generation algorithms are proving increasingly useful to fill in missing data. Indeed, thanks to their ability to understand and generate data sets these algorithms can be used to create realistic data. This is useful for many industries, for instance for insurance companies to calculate the risk of natural disasters, the risk of pandemics or any other type of event for which we lack historical data.

[34]https://consumer.huawei.com/uk/campaign/unfinishedsymphony/

2. Les enjeux éthiques liés aux algorithmes de génération de contenu

2.1. Background

"Every jump of technical progress leaves the relative intellectual development of the masses a step behind, and thus causes a fall in the political-maturity thermometer. It takes sometimes tens of years, sometimes generations, for a people's level of understanding gradually to adapt itself to the changed state of affairs [...]"

Arthur Koestler, Darkness at Noon

Content generation algorithms pose many questions regarding individual freedoms, respect for others and more generally on human co-existence in a democratic society.

The development and use of programs that can generate deepfakes content should lead us to ask the question (again): should everything that is technically possible be undertaken? What needs to be regulated?

2.1.1. History

The dissemination of false information or forged documents is nothing new[35]. Pharaoh Ramses II pretended in inscriptions that the Egyptians had defeated the Hittites at the battle of Kadesh when in reality there had been no victor. One of the first forged documents used to justify an action is the *Donation of Constantine*, a letter by which the Emperor Constantine I in the year 315 purportedly transferred all his power over the West to Pope Sylvester. The Donation was first mentioned in 979, and the forgery would not be proven until 500 years later.

The Protocols of the Elders of Zion, texts published in Russia at the beginning of the 20th century, purport to reveal an alleged plot by the Jewish and Freemason communities to

[35] Wikipedia, *Fake news*, https://en.wikipedia.org/wiki/Fake_news#Ancient

dominate the world. They were invoked by Hitler, notably in *Mein Kampf*, to justify the theory of the Jewish conspiracy. They are still disseminated in antisemitic circles.

As for photographic manipulation, we can mention the photos of Soviet dignitaries which were later "photoshopped" to remove any disgraced leaders, as was this photo from the 1930s[36]:

Photos of Stalin with and without Nikola Yezhov, Stalin's former right-hand man.

Source: Wikimedia, 1930

A few dozen years earlier and in another context, we can mention this photo of General Ulysses Grant from the American Civil War, which is actually a composite photo made up of three different images[37].

[36] Wikimedia, *URSS, 1930,* https://curiummag.com/wp-content/uploads/2016/08/staline-1024x462.png

[37] Wikpedia, *Photo manipulation*, https://en.wikipedia.org/wiki/Photo_manipulation

Photo of General Ulysses Grant

Source: Wikimedia, 1930

The manipulation of texts and images for political objectives is therefore nothing new.

2.1.2. Algorithms to generate content

Content generation algorithms are used to create content that can be playful, funny, and are of course not only applied to create deepfakes.

In addition to creating images, sounds, and videos, these algorithms can be used to complete a set (picture, sound...) even with significant data gaps. This can be very useful for improving models that learn from small datasets, for example predicting the impact of an epidemic, or assessing a cybersecurity risk to better estimate the cybersecurity insurance premium.

In the field of medical imaging, where the number of cases available is often limited, they allow the creation of new images that can be used to train deep learning models[38].

In the field of art, design and fashion, GANs allow the creation of new designs as well as personalized and unique designs. They can also find many applications in the field of video games[39].

GANs allow for the creation of useful makeup environments that can be very useful for training machines or even humans for new tasks.

The CEO of Ekimetrics, Jean-Baptiste Bouzige, considers that GANs would help to reduce construction costs by modeling buildings well before their formal design, for example to model high energy efficiency or to overcome other complex technical constraints.

2.1.3. The democratization of content production

Another critical difference with the past is that every individual can be at the same time creator, consumer, and sharer of content and, of course, citizen.

Let's look at journalism, for example. Although it is difficult for journalists to be impartial and to communicate information without bias, the profession has gradually become organized and developed a code of conduct for the treatment of information, for example with the *Munich Charter* in Europe[40]. In particular, journalists have a moral obligation to verify the facts mentioned. In the current environment, not all Internet content producers adhere to this code of ethics, and much fake news circulates on the web.

---

[38] Maayan Frid-Adar et al., "Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification", Department of Biomedical Engineering, Tel Aviv University, 2018

[39] Wang-Cheng Kang et al., "Visually-Aware Fashion Recommendation and Design with Generative Image Models" , 2017

[40] Charte de Munich: Déclaration des droits et des devoirs des journalistes (1971), and Charte du Syndicat National des journalistes (1918, revised in 1938)

It is impossible to force platforms to ensure the authenticity of all the content they produce because, as we have seen, it is sometimes impossible to separate truth from fiction. Besides, false content is not necessarily malicious: for example, a cartoon does not necessarily aim to harm. Nevertheless, a simple measure could better regulate false content: an obligation for authors to declare altered or modified content (beyond simple filter and cosmetic corrections). This would make the authors of false content accountable and also better protect platform users.

2.1.4. Social networks: accelerated dissemination, truth bias by highlighting false content

There is no denying that today's content can be spread almost instantaneously, and the number of people that can be reached at once has exploded. According to a study published by the *International Business Times*, 60% of American millennials use social media platforms like Twitter and Facebook as their source of news rather than traditional media. Therefore, they are much more exposed to viral content (fake or true), which can be promoted by platforms using models to increase sharing and the number of likes.

Indeed, platforms have a self-interest in promoting the distribution (or virality) of videos attracting the viewers (and controversy), which generate multiple sharing between users. The more viewers, the higher the profits.

We note for example the phenomenon of memes[41] on the Internet: exciting and fun, they spread quickly and turn out to be an excellent communication tool, which many advertisers have understood, but also many people seeking to defend their political cause. The process itself is not that different from the cartoons that have been used for hundreds of years. The

[41] A meme is a "concept (text, image, video) massively picked up, declined and diverted on the Internet in an often parodic way, which spreads very quickly, thus creating a buzz." Larousse, https://www.larousse.fr/dictionnaires/francais/m%C3%A8me/10910896

difference lies not so much in the appeal of scandal, the use of laughter or parody to discredit or convey a message, but rather, on one hand, in the increased speed in the diffusion which prevents any effective reaction, and in the scandal bias, which means that such content attracts more than a less exciting content that tries to inform, on the other.

When a deepfake of President Obama is posted online, even though it is known to be fake, and many confirm it to be fake, it is very difficult to limit its impact.

Former French minister Nathalie Kosciusko-Morizet[42], who met with us, points to the existence of two major risks: the speed of dissemination of information and virality:

"The first risk is the speed with which information is disseminated. Precisely, the gap between the speed of dissemination of lies and that of denials. There are documented studies on this topic. The second risk is that of virality. With digital comes the decline of intermediaries, who served as moderators. This is the premium on emotion, on radicalism, on excess. The appetite for scandal and controversy is in full swing, especially as this natural propensity of the human soul is fostered by algorithms that push the most controversial opinions to the top of the pages. The problem is compounded by the persistence of presence on the Internet. Even when denied or outdated, the information, true or fake, remains online."

A study published in 2018 confirms her words, saying *"fake news spread six times faster on Twitter than information from fact-checking."*[43]

2.2. What is "ethics"?

"Ethics" is one of those ready-to-use concepts that are handy for shifting a debate to another level, without sometimes knowing what the word means. Without going into long

[42] Nathalie Kosciusko-Morizet, *Cyberespace et déstabilisation de la démocratie*, 2020 https://www.linkedin.com/pulse/cyberespace-et-d%C3%A9stabilisation-de-la-d%C3%A9mocratie-kosciusko-morizet/

[43] Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online*. Science, 359(6380), 1146-1151.

explanations, it seems important to clarify its meaning and remember how it was able to find a concrete and practical meaning in the case of medicine.

In1990, the philosopher Paul Ricœur defined ethics as *"the aim of a life accomplished under the sign of actions considered as good."*[44] The philosopher Dominique Lecourt defines ethics as *"a reflection, a questioning which deals with dilemmas. Faced with complex situations, there is a choice to be made by an individual between several answers which are all unsatisfactory."*[45] Ethics is therefore an individual reflection on what seems good and must be chosen, even if no external constraint (legal or moral) imposes this choice.

This notion of "good" in a democratic society is not easy to define, but we feel that it can easily enter into contradiction with business ethics, which sets as a limit to do whatever is legally permitted to optimize economic performance. Thus, platforms can find economic justification for letting a deepfake circulate even though executives may individually believe that it is not ethically correct.

2.3. First steps for introducing ethics to the net

In recent years, several initiatives by both Internet actors and public authorities have emerged aimed at drawing up ethics charters.

On the business side, we can cite in particular *Partnership on AI*[46], which includes the giants, Amazon, Facebook, Google, Microsoft, IBM, Apple and also, since 2018, the Chinese giant Baidu. This coalition, founded in 2016, aims to educate the public on AI and define good practices in artificial intelligence.

Other actors have also proposed ethical principles to follow when designing algorithms.

---

[44] Paul Ricœur, *Éthique et morale*, 1990

[45] Simon Sutour and Jean-Louis Lorrain on behalf of the European Commission, Report to the Senate, 2013

[46] The Partnership on AI, https://www.partnershiponai.org/

European Commission

The European Commission recently published a white paper on artificial intelligence which includes among other things the conclusions of its report of June 2018. According to it, *"While AI can do much good, including by making products and processes safer, it can also harm. This harm might be both material (safety and health of individuals, including loss of life, property damage) and immaterial (loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination for instance in access to employment), and can relate to a wide variety of risks."* [47]

The ethical principles put forward by the European Commission are:

1. Human agency and oversight,
2. Technical robustness and safety,
3. Privacy and data governance,
4. Transparency,
5. Diversity, non-discrimination, and fairness,
6. Societal and environmental wellbeing, and
7. Accountability.

[47] European Commission, *On Artificial Intelligence - A European approach to excellence and trust*, 2020, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Source: European Commission, *Ethics Guidelines for Trustworthy AI*

The EU is also proposing to ensure that AI ethics training is systematically provided to a new generation of experts and to carry out a Trustworthy AI Assessment at the development, deployment, or use of AI systems, tailoring it to the specific use case of the system.

OECD

The OECD has also taken part in the debate around artificial intelligence by publishing a list of recommendations and principles which appear to be very similar[48] to those set out by the European Commission:

- **Inclusive growth, sustainable development, and well-being:**

  Stakeholders should proactively engage in responsible management of reliable AI in the pursuit of results that benefit people and the planet, such as increasing human capacities and enhancing creativity, promoting inclusion of under-represented

[48] OECD Principles on AI, https://www.oecd.org/going-digital/ai/principles/

populations, reduction of economic, social, gender and other inequalities, and protection of natural environments, thereby boosting inclusive growth, sustainable development, and well-being.

- **Human-centered values and equity:**

  AI actors should respect the rule of law, human rights, and democratic values throughout the lifecycle of the AI system. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, equity, social justice, and internationally recognized labor rights.

  To this end, AI actors should implement mechanisms and guarantees, such as the human determination capacity, adapted to the context and in line with the state of the art.

- **Transparency and explainability:**

  AI actors should commit to ensuring transparency and responsible disclosure of AI systems. To this end, they should provide meaningful, context-appropriate, and state-of-the-art information:

    - To foster a general understanding of AI systems.
    - To raise awareness among stakeholders of their interactions with AI systems, including in the workplace.
    - Enable people affected by an AI system to understand the outcome, etc.
    - Enable those affected by an AI system to challenge its results based on clear and easy-to-understand information about the factors and logic that served as the basis for the prediction, recommendation, or decision.

- **Robustness, security, and safety:**

AI systems must be robust, secure, and secure throughout their lifecycle so that under normal, expected, or abusive use, or any other adverse conditions, they perform appropriately and not do not present an unreasonable security risk.

To this end, AI actors should ensure traceability, including for data sets, processes, and decisions made during the lifecycle of the AI system, to enable analysis of AI system results, and survey responses, appropriate to the context and consistent with the state of the art.

AI actors should, depending on their roles, context, and ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on an ongoing basis to address risks associated with AI systems, including privacy, digital security, safety, and bias.

- **Accountability:**

  AI actors should be accountable for the proper functioning of AI systems and adherence to the above principles, depending on their roles, context, and following the state of the art.

The CNIL Report

In 2017, the French authority for personal data control and liberties, the Commission nationale de l'informatique et des libertés (CNIL), published a report on the ethical issues of algorithms and artificial intelligence[49] that defined two founding principles "for artificial intelligence at the service of citizens": a principle of loyalty and a principle of vigilance (see below).

[49] CNIL, *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence*, 2017,
https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

Towards the affirmation of two founding principles: loyalty and vigilance

A principle of *loyalty* applied to all algorithms and integrating the collective, and not just personal, impacts of the latter. Any algorithm, whether or not it processes personal data, must be loyal to its users, not only as consumers but also as citizens, even towards communities or large collective interests whose existence could be directly affected. **The interests of users must come first**. For example, such a principle could be applied to the potential impact of social networks on the structure of public debate in our democracies (segmentation of the body politic by the targeting of information) or to that of algorithms of predictive policing for entire communities or neighborhoods.

A principle of *vigilance/reflexivity*: this involves organizing a form of regular, methodical, and deliberative questioning with regard to these moving objects. This principle constitutes a direct response to the requirements imposed by these technological objects due to their unpredictable nature (inherent in machine learning), the highly compartmentalized nature of the algorithmic chains within which they are inserted, and, finally, the excessive trust in which they often give rise to. It is all the links in the algorithmic chain (designers, companies, citizens) that must be mobilized to give substance to this principle, by means of concrete procedures (for example, ethics committees ensuring a systematic and continuous dialogue between the different stakeholders).

European ethical charter for the use of artificial intelligence in judicial systems

In December 2018, a European Commission working group "for the efficiency of justice" published a European ethical Charter on the use of AI in legal systems[50]. This charter highlights five key principles to fight against abuses of artificial intelligence and to set an ethical framework for the use of artificial intelligence:

- **Principle of respect for fundamental rights:** aimed at ensuring the design and implementation of artificial intelligence tools and services that are compatible with fundamental rights.
- **Principle of non-discrimination:** aimed specifically at preventing the creation or reinforcement of discrimination between individuals or groups of individuals.
- **Principle of quality and security:** aiming, with regard to the processing of judicial decisions and judicial data, to use certified sources and intangible data with models designed in a multidisciplinary manner, in a secure technological environment.
- **Principle of transparency, neutrality, and intellectual integrity:** aimed at making data processing methodologies accessible and understandable and at authorizing external audits.
- **Principle of "under user control":** aiming to ban a prescriptive approach and allow the user to be an enlightened actor and master of his choices.

It concludes its report by explaining that technological developments must be framed by "cyberethics" which *"must be accompanied by large-scale training of stakeholders, from algorithm designers and legal tech companies to their users. New transdisciplinary humanities should be made available to all so that AI becomes a vector of positive development for humankind."*

[50]European Commission for Justice Efficiency (CEPEJ), European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment, 2018 https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

2.4. What are the online platforms doing about this??

Online platforms, especially social networks, play a major role in the dissemination of information, and by extension in deepfake content. Whether in France[51], the United Kingdom[52], or the United States, studies have shown that people tend to get their news online and from social media rather than via traditional channels such as radio or the newspapers. We have already mentioned the study published by the *International Business Times*. Among French under-34s, social networks are even the first source of information, ahead of television. However, based on their business model and by personalizing the content they disseminate, social media help to polarize opinions. Indeed, polarized individuals and groups interested in false or manipulated information - whether they know it or not - tend to be exposed to more content of the same type. In doing so, social media help to reinforce beliefs and to polarize further.

However, if the Internet giants are aware of the ethical issues and risks posed by fake news, their response to this problem does not seem unanimous. The universal adoption of an ethical charter does not appear to be on the agenda as each platform takes a different stance and therefore takes its own measures.

2.4.1. Facebook

In May 2019, when Facebook circulated a distorted video of Nancy Pelosi, a Democrat and Speaker of the United States House of Representatives, in which she appeared to be drunk, the social network came under fire from critics. Going viral, the video was seen by more than

[51] Noémie Bonnin, *Les réseaux sociaux première source d'info en ligne chez les personnes sensibles aux théories du complot, 2019,* https://www.francetvinfo.fr/internet/reseaux-sociaux/info-franceinfo-les-reseaux-sociaux-premiere-source-d-info-en-ligne-chez-les-personnes-sensibles-aux-theories-du-complot_3191963.html

[52] *CCFI, UK: la moitié de la population s'informe par les réseaux sociaux, 2019,* https://www.ccfi.asso.fr/uk-la-moitie-de-la-population-sinforme-par-les-reseaux-sociaux/

2 million people in two days and allowed Donald Trump's team to discredit her. In response, Facebook announced it considers itself neither a guarantor nor a judge of the veracity of videos posted on its site, that it advocates freedom of expression, and confirmed it will not remove deepfakes. In June 2019, Mark Zuckerberg, founder and CEO of Facebook, was himself the victim of such a video in which he appears to announce that he is ready to use billions of stolen personal data to control the future and that he owes everything to "Spectre".

However, in September 2019, Facebook, in partnership with Amazon and Microsoft, launched the Deepfake Detection Challenge (DFDC) to develop algorithms capable of identifying deepfake videos and images.[53]

Since then, Facebook has implemented a policy against deepfakes. As of January 2020, a document will now be deleted from the platform:[54]

- if *"it has been edited or synthesized – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say", and*
- if *"it is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic."*

This policy "*does not extend to content that is parody or satire, or video that has been edited solely to omit or change the order of words*." Caricatures and deepfakes for parodies are allowed.

However, Facebook is not asking authors for more transparency to warn users about the use of content generation algorithms.

[53] Facebook AI, *Deepfake Detection Challenge Dataset*, 2020, https://deepfakedetectionchallenge.ai/

[54] Facebook, *Enforcing Against Manipulated Media*, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

2.4.2. Twitter

Twitter announced a new anti-deepfake policy in February.[55] The company will report accounts that share deepfake documents, notify user accounts that wish to share these documents ("retweet"), and provide the reasons why the document may be a deepfake.

Source: Twitter

For a start, Twitter mentions that to identify fake content, the following factors will be taken into account:

- If the content has been significantly altered in a way that fundamentally alters its composition, sequence, pace, or framing;

[55] Yoel Roth & Ashita Achuthan, *Building rules in public: Our approach to synthetic & manipulated media*, 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

- Any visual or auditory information (such as new video images, dubbed audio data, or edited subtitles) that has been added or deleted;
- Whether media representing a real person has been fabricated or simulated.

Where content meets the conditions listed above,Twitter may take any of the measures listed in the table below to limit the scope and negative impact of bogus content. Where appropriate, Twitter will be able to apply warning labels to posts such as they did with Donald Trump's tweets denouncing absentee voting.[56]

The following measures will be implemented where and when required:

- Apply a label to the tweet.
- Show users a warning before retweeting or liking a tweet.
- Reduce the visibility of the tweet on the network and/or prevent it from being recommended.
- Provide additional explanation or clarification, where applicable, such as a home page with more context or a link to a certified source.

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

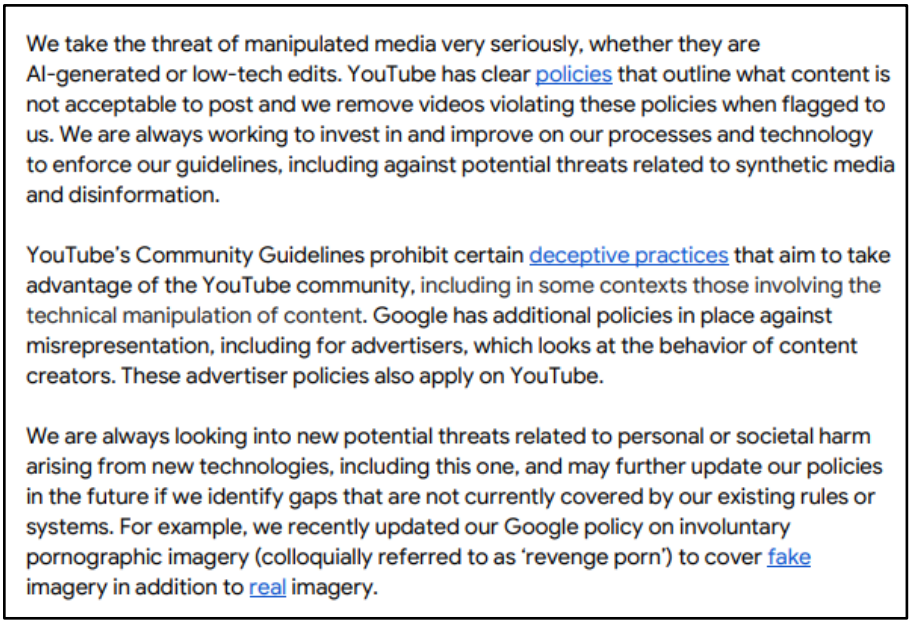
Source: Twitter

2.4.3. Google & YouTube

[56]https://www.franceinter.fr/monde/pour-la-premiere-fois-un-message-de-donald-trump-signale-comme-mensonger-par-twitter

As the largest video platform in the world, YouTube necessarily plays a major role in the distribution of deepfake videos. YouTube was the first major network to remove the doctored video of Nancy Pelosi. YouTube officials say they want to *"become a more reliable source for election-related news and information as well as an open platform for healthy political discourse"*[57]. The network further announced that it was going to remove *"videos that are technically manipulated or doctored in a way that misleads users [...] and may pose a serious risk"*. A position that is similar to that of other major platforms.

More broadly, Google (which owns YouTube) has a policy of not allowing deceptive practices as they claim in a response to the House Permanent Select Committee on Intelligence (a standing committee of the U.S. Congress under the responsibility of the House of Representatives)[58]. In addition, they collaborate with the FBI to identify malicious users and deactivate their accounts.



We take the threat of manipulated media very seriously, whether they are AI-generated or low-tech edits. YouTube has clear policies that outline what content is not acceptable to post and we remove videos violating these policies when flagged to us. We are always working to invest in and improve on our processes and technology to enforce our guidelines, including against potential threats related to synthetic media and disinformation.

YouTube's Community Guidelines prohibit certain deceptive practices that aim to take advantage of the YouTube community, including in some contexts those involving the technical manipulation of content. Google has additional policies in place against misrepresentation, including for advertisers, which looks at the behavior of content creators. These advertiser policies also apply on YouTube.

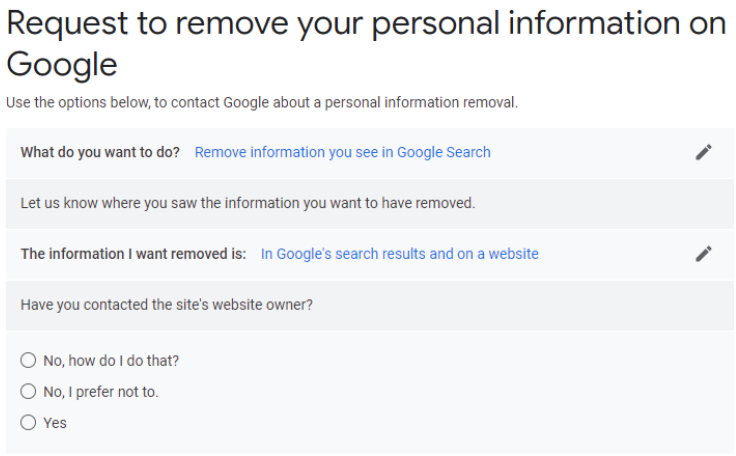
We are always looking into new potential threats related to personal or societal harm arising from new technologies, including this one, and may further update our policies in the future if we identify gaps that are not currently covered by our existing rules or systems. For example, we recently updated our Google policy on involuntary pornographic imagery (colloquially referred to as 'revenge porn') to cover fake imagery in addition to real imagery.

Source: extract of the response by Google to Adam B. Schiff

[57] Youtube, *How Youtube supports elections*, 2020, https://blog.youtube/news-and-events/how-youtube-supports-elections/

[58] KaranBhatia, 2020, https://schiff.house.gov/imo/media/doc/KB%20Schiff%20response%207-2019%20(1).pdf

Google also offers the possibility for users to remove deepfake images or videos from its search results (but not from the Internet) by filling out an online form[59]. However, it is possible that the volume of these documents is too large to make this method effective, especially when it comes to public figures or celebrities.



Screenshot of the form by Google to request withdrawal of content

Google also contributes to the *technological* fight against deepfakes. The company is already distributing an experimental tool to detect deepfake photos (but not yet for videos)[60]. And in 2019, Google released a database of 3,000 deepfake videos (which Google will expand over time) to encourage and help research when it comes to detecting this type of content. In particular, these videos make it possible to train deepfake detection algorithms.

2.4.4. Microsoft

[59] Google, *Demande de suppression de vos informations personnelles de Google*, https://support.google.com/websearch/troubleshooter/9685456#ts=2889054%2C2889099

[60] https://www.technologyreview.com/2020/02/05/349126/google-ai-deepfakes-manipulated-images-jigsaw-assembler/

In order to regulate artificial intelligence systems and to fight against deepfake content, Microsoft has adopted the following ethical principles:

- **Fairness**: AI systems should treat everyone fairly.
- **Inclusiveness**: AI systems should empower everyone and involve people.
- **Reliability and security**: AI systems must operate reliably and securely.
- **Transparency**: AI systems must be understandable.
- **Confidentiality and security**: AI systems must be secure and respect privacy.
- **Accountability**: AI systems should have algorithmic accountability.

Microsoft President Brad Smith has called on governments to regulate artificial intelligence, especially facial recognition, which identifies individuals in public places[61]. In a 2018 report he wrote this:

Today, there are some people who might say that ethical principles and best practices are all that is needed as we move forward. They suggest that technology innovation doesn't really need the help of regulators, legislators and lawyers.

While they make some important points, we believe this view is unrealistic and even misguided. AI will be like every technology that has preceded it. It will confer enormous benefits on society. But inevitably, some people will use it to cause harm. Just as the advent of the postal service led criminals to invent mail fraud and the telegraph was followed by wire fraud, the years since 1998 have seen both the adoption of the internet as a tool for progress and the rise of the internet as a new arena for fraud, practiced in increasingly creative and disturbing ways on a global basis.

We must assume that by 2038, we'll grapple with the issues that arise when criminal enterprises and others use AI in ways that are objectionable and even harmful. And undoubtedly other important questions will need to be addressed regarding societally acceptable uses for AI. It will be impossible to address these issues effectively without a new generation of laws. So, while we can't afford to stifle AI technology by adopting laws before we understand the issues that lie ahead of us, neither can we make the mistake of doing nothing now and waiting for two decades before getting started. We need to strike a balance.

Source: Microsoft, *The Future Computed*, 2018 [62]

[61] Nitasha Tiku, *Microsoft Wants to Stop AI's 'Race to the Bottom'*, 2018, https://www.wired.com/story/microsoft-wants-stop-ai-facial-recognition-bottom/
[62] Brad Smith & Harry Shum, *Artificial Intelligence and its role in society*, 2018, https://news.microsoft.com/uploads/2018/02/The-Future-Computed_2.8.18.pdf

In light of the U.S. presidential elections in November 2020, Microsoft has deployed new tools to fight against deepfakes[63]. A tool called Video Authenticator allows media and campaign teams to recognize deepfakes. Another, Azure Information Protection, allows documents to be tracked using digital markers and certificate-based authentication systems. As these technologies are new, it is difficult to know if they will have a significant impact. For example, it is not certain that many users will voluntarily install the Internet browsing extensions suggested by Microsoft.

In parallel with the launch of these new tools, Microsoft has announced that it is joining several initiatives:

- A partnership with the AI Foundation, to make Video Authenticator available to organizations through its Reality Defender 2020 program. The challenge is to better guide organizations within the limits and ethical considerations relating to deepfake detection technologies.
- A partnership with a consortium of media companies under the name Project Origin will test Microsoft's authentication technology. An initiative of several publishers and social media companies, the Trusted News Initiative, has also agreed to work with Microsoft to test its technology.
- A partnership with the University of Washington, the deepfake-detection platform Sensity, and *USA Today* with the aim to expand media literacy and help individuals better identify disinformation and manage the risks associated with both deepfakes and cheap fakes[64]. As part of this partnership, a public service announcement campaign encourages the public to take a "reflective pause". The campaign further urges people to verify that information comes from reputable news organizations before sharing it on social media, especially during an election period.

[63] https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/

[64] Le monde Informatique, *Microsoft dévoile des outils pour lutter contre les deepfakes*, 2020

Through these initiatives and partnerships, Microsoft is promoting education to help people sort misinformation from real facts.

2.4.5. Platforms based in Asia

This attitude towards deepfake content is far from universal among platforms, however. For example, Snapchat and TikTok[65] incorporate deepfake technologies into their respective platforms, in particular to offer filters to modify their videos. It is in this sense that Snapchat acquired AI Factory[66], a company specializing in the creation of deepfakes, based partly in Ukraine. However, if these platforms adopt deepfake technologies, they are putting in place arrangements so that their users can easily understand that these are not real videos. It is also interesting to note that the Chinese version of TikTok, Douyin, implements the more stringent Chinese standards in this area (e.g. labeling of deepfakes, real user identities, data protection). However, this is not the case with TikTok[67]. This tends to illustrate a lack of voluntarism among some platforms vis-à-vis the protection against the misuse of deepfakes. Like Bytedance (the parent company of TikTok and Bayin), platforms can choose to adapt to the legal standards of different countries and impose barriers only if they are forced to do so.

[65] Michael Nuñez, *Snapchat and TikTok Embrace 'Deepfake' Video Technology Even As Facebook Shuns It*, 2020, https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/#7782d70042c0

[66] Ingrid Lunden, *Snapchat quietly acquired AI Factory, the company behind its new Cameos feature, for $166M*, 2020, https://techcrunch.com/2020/01/03/snapchat-quietly-acquired-ai-factory-the-company-behind-its-new-cameos-feature-for-166m/

[67] Josh Constine, ByteDance & TikTok have secretly built a deepfakes maker, 2020 https://techcrunch.com/2020/01/03/tiktok-deepfakes-face-swap/?guccounter=1

> ***This tends to illustrate a lack of voluntarism on some platforms vis-à-vis the protection against the misuse of deepfakes.***

For others, deepfake technology is the basis of their service. This is for example the case of Zao, a Chinese app released in August 2019 that very quickly found itself among the most downloaded apps[68]. The app allows, thanks to deepfake algorithms, to embed one's face in a video sequence, especially in movies. However, the app has raised a lot of criticism because of its terms and conditions which stipulated that Zao retained a permanent and irrevocable right to the photos provided by users. These image rights issues raise questions in an Asian society that increasingly uses facial features to provide various services, such as payment by facial recognition. These technologies, still not widely democratized in the West, could quickly develop there. This is why the regulations of Western countries must anticipate the potential emergence of these new services.

Tencent, the company behind the app WeChat, sees many opportunities for deepfakes, which it calls deep synthesis[69]. Its research arm Tencent AI Lab explores the application of the technology in content creation, marketing, social media, entertainment, e-commerce, and even communication. For example, researchers at Tencent AI Lab recently published an article showing that it is possible to create a singing voice from samples of a normal voice[70]. It is therefore clear that Chinese companies will continue to invest in this area.

---

[68] Pierrine Sigoret, *Ce qu'il faut savoir sur Zao, l'application au succès fulgurant qui vous transforme en Leonardo DiCaprio*, 2019, https://www.numerama.com/tech/545092-ce-quil-faut-savoir-sur-zao-lapplication-au-succes-fulgurant-qui-vous-transforme-en-leonardo-dicaprio.html

[69] https://chinai.substack.com/p/chinai-104-tencent-2020-ai-white

[70] https://www.arxiv-vanity.com/papers/1912.10128/

Of course, the authorities are more circumspect. Thus, on January 1, 2020, the Internet regulatory authority in China, aware that deepfakes can *"endanger national security, disrupt social stability, disrupt social order and undermine rights and legitimate interests of others"*, made mandatory the marking and reporting of deepfakes and virtual reality documents by their authors under penalty of sanctions.

Internet platforms are not, however, the only players to be taken into account in this debate. Other stakeholders may have a role to play.

2.5. Beyond the Internet players

Many ethical questions will find answers in the principles adopted by Internet players. But as the president of Microsoft and other industry leaders point out, these principles will not be enough. New laws and regulations are needed, especially as some platforms show a lack of willingness to self-regulate.

But beyond these new laws, other actors have to be involved. Citizens, employees, consumers, and investors can also take action on their own to limit the harmful effects of ethical violations. Taken as a whole, they can, in particular, promote the companies that are most active in solving the ethical problems caused by their products and services. Employees can share their concerns with management, refuse to work on unethical projects, or decide to leave the company. Consumers can boycott companies that have questionable ethical practices. Investors can request that ethical considerations be taken seriously by the companies in which they invest, or direct their investments to companies that place greater emphasis on these issues.

All stakeholders have their role to play in the fight against the misuse of content generation algorithms.

Proposition No. 12
Involve employers in the fight against deepfakes. Encourage them to adopt ethical charters, to inform their employees, customers, suppliers, and partners of their good ethical practices on the subject. Encourage the development of training courses around deepfakes.

3. Cadre légal, régulation et gouvernance

3.1. Mapping of players

The legal challenges posed by content generation algorithms are not new and fall within well-established areas of law: freedom of expression, image rights and property rights, in particular. However, legal tools do not always seem able to adapt to the evolution of technology (speed of propagation, volume, low production cost).

Four major players can be identified:

1. The creators of algorithms (companies, independent developers, etc.).
2. The authors of deepfake content (content based on algorithms).
3. Internet platforms (Facebook, YouTube, Reddit, among others).
4. Users (who watch and share deepfakes online).

This list makes it possible to assess the degree of legislation needed to regulate the algorithms of content generation. If laws governing content producers can be dealt with at a national or European level, issues linked to destabilization operations carried out by States or the regulation of platforms with a global reach, should, in our view, be dealt with at supra-state level.

3.2. Legal framework

There is a pre-existing legal framework applicable to content generation algorithms, which are traditional rules relating to freedom of expression, image rights and property rights. More recently, laws pertaining to personal data and fake news were added to this list.

- **Freedom of expression**: Content generation algorithms are a new tool of expression allowing people to create art or disseminate political or humorous ideas. This type of information is protected by freedom of expression, a fundamental freedom enshrined at international level (by the Universal Declaration of Human Rights of

1948), at European level (by the European Convention on Human Rights), and at national level (by the Declaration of the Rights of Man and of the Citizen of 1789). This freedom implies that everyone has the right to their own opinion and ideas and can express them by any means, in any format and through any medium: oral, written, audiovisual, cultural, virtual, artistic, etc. Considered a "universal right which constitutes an element of any democracy"[71], freedom of expression may nevertheless be limited when the following apply:

  - Public order (national security, territorial integrity, public safety, prevention of offenses, protection of public health, protection of democratic order, fight against racism, etc.).
  - Other rights (protection of privacy, protection against defamation and insult, right of reply, confidential information and business secrecy, image rights, presumption of innocence, etc.).

- **The right to one's own image:** Content generation algorithms are often linked to the distribution of images (photos or videos). Here the right to one's own image ensures the right to a private life. The right to one's own image is the exclusive right that everyone has over their image and its use, regardless of medium (photos, videos, etc.) or context (holidays, family events, cultural or religious events, etc.). This right (like freedom of expression) also has limitations. The main ones are:
  - A person in the background of an image taken in a public place.
  - The right to information when the image illustrates current events, a general debate on democracy or a social phenomenon, a historical subject or a public figure.
  - Parody, caricature or pastiche.

[71] Constitutional Council, 1994: freedom of expression is a "fundamental freedom all the more precious as its existence is one of the essential guarantees of respect for other rights and freedoms."

Apart from these exceptions, the content creator must, before any dissemination, obtain prior consent from the person depicted, specifying the purpose for which consent is given. Consent needs to be clearly defined. For instance, accepting to be photographed does not automatically imply consent to the dissemination of the image. The same rules apply to users of content generation algorithms. Thus, the author of a content generation algorithm sharing an image of a person, without having obtained prior consent, becomes liable according to article 9 of the French Civil Code. The author also becomes criminally liable (and risks up to one year in prison and a fine of 45,000 euros) if the offenses are as defined in articles 226-1 (invasion of privacy by taking or recording the image of a person in a private place) and 226-2 (sharing an image of a person in public taken without their consent) of the French penal code.

- **Right of ownership**: Content generation algorithms raise issues about the nature of ownership: who owns the object created and disseminated?

  The Bellamy painting (a painting created using artificial intelligence), threw up questions of copyright and authorship. Who was the author and copyright holder of the artwork?

  - Ian Goodfellow, inventor of content generation algorithms, whose mathematical equation appears at the bottom of the painting?
  - The creator of the code that was used and which is accessible in open source on the Internet?
  - The members of the French art collective Obvious?

  We had the opportunity to talk to Hugo Caselles-Dupré, member of the Obvious collective. He told us that while Ian Goodfellow never claimed ownership, the creator of the open source code attempted to claim ownership of the code and hence the painting itself. In this particular case, a consensus on the ownership of the eleven

paintings in the Bellamy series was reached between the collective and the creator of the open source code, granting copyright to the Obvious collective. However, as the courts were not involved in the case, the questions surrounding copyright and authorship of works created using artificial intelligence systems, remain, to date, a doctrinal debate. Current discussions (the European Parliament Resolution of February 16, 2017, public consultation on artificial intelligence initiated by the WIPO in December 2019) favor the protection of these works through copyright, which would require adapting this right, as it is currently mainly attached to the author. It should be noted that a first step in this direction has been taken in China. The Chinese court of Nashan recently settled a dispute relating to a product generated by an intelligent writing system that was granted protection under copyright law (Nanshan District People's Court, Shenzhen Tencent Computer System Co., Ltd. v. Shanghai Yingxun Technology Co., Ltd., December 24, 2019).

- **Protection of personal data:** Since the implementation on May 25, 2018, of Regulation (EU) 2016/679 of the European Parliament and the Council of April 27, 2016 on the protection of individuals with regard to processing of personal data and the free movement of data ("General Data Protection Regulation" or "GDPR"), the processing of personal data is lawful under certain circumstances as listed in article 6 (including, but not limited to, legitimate interest, execution of a contract, or consent).

  Data controllers, i.e. the employees within an organization deciding the purpose for which personal data is used, must also comply with the following principles:

  - Lawfulness, fairness and transparency (processing must fall within one of the legal bases listed by GDPR and information relating to processing must be easily accessible and formulated in clear and simple terms).
  - Purpose limitation. Data must be relevant and limited to what is necessary for processing.

- Data accuracy.
- Storage limitation.
- Data minimization.
- Data security principle, in particular against any risk of unlawful processing, loss, destruction or accidental or willful damage.

The GDPR also provides the following rights for data subjects:

- The right to be informed about the characteristics of the processing.
- The right to access data and change if it is inaccurate.
- The right to erasure (or "right to be forgotten"), if the processing was based on the consent of the person and the latter withdraws it or when the data is no longer necessary for the purposes processing, for instance.
- The right to restrict processing, for example when the data subject disputes the accuracy of the data, for the time necessary to verify this accuracy.
- The right to data portability in order to allow the data subject to receive their data in a structured, commonly used and machine-readable format to transmit it to another data controller.
- The right to object, in particular if the processing is aimed at profiling or commercial prospecting actions.
- The right not to be the subject of an automated decision (unless this decision is necessary for the conclusion or performance of a contract or if the person has explicitly consented to it).

- **Protection against fake news:** Art. 27 of the French Press Law of 1881 provides that publication, dissemination or reproduction of fake news and / or news items that are fabricated, falsified or deceptively attributed to third party is punishable with a fine of 45,000 euros, where it was done in bad faith and disturbs or is likely to disturb public order.

It is, however, a crime against the state and not a crime against individuals, so that only entities having an interest in defending it can invoke it (e.g. the Public Prosecutor on behalf of the State, but also associations having for object the defense of the public good for example). A private individual, therefore, cannot file a complaint or initiate legal action on the basis of this law.

In addition, the French “anti fake news law” passed in 2018 provides for the following measures:

- The obligation for platforms to disclose the identity of authors of advertisements disseminated during elections as well as the money spent on such advertisements. This transparency obligation is not, however, subject to any sanctions.
- During the three months preceding a national election , summary proceedings can be taken (by emergency procedure) to halt the distribution of fake news. Courts can rule within 48 hours whether (i) the false information is disseminated in an "artificial or automated" manner and "on a large scale", (ii) the inaccuracy or misleading nature of the information is proven, and (iii) there is evidence that the fairness of the ballot is at risk.
- The Conseil Supérieur de l'Audiovisuel (CSA) can prevent or suspend the broadcasting of television services controlled by a foreign state or under the influence of a foreign state that undermine the fundamental interests of the nation. The CSA can also issue recommendations, carry out investigations, and in some cases apply sanctions to platforms that do not comply with the obligation of cooperation defined below.
- Online platforms are obliged to provide their users with an easily accessible and visible device allowing them to report false information as well as to implement additional measures relating in particular to the transparency of their algorithms, the fight against accounts spreading false information or to the notification of users regarding the nature, origin and methods of content

distribution as well as the identity of the people who pay them in return for promotion of factual content relating to a debate of general interest. These measures must be made public.

3.3. What tools does French law offer today to deal with content generation algorithms that violate people's rights?

France does not have a specific law on content generation algorithms, but the existing legal framework makes it possible to regulate algorithms and their possible infringements of human rights. In this chapter, we are particularly focusing on deepfakes, where the image of an identifiable person is used and may infringe copyright laws, or may be defamatory, offensive and / or convey false information.

In this context, the victim has two options:

- **Take action against the platform,** using the 2004 law on trust in the digital economy ("LCEN"), in order to request all measures to prevent damage or to put an end to the damage caused by content of an online communication service. Most frequently the request will relate to the withdrawal of the contested publication.

  Art. 6, section I. of "LCEN stipulates that any person can:

  - Notify the host of a website to remove contentious content (article 6., section I.-5);
  - Request the hosting or Internet access provider to take "all appropriate measures to prevent damage or to put an end to damage caused by the content of an online communication service to the public", in particular the withdrawal of content ("Internet referral" provided for in article 6, section I.-8).

- **Take action against the publisher of the content** and seek compensation in order to repair its damage.

The above scenarios are evidenced in the case studies below.

3.3.1. Case study 1: deepfake infringing the right to one's image

This happens when a person's image is used without permission.

In such a case, the person whose image was used may:

- Contact the manager of the site on which the image is published and ask them to unpublish it on the basis of their right to erase personal data. In the absence of a satisfactory response within 30 days, the person can appeal to the CNIL free of charge, which may impose several types of sanctions (warning, injunctions, financial sanctions, etc.).
- Refer the matter to the civil courts, if necessary in an emergency procedure, in order to obtain the withdrawal of disputed images and compensation (article 9 of the French Civil Code). However, a decision is often only obtained after several weeks.
- File a complaint against the author, on the basis of articles 226-1 to 226-2 of the French Penal Code.

3.3.2. Case study 2: defamatory deepfake (Art. 29ff of the French Press Law of 1881)

A deepfake is defamatory when it alleges a fact relating to the private life of a person that harms their honor or respect.

In the context of deepfakes cited above, the videos representing Mark Zuckerberg and Nancy Pelosi could be considered defamatory.

The victim can then file a complaint against the author of the broadcast (who incurs, when the broadcast is public, a fine of up to € 12,000 or € 45,000, depending on the case, for defamation committed against individuals) and seek damages to repair the damage suffered as a result of the defamatory publication.

3.3.3. **Case study 3: offensive deepfake** (Art. 33ff of the French Press Law of 1881)

A deepfake is offensive when the accusation it contains is not a verifiable fact and its purpose is to hurt or offend the person represented. In this context, the author of the deepfake will not be able to invoke good faith to exonerate themselves.

As with a defamatory act, the victim will have to file a complaint, and the perpetrator incurs the same penalties.

3.3.4. **Case study 4: deepfake disseminator of false news (**Art. 27 the French Press Law of 1881 **and "anti fake news" law of December 22, 2018)**

Fake news is defined as an item of information (potentially defamatory, abusive, disturbing public order, etc.) whose purpose may be to manipulate public opinion (e.g. lobbies, pressure groups), to distort a stock market price, to discredit a competitor, to influence political choices, to instill hatred against a community or to attract traffic by disseminating extravagant or provocative information.

- Where a deepfake is (i) a vector for disseminating fake, fabricated or falsified news or news deceptively attributed to third parties and (ii) disturbs or is likely to disturb public order, it falls under the fake news offense stipulated by article 27 of the 1881

press law. However, as explained above[72], only the Public Prosecutor's Office can initiate an action on the basis of this article.

- In addition, the emergency procedure created by the 2018 law on the fight against the manipulation of information (known as the “anti-fake news” law), makes it possible to ask to stop the dissemination of false information during three months before a national election. When called upon, the judge must assess, within 48 hours, whether the false information is disseminated in an "artificial or automated" manner and "on a large scale", (ii) the inaccuracy or misleading nature of the information is proven, and (iii) there is evidence that the fairness of the ballot is at risk. It is important to note here that despite its extremely broad wording, the anti-fake news law is strictly limited to electoral periods and consequences.

But if a deepfake is likely to have consequences or implications in a particular industry, other specific French laws are likely to apply:

- The offense of deceptive commercial practice (article L.121-2 to L.121-4 and L.132-2 of the Consumer Code.
- Unfair competition, including free riding and denigration.
- The offense of "false facts likely to damage the credit of the nation" (article 410-1 of the Penal Code).
- The offense of infringing market transparency (articles L.465-3-1ff of the Monetary and Financial Code).
- Changes made to the words or image of a person without their consent (article 226-8 of the penal code).
- The crime of identity theft (article 226-4-1 of the Criminal Code).
- The offense of slanderous denunciation (article 226-10 of the Penal Code).
- The offense of disclosing threats and false alarms (article 322-14 of the Penal Code).

[72] section 3.2 of this report

- The right established by GDPR not to be the subject to a decision exclusively based on automated processing, including profiling, producing legal effects concerning them or significantly affecting them, subject to the exceptions referred to above[73].

French law is therefore not completely helpless in the face of deepfakes and has several instruments allowing victims to request the withdrawal of fraudulent publications and / or - at least in theory in the face of the difficult identification of the perpetrators - compensation for the damage suffered. However, few victims know the extent of their rights. The existence and visibility of existing support structures should therefore be developed.

Proposition No. 7
Promote access of existing rules for victims of deepfakes, identity theft or harassment to legal remedies and their rights through the development and exposure of dedicated platforms.

Most of the aforementioned rules are intended y for the benefit of consumers of content generation algorithms (in order to protect them) and against the producers (in order to sanction them). Few of the existing rules (with the exception of the law on trust in the digital economy) directly target platforms which, as their name suggests, play a central role in the large-scale distribution of content (including algorithms generation of content).

[73] section 3.2 of this report

3.4. What responsibility should be incurred by the platforms?

Digital platforms now play a major role in the distribution of content. The examples presented above show to what extent platforms accelerate distribution and increase the volume of content. The debate on the liability of platforms is extremely heated because of the legal exemption they are subject to. The LCEN law, which transposes directive 2000/31 / CE known as "e-commerce directive", provides for a derogatory liability regime for hosting providers, that is organizations providing services storing and making the data produced by third parties available to the public.

Hosting providers cannot be held liable for criminal liability due to content uploaded by third parties, most likely the producer of such content, if they were not aware of illegal activities or complaints. Hosting providers may not be held liable if, from the moment they became aware of illegal activities, they acted promptly to withdraw content or make it impossible to access it. This derogatory regime applies to platforms, as long as they do not act as content publishers, have not determined or selected content and have not been aware of or have not verified it. Originally, this legal framework was aiming at developing trust in the digital economy by increasing transparency and reducing obligations of hosting providers, and thus limiting compliance costs.

But platforms have come under additional scrutiny, especially regarding the spread of misinformation through deepfakes. To anticipate public authorities' responses, large platforms have voluntarily taken preventive measures, which are detailed below.

Voluntary measures taken by platforms against deepfakes
Facebook, Amazon and Microsoft have launched the Deepfake Detection Challenge (DFDC) to develop algorithms that can identify deepfakes and

media which have been manipulated[74]. Facebook has also decided to remove deepfake documents from its platform[75] based on the following criteria:

- The material has been edited or synthesized, beyond adjustments for clarity or quality, in a way which is not obvious to the average individual and which could be misleading by suggesting that a subject in the video said words it didn't actually say;
- The material is the product of artificial intelligence or machine learning that merges, replaces, or overlays content on a video, making it appear authentic;
- Parody or satirical content is excluded from restriction.

Twitter is also in the process of implementing its policy against deepfakes[76]. Twitter will flag users of accounts that will share deepfake documents, notify accounts that wish to share those documents ("retweet"), and provide reasons why it believes the document to be a deepfake.

Google offers the option to remove deepfake images or videos from its search results (but not from the Internet) by filling out a form online[77]. It is possible that the volume of requests is too large to make this method effective, especially for public or famous figures.

[74] Facebook AI, *Deepfake Detection Challenge Dataset*, 2020, https://deepfakedetectionchallenge.ai/

[75] Facebook, *Enforcing Against Manipulated Media*, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

[76] Sarah Perez, *Twitter drafts a deepfake policy that would label and warn, but not always remove, manipulated media*, 2019, https://techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/

[77] Google Search Help, https://support.google.com/websearch/answer/9116649?hl=en

These approaches are not necessarily followed by all platforms. **Snapchat** and **TikTok**[78] plan to incorporate a deepfake creation feature to increase the use of their platforms. Snapchat has notably acquired AI Factory[79], a company specializing in the creation of deepfakes, partly based in Ukraine.

Nevertheless, the regulation of platforms keeps increasing, in particular with the implementation of the General Data Protection Regulation (GDPR) of 2018. If GDPR is not intended to regulate deepfakes or fake news, it is nonetheless an effective tool that regulators can use to curb the content of deepfakes, especially those using images of people whose consent has not been obtained. The Zao app, which became very popular in China during the summer of 2019, offered deepfake services by replacing the faces of stars with that of users[80], but its terms of use stipulated that the rights granted by Internet users on their images (e.g. their copyright) were *"free, irrevocable, and permanent"*. If this type of practice is common with mobile apps, it does not mean it's legal. Under GDPR, the purpose for processing personal data must be clearly established in the terms of use and/or a privacy notice provided to individuals.

GDPR may be leveraged for content generation algorithms for the following purposes:

- To sanction producers who illegally collect images of people (without their prior consent)

[78] Michael Nuñez, Snapchat and TikTok Embrace 'Deepfake' Video Technology Even As Facebook Shuns It, 2020, https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/#5995495242c0

[79] Ingrid Lunden, *Snapchat quietly acquired AI Factory, the company behind its new Cameos feature, for $166M*, 2020

[80] Alexis Zema, *Zao, l'application de vidéos « deepfake » qui inquiète les internautes chinois,* 2019, https://www.lefigaro.fr/secteur/high-tech/vie-privee-les-videos-deepfake-de-l-application-zao-inquietent-les-internautes-chinois-20190902

- To make platforms accountable when distributing content generation algorithms. The General Data Protection Regulation (GDPR) is based on the principle of the accountability of data controllers (there is no prior authorization before processing data) who must set up privacy programs
- To use the current resources of the CNIL for education, audit and formal notice purposes. GDPR lists the obligations that national authorities, such as the CNIL, must meet, which are, amongst others, education, audit and formal notice obligations for non-compliance with the regulations. For example, the CNIL has produced reports on web-enabled devices and artificial intelligence. Automated content generation algorithms could also be investigated and become the subject of a public debate.

While GDPR can be an interesting public policy tool, it remains limited for several reasons:

- It only applies to the processing of personal data.
- The application of GDPR outside the territory of the European Union is problematic. If regulatory authorities, such as the CNIL in France, have some level of enforcement power, the implementation of the extraterritoriality principle is not easily enforceable. In the case of foreign actors using data from European citizens outside the European Union, the question may arise as to what extent a decision of a European court is applicable.

A more effective public policy tool may consist of making platforms more accountable.

The 2000 e-commerce directive, transposed into French law in 2004 by the law on trust in the digital economy, provides for an accommodating legal regime for platforms like Facebook and Google. Today, the following platform obligations no longer seem sufficient in the face of the challenges related to content generation algorithms:

- Obligation to *“promptly”* remove any data whose content is *“manifestly illegal”*.
- Obligation to inform the competent public authorities *"promptly"* of reports received by users for limited illegal activities (incitement to hatred, procuring, child

pornography, corruption of minors). The government platform Pharos (Platform for Harmonization, Analysis, Cross-referencing and Referral of Reports) includes a team of 20 investigators to deal with complaints from platforms but also directly from users.

- Obligation to provide a system that is *"easily accessible and visible, allowing anyone to bring to their attention"* content contrary to the general interest.
- Obligations to communicate on the means implemented in the fight against illegal activities.
- Transparency obligations when disseminating content in return for payment. Those who exceed a certain volume of connections per day must have a legal representative in France, implement measures that may relate to the transparency of their algorithms and publish aggregated statistics on the operation of their algorithms for recommending, ranking or referencing content.

Non-binding opinions have been issued by public authorities. The CSA recommendation (May 15, 2019) suggests that platforms exceeding a threshold of 5five million visitors a month take certain concrete steps intended to ensure the transparency of information. Similarly, the 2019 report by the National Agency for the Security of Information Systems (ANSSI) warns against actions of a destabilizing and influencing nature.

Considering the importance of platforms in the distribution and production of content (in particular by content selection algorithms or platform design), an intermediate liability status may better address challenges stemming from content generation algorithms. The European Commission is currently discussing such measures.[81] The French Senate proposed an intermediate liability status (called “publisher of services”) as part of a recommendation for a European resolution in 2018.[82] A publisher of services could be *"a company which obtains a*

[81] Julien Lausson, *Bruxelles entend accentuer la responsabilité judiciaire des géants du net*, 2020, https://www.numerama.com/politique/598046-bruxelles-entend-accentuer-la-responsabilite-judiciaire-des-geants-du-net.html

[82] Sénat, *Proposition de résolution européenne en application de l'article 73 quinquies du Règlement, sur la responsabilisation partielle des hébergeurs,* 2018,http://www.senat.fr/leg/ppr17-739.html

direct economic advantage from the access of hosted content", subject to intermediate obligations between those borne by publishers and those of hosts. The table below summarizes this intermediate status:

Editeur *(sans changement)*	Editeur de services *(nouvelle catégorie)*	Hébergeur *(sans changement)*
Régime de responsabilité de **droit commun** : - a une obligation d'identification des personnes qui ont créé un contenu qu'il héberge ; - a une obligation de surveillance du contenu des pages éditées : il s'agit d'une **obligation de résultat,** et non de moyens.	Régime de **responsabilité intermédiaire** L'éditeur de services : - a une obligation d'identification des personnes qui ont créé un contenu qu'il héberge ; - a une obligation de mettre en place les moyens, conformes à l'état de l'art, de surveillance des informations qu'il transmet ou stocke, et de recherche des faits ou des circonstances révélant des activités illicites ; il s'agirait donc d'une obligation de moyens, non de résultat ; - peut être tenu civilement ou pénalement responsable s'il a connaissance d'activités ou d'information manifestement illicites et qu'il n'agit pas promptement pour retirer ces informations ou en rendre l'accès impossible.	Régime de **responsabilité aménagée** L'hébergeur : - a une obligation d'identification des personnes qui ont créé un contenu qu'il héberge ; - n'a pas d'obligation générale de surveillance des informations qu'il transmet ou stocke, ni d'obligation générale de rechercher des faits ou des circonstances révélant des activités illicites ; - n'a pas d'obligation générale de surveillance des informations qu'il transmet ou stocke, ni d'obligation générale de rechercher des faits ou des circonstances révélant des activités illicites ; - ne peut être tenu civilement ou pénalement responsable que s'il a connaissance d'activités ou d'information manifestement illicites et qu'il n'agit pas promptement pour retirer ces informations ou en rendre l'accès impossible.

8384

If the ethical and non-binding measures taken by major Internet platforms are getting closer to the intermediary liability status listed above, this could indicate that this regime may benefit large American players. A more flexible system, that of limited liability, could be considered for platforms below a certain threshold (in number of users, traffic, etc.).

[83] NextInpact, *L'éditeur de services, nouveau statut pour surveiller les contenus,* 2011, https://www.nextinpact.com/archive/61904-statut-hebergeur-editeur-service-ligne.htm

[84] Gérard Longuet, *Le devoir de souveraineté numérique*, 2020, http://www.senat.fr/rap/r19-007-1/r19-007-13.html

In addition to the significant self-regulation efforts made by large platforms, public policy tools exist, such as the GDPR, or are in the process of development, such as intermediary responsibility.

3.5. Projects in progress

French law, however, mainly comes up against two major problems:

- **The inadequacy of criminal sanctions incurred by the perpetrators with regard to the consequences on public order** (let alone the consequences for the victims, which the criminal sanction is not intended to repair).

  Indeed, the sanctions provided for today are mainly contained in legal instruments dating from the 19th century, at a time when it was less easy to have access to large-scale distribution tools (mainly newspapers, which implied having a contact with a publisher) and especially when information distribution was very limited from a geographical perspective (most newspapers having local or national circulation).

  Today, where all it takes is a smartphone and an Internet connection to broadcast an image or other content to the whole world in seconds, a perpetrator of illegal broadcasts is less aware of any illegal acts he may commit. The consequences on public order and for the victim are inversely proportional to the steps taken by the author to disseminate his content.

  This is why we suggest strengthening the criminal penalties incurred by the perpetrators with two objectives: (i) as a deterrence and (ii) to limit the disturbance caused by the disputed broadcasts to public order.

Proposition No. 8

Strengthen existing criminal sanctions against authors of deepfake content and reassess them taking into account the new technological and digital environment where the creation of deepfakes is made possible.

We also suggest, as laid out below in propostion No. 11, to require content publishers to report any deepfake content in order to limit the impact on victims, and that in the absence of reporting, the criminal penalties incurred are aggravated.

- **Broadcasters' liability.** Identifying authors of news content is complex. This is the reason why new legal instruments must aim at reaching broadcasters and not directly authors. This also seems to be the most effective way to fight, not against the publication of such content, which is necessarily difficult to contain, but against their wide dissemination r, which generates harm.

  For several reasons, however, the liability of broadcasters, remains largely theoretical:

  - Hosts are only liable if they have not acted *"promptly"* in withdrawing an illegal publication or blocking access to a site that has been reported to them. No sanction is therefore incurred in the event of prompt compliance, but above all, no quantitative definition is given of this requirement for responsiveness, which may create a certain flexibility in the interpretation made by platforms.
  - The means available to platforms to exercise a potential recourse against the publisher of content (the one who is at the origin of the publication and the information, e.g. the author of a tweet).

Many people are therefore campaigning to strengthen the liability of large platforms, in particular in view of their market power. As such, it is interesting to look at the systems implemented across the world.

3.6. International illustrations

With fake news spreading all over the world, most states seem to have seized on this problem, each finding its own "anti-fake news" bulwark[85].

3.6.1. In Europe

Several European countries have adopted measures to fight against disinformation and more specifically fake news, for example:

- Germany has passed a law known as the "NetzDG measure" providing for a fine of € 50 million against social media companies that do not remove hate posts or *fake news*.
- Great Britain has distinguished itself by creating a special agency called the "National Security Communications Unit" to fight fake news.
- In January 2018, Italy launched a website, linked to the Police, where anyone can report potential fake news that they may have identified.

At this stage, only Switzerland seems to be an exception, concluding in a report of May 2017 that, *"for lack of hindsight, it is however not possible at present to assert to what extent state regulation is necessary."*

3.6.2. Outside Europe

[85] Corentin Lacoste, *À l'étranger, quelles législations contre les fake news ?*, 2018, https://www.liberation.fr/politiques/2018/06/09/a-l-etranger-quelles-legislations-contre-les-fake-news_1657644

The United States does not have a federal law on deepfakes, but discussions have been held in Washington D.C. for the past two years. However, some states went ahead on their own. For instance, the state of Virginia, which adopted a law in July 2019 which includes ″deepfakes″ in its fight against ″revenge porn″ (the practice of sharing pornographic content without the individual's consent). People found responsible for sharing content (true or falsified via content generation algorithms) will be fined € 2,500 and up to twelve years of imprisonment.

Additional examples include the followings:

- In a 2019 law, Kenya listed 17 cyber crimes including the publication of fake news, which can lead to two years in prison and a fine of five million Kenyan shillings (approximately € 42,000).
- In Malaysia and Indonesia, perpetrators of fake news face six years in prison, which, as some commentators have noted, raises fears of potential political censorship.
- Prompted largely by the reaction to the Chinese app Zao, China recently adopted, a law that as of January 1, 2020, which prohibits fake news and requires that content generated by algorithms is distinctly identified as such. It also requires platforms to identify the real identity of users and immediately remove content identified as illegal. The content may be restored if the algorithm behind the content is revealed.

In other countries[86], bills are underway (Croatia, South Korea) or the issue is under consideration (Belgium and Singapore).

In view of the practices implemented in other countries of the world we have therefore formulated the following recommendations.

[86] Harold Grand, *En Estonie, une intelligence artificielle va rendre des décisions de justice*, 2019, https://www.lefigaro.fr/secteur/high-tech/en-estonie-une-intelligence-artificielle-va-rendre-des-decisions-de-justice-20190401

3.7. Our recommendations

Our recommendations are as follows:

- Require platforms to verify identification data of new accounts, and, in the absence of real identification of the author, close the account without delay.
- Limit and legally quantify the withdrawal period for illegal content reported to platforms under penalty of a heavy fine.
- Require hosts located outside France to appoint a representative in France who assumes the responsibility of the host who has not withdrawn within this legally fixed period content reported as illegal.
- Allow in certain exceptional cases (legally defined), where there is clear evidence of illegal content, to order the permanent removal of a site, a profile, an account, a page, a group, without giving the site in question the opportunity to defend itself Ceci ne vaudrait, à ce stade, en l'absence de solution supranationale, que pour les sites, profils, groupes hébergés sur le territoire national.
- Systematically combine the financial sanctions imposed on platforms with publicity measures.
- Develop a supranational organization for cyber-policing and judicial cooperation focused on the matter in order to facilitate the exchange of information between countries based on the model of Europol which exists in the fight against crime and terrorism, or even of iPROCEEDS, which currently focuses on financial investigations in order to search, seize and confiscate the proceeds from cybercrime and prevent money laundering on the Internet.

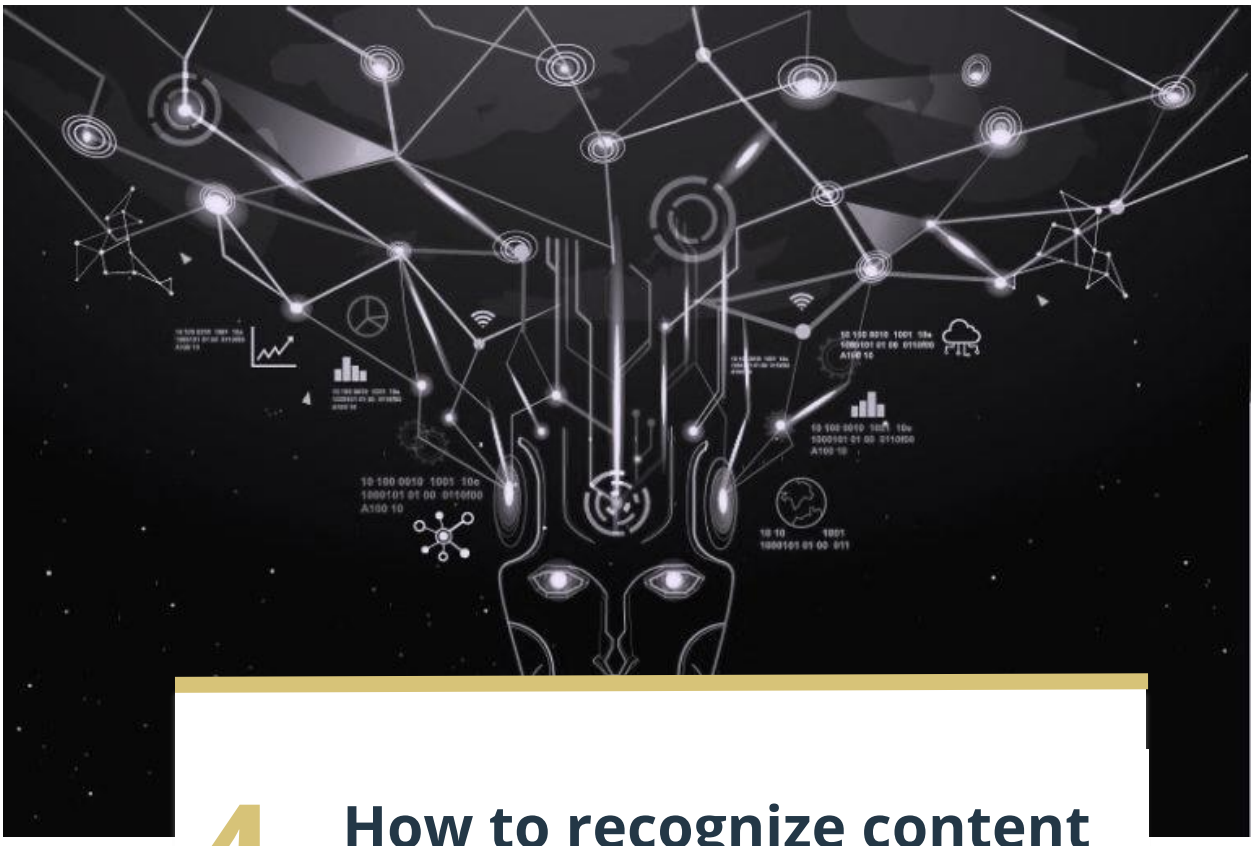
In this context, given the cross-border nature of the Internet and the content passing through it, such a comprehensive and protective national law will never be enough to fight truly effectively against deepfakes. This is why we suggest strengthening at state level but above all at supra-state level (the transmission of information knowing no borders) the responsibility of the platforms.

Proposition No. 9

Strengthen at state level but above all at supra-state (the transmission of information knowing no borders) the responsibility of platforms.

In this way:

- Develop a supranational organization for cyber-policing and judicial cooperation focused on the matter in order to facilitate the exchange of information between states.
- Allow in certain exceptional cases, where there is clear evidence of illegal content, to order the permanent removal of a site, a profile, an account, a page, a group, without giving the site in question the opportunity to defend itself
- Limit and legally quantify the withdrawal period for illegal content reported to platforms under penalty of heavy fines
- Force platforms to verify identification data when a new account is created, and, in the absence of real identification of the author, close the account without delay
- Require hosts located outside France to appoint a representative in France who assumes the responsibility of the host who has not withdrawn within this legally fixed period content reported as illegal
- Systematically match the financial sanctions imposed on platforms with publicity measures.



4. How to recognize content generated by an algorithm? How to go further to protect yourself from *fake content*?

Despite extremely rapid progress and our desire to improve our ability to detect fakes, “forgers” will adapt and manage to generate fakes that are even less detectable. There are a few tools that can help us have more or less confidence in the validity of a photo, video, voice or text.

We will first discuss the most basic means of detection, which is for the authors of deepfakes to report them as such. Then we will unveil some light-hearted techniques that allow, with a little practice, to detect an image generated by an algorithm. We will then discuss the delicate subject that is the detection of fakes by algorithms. Indeed, content generation algorithms create fakes, but also inherently rely on fake detectors. It is therefore possible to detect a forgery thanks to an algorithm similar to the one used for its creation. Subsequently we will discuss solutions that are more accessible but nevertheless increasingly essential to authenticate and certify content. Finally, our conclusion will stress the need to raise awareness and invest in scientific research.

Proposition No. 1
Support the improvement and development of techniques for classifying content, whether true, false or suspect (fact-checking, watermarks, algorithms, blockchain), while defining European standards in order to support Internet users in identifying reliable content.

4.1. Detection of fakes

4.1.1. Fact-checking

In a world where creating fake content is getting easier and easier, a first response appeared in the 1990s: fact-checking sites. We note for instance the creation of Snopes[87] in 1995, TruthOrFiction.com in 1999, then FactCheck.org[88] in 2003 and Politifact in 2007. We should also mention www.eufactcheck.eu, an initiative by the European Journalism Training Association (EJTA). The project brings together some 20 European schools of journalism and offers a methodology developed by teacher-journalists for teaching fact checking.

Various newspapers now have their own fact-checking blogs: “Les Décodeurs” for *Le Monde*, “Political Scan” for *Le Figaro*, “Désintox” for *Libération*.

Given the mass of information available on the Internet, investigative journalists cannot fact-check everything. With the goal of automating these tasks, eight years after launching its fact-checking blog "Les Décodeurs", *Le Monde* announced in February 2017 the launch of its automatic source verification tool "Décodex". Décodex offers a plug-in that allows users to verify a source of information. The plug-in checks whether a source is referenced or not in its database and assigns it a classification.

[87] **Snopes** is an English-speaking website launched in 1995 by Barbara and David Mikkelson which aims to limit the spread of false news and unfounded rumors circulating on the Internet. In 2016, Facebook partnered with Snopes to tackle fake news. In 2019, Snopes withdrew from this collaboration, believing the workload on their teams was too high, that the activity was not paid enough and that its impact was not as expected.

[88] **FactCheck.org** is an American fact-checking website focusing on news and run by the University of Pennsylvania (Annenberg Public Policy Center of the University of Pennsylvania).

source: http://www.lemonde.fr/les-decodeurs/article/2017/01/23/l-annuaire-des-sources-du-decodex-mode-d-emploi_5067719_4355770.html

However, we observe several limitations to this kind of tool. First, a reliable site like Le Monde may very well fall victim to fake news and relay it without realizing it. On the other hand, it is very complicated to define whether a source is reliable or not. Source classification tools are just one tool among many for spotting fake news, but relying entirely on them to have confidence in content does not seem sufficient.

Journalists need more tools to help them identify fake content beyond the reliability of the source. In this perspective, we see the emergence of augmented journalism where journalists rely on new tools based in particular on artificial intelligence.

The *Washington Post* prototype Truth Teller, launched in 2013, for instance aims to analyze political discourse in real time. The tool is based on a deep neural networks algorithm that translates dialogue into text and another algorithm that extracts fact from generated text and compares it to other texts.

Jigsaw[89] also announced in 2016 the launch of the experimental ProjectAssembler[90] in partnership with Google and a number of universities. Its goal is to create innovative tools for detecting fake content. *"We built two new detectors to test on the platform. The first is the StyleGAN detector to specifically address deepfakes. This detector uses machine learning to differentiate between images of real people from deepfake images produced by the StyleGAN deepfake architecture. Our second model, the ensemble model, is trained using combined signals from each of the individual detectors, allowing it to analyze an image for multiple types of manipulation simultaneously."*[91]. The tool was released in early February 2020, and tests are currently being carried out by a panel of journalists.

Developing this kind of tool seems essential to better support journalists in the fight against false content. Journalistic investigation alone is not enough to verify the mass of information circulating on the web, particularly on social networks like Twitter and Facebook.

4.1.2. Visual detection of false images, some light-hearted examples

To create perfect fake images, algorithms still need a lot of computing time and data to train and perform well enough. Therefore, in many cases, it is still possible to visually detect fakes

[89] subsidiary of Alphabet, formerly *Google Ideas*

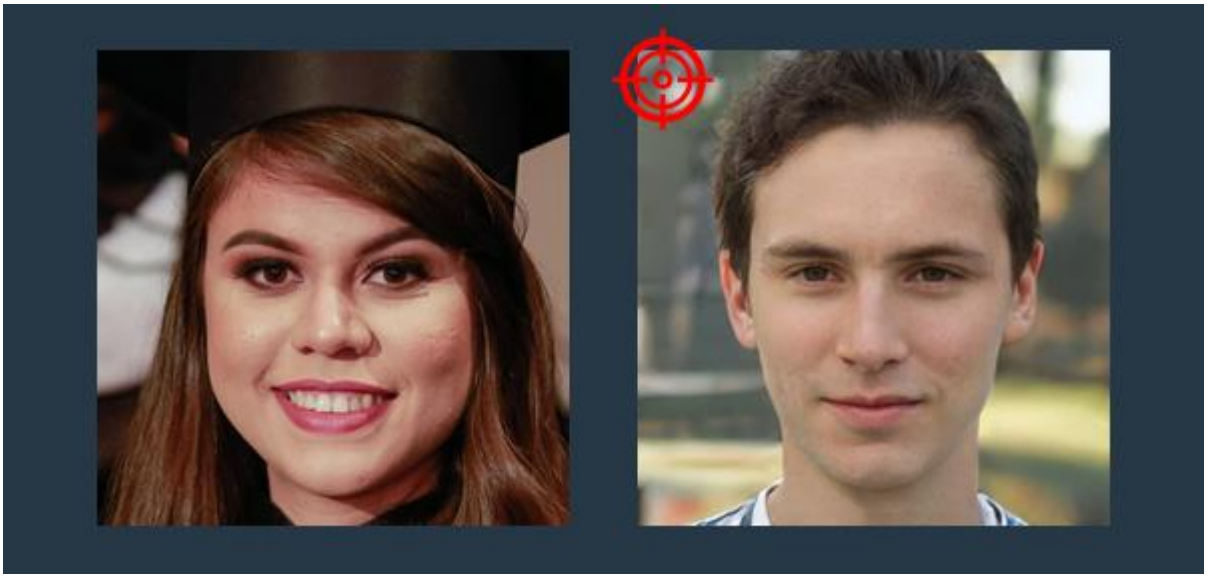
[90] Jay Peters, *Alphabet's Jigsaw unveils a tool to help journalists spot deepfakes and manipulated images*, 2020, https://www.google.com/url?q=https://www.theverge.com/2020/2/4/21122778/alphabet-jigsaw-assembler-tool-news-journalists-deepfakes-images&sa=D&ust=1582797385682000&usg=AFQjCNEbAzrRVVSVtqxnvGSISLsPXlLWwQ

[91] https://medium.com/jigsaw/disinformation-is-more-than-fake-news-7fdd24ee6bf7

with the help of a few recurring defects. In view of the rapid progression of algorithms, these defects may not be so visible any more in the future, but it is still relevant to know some clues that allow us to visually identify false images. This light-hearted section always seems very useful to us to raise awareness of the issue of deepfakes.

> ***In many cases, it is still possible to visually detect fakes using a few recurring defects.***

For instance, during the presentation made by Ekimetrics at the AI Paris forum in September 2020, the speakers asked the audience to select the photo that they believed to be fake from a selection. Here is a first example:



In this first example, the photo on the right is counterfeit while the one on the left is genuine. We can deduce it in particular thanks to the following clues:

- The background on the right does not look real.
- The ears of the individual on the right are not symmetrical and do not appear to belong to the same person.

Here is a second example:

Here the photo on the left is fake, and the one on the right is authentic. We can detect it with the naked eye using a few clues:

- The hair on the left does not look realistic, especially on the left side of the face.
- The clothes in the photo on the left arouse suspicion.
- The photo on the right shows a painting in the background, which is very complicated to generate with an algorithm.

Sometimes the rendering is very high quality and cannot be detected at first glance:

Source: Jonathan Hui, 2020[92]

[92] Jonathan Hui, Detect AI-generated Images & Deepfakes (Part 1), *Medium*, 2020

The defects here are on the asymmetry of the shoulders; the right shoulder is a different height from the left one. The background of the photo is also suspect.

Although many fake photos are not detectable with the naked eye, we have identified six "classic" signs that suggest that a face was probably generated by an algorithm: stains, background, glasses, lack of symmetry, hair and teeth. Conversely, some signs such as the symmetry, the complexity of an image or integrated text are elements that should reinforce our belief that a photo is genuine.

Symmetry

Complexity

Text

Fake images are a simple example to introduce the topic of deepfakes. The influencing power of images is such that they constitute the preferred tool for manipulating an audience. In practice, these methods can be used to create fun apps such as aging a face or making it smile. However, fake faces have been generated to create fake profiles on social networks, such as the profile of Russian Katie Jones mentioned in the introduction.

❝

We have identified six "classic" signs that suggest a face was probably generated by an algorithm: stains, background, glasses, lack of symmetry, hair and teeth.

4.1.3. Algorithmic detection

There is a discriminating algorithm within the GANs. The models are also capable of detecting some fakes, as do other content generation algorithms. They are not commonly used yet for this purpose, and it would seem useful to invest in research into these essential applications.

Deepfake Detection Challenge

In 2019, Facebook, Amazon, Microsoft, the Partnership on AI as well as several academic institutes (including MIT) launched the Deepfake Detection Challenge (DFDC)[93]. It aimed to develop deepfake detection tools. This challenge, which mainly targeted deepfake videos, was organized on the Kaggle platform and brought together more than 2000 participants. The general idea of the competition was to build a model capable of identifying false content within a video dataset; these had been specially designed for the purpose through several content generation algorithms. After a first model was selected with an accuracy of 85.26%, it was found not to perform as well on a black box dataset. The winning model reached an accuracy of 65.15% on this dataset. The objective of this challenge was to make researchers aware of the problems associated with deepfakes and to encourage them to implement innovative technologies in order to detect all kinds of manipulation. As a result of this challenge, Facebook's CTO announced that the platform would be building its own technology to deal with fake content issues[94]. We provide more detail about the Deepfake Detection Challenge in the appendix.

Facebook's CTO announced that the platform would be building its own technology to deal with fake content issues.

[93]Facebook AI, *Deepfake Detection Challenge Dataset*, 2020, https://deepfakedetectionchallenge.ai/

[94] James Vincet, Facebook contest reveals deepfake detection is still an 'unsolved problem', *The Verge*, 2020, https://www.theverge.com

Algorithmic detection of false texts

Several entities have joined forces to develop algorithms capable of detecting false texts generated by algorithms. The U.S. government, for example, has called — through its Defense Advanced Research Projects Agency (DARPA) — for projects on the topic. In particular, the SemaFor program, or Semantic Forensics, aims to develop an innovative technology that will analyze media in order to detect whether the content is generated or modified algorithmically (such as by GANs). The project will integrate several types of algorithms:

- Semantic detection algorithms as well as attribution models will be able to determine whether a fake originated from a company or from an identified individual.
- Characterization algorithms will make it possible to determine whether fake content was generated for malicious purposes or not.

All of these technologies will help the government identify and detect disinformation campaigns[95].

A simple approach to detecting dummy text is TF-IDF, the *t*erm frequency-inverse document frequency method. This method makes it possible to determine the most important words by comparing the relative frequency of the terms (term frequency, or TF) against the frequency and distribution of these same terms in all the other documents (inverse document frequency, or IDF). The importance of the word increases proportionally with the number of times it appears in the text, which is however compensated by the frequency of that word in the corpus. Thus, common words found in any document, including logical connectors, will be ranked lower. If a suspicious word, however, appears several times in a document without it being repeated in the corpus, it will be ranked higher. Since the TF-IDF method only takes into account raw, high-ranking weightings, it is not able to identify semantic relationships. For some this method remains non-optimal, useful for information

[95] Dr. Matt Turek, Semantic Forensics (SemaFor), DARPA, https://www.darpa.mil/program/semantic-forensics

research rather than content optimization. However, the weakness of this method lies mainly in the fact that it requires a high number of words to be effective. Thus, the results are inconclusive on short texts describing products, or even on blog articles.

Another type of algorithm from the family of transformers, such as GPT-3, BERT, or Bidirectional Encoder Representations from Transformers, was published by Google AI Language. BERT comes with a number of features, including question answering and natural language inference. The main innovation in this algorithm is its ability to read text not only from left to right, but also from right to left. Thus, with the help of transformers, BERT is able to learn contextual relationships between two words by looking at the words which surround them. Because of this, its decoder is able to predict the next word[96]. By hiding words randomly, BERT is able to predict false content. In a first stage of pre-training, two unsupervised learning stages are applied. In the first, 15% of the words in a document are hidden to be predicted later using the Masked LM method, by looking at the rest of the content. The model then proceeds to predict the following sentence (NSP or Next-Sentence Prediction), with methods such as natural language inferences and question answering. In a second step, BERT applies the self-attention process in order to learn the correlations between current and previous words, thus allowing "to capture the two-way cross-attention between pairs of sentences"[97].

University research

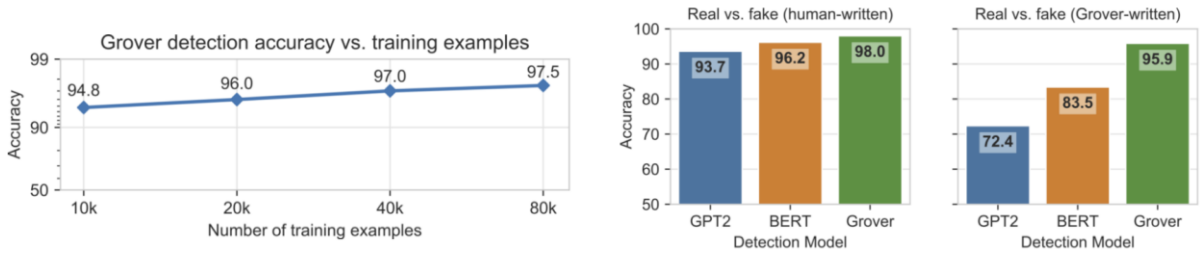
Recent work by a group of researchers, the Allen Institute for AI, allowed to develop a model called Grover[98] that generates and detects fake news (for text) with much greater accuracy than any other model in existence (92% compared to 73% for the best detection models). In a report dated October 29, 2019, the research team detailed how to defend against fake

[96] Rani Horev, BERT Explained: State of the art language model for NLP, towards data science, 2018

[97] Sadrach Pierre, Fake News Classification with BERT, *towards data sicence*, 2019: https://towardsdatascience.com/

[98] Allen Institute for AI, *Grover- A State-of-the-Art Defense against Neural Fake News,* https://grover.allenai.org/

news[99]. According to the Allen Institute for AI, fake news generators like Grover are the most likely to detect fakes, since they already have examples of fakes. Out of 5,000 examples of fake news generated by its neural network, Grover was able to differentiate human content from that generated by the network with 92% accuracy. The ability of the model to detect fakes increases with the number of examples provided.



Source: Rowan Zellers, *Medium*, 2019

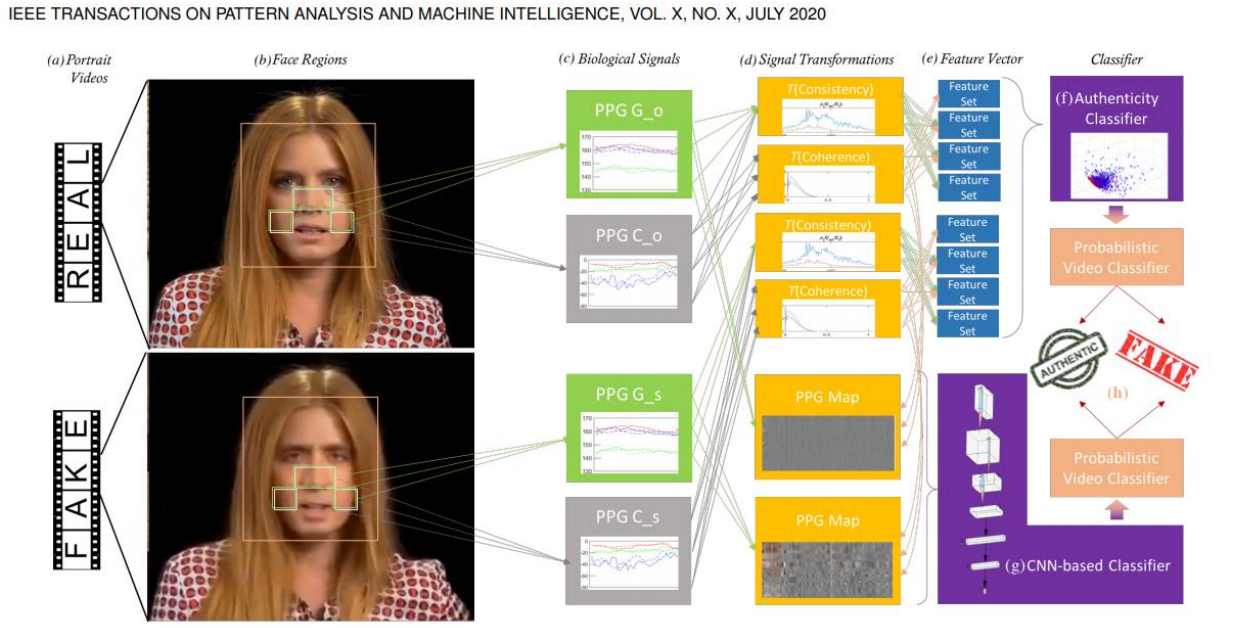
Grover's ability is not restricted to detecting self-generated content, as it is also able to report fake news generated by GPT-2 (before the release of GPT-3), with an accuracy of 96.1%. Grover is also able to detect fake news created by humans[100].

Three researchers, two from Binghamton University and one from Intel, recently published an article in which they described the possibility of detecting fakes in videos, following their creation of FakeCatcher, a software capable of pointing out genuine videos of people through physiological signals which are not visible to the naked eye. Their main claim comes from the fact that biological signals, which are not preserved, neither spatially nor temporally in false content, can be deployed as implicit descriptors of authenticity. In the first step of the process, video segments of fixed lengths with regions of interest within the face are collected. Subsequently, the biological signals are analyzed, including how they change and how they relate to each other. Then the results are used to create a classifier of fake content based on these transformations and the corresponding characteristics in order to create a general authenticity. The traditional classifier is further improved by the generation of new

[99] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi & P. G. Allen, *Defending Against Neural Fake News*, 2019, https://arxiv.org/pdf/1905.12616.pdf

[100] Rowan Zellers, Counteracting neural disinformation with Grover, *Medium*, 2019.

signal maps and the use of a Convolutional Neural Network (CNN), capable of detecting synthetic content. As a final step, the probabilities of authenticity are aggregated and classified allowing to generate a binary decision distinguishing the false content from the real[101].

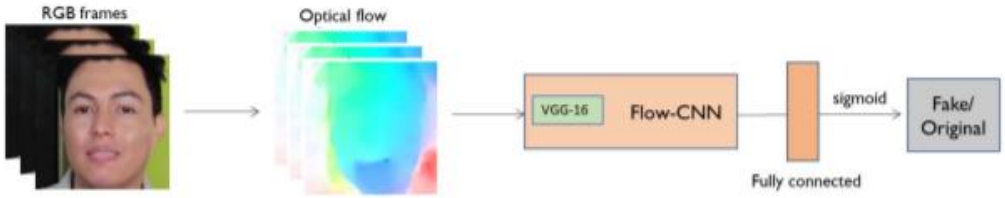


Source: Umur Aybars Ciftci et al., *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020

Continuing in the field of research, an article published in 2019 by the Media Integration and Communication Center (MICC) of the University of Florence and the National Inter-University Consortium for Telecommunications (CNIT) of Parma in Italy, exploited the detection of deepfakes in videos through optical flow applied to convolutional neural networks[102]. The researchers presented a new, sequential approach in order to investigate the existence of dissimilarities between the temporal structures of the videos.

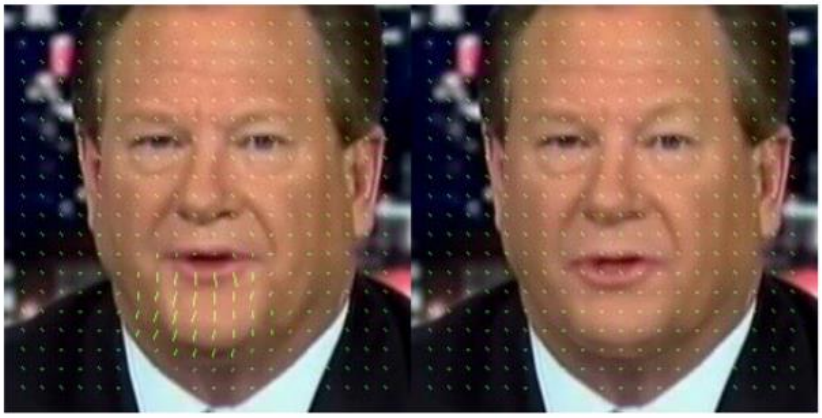
[101] Umur Aybars Ciftci et al., FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. X, NO. X, 2020

[102] Irene Amerini et al., Deepfake Video Detection through Optical Flow based CNN, 2019.



Source: Irene Amerini et al., Deepfake Video Detection through Optical Flow based CNN, 2019.

Optical flow is the apparent movement on a visual scene, objects or surfaces. It is caused by the relative movement between the scene itself and an observer (the eye or the camera, for instance). The objective of their study was therefore to use optical flow to extract the disparities between two images in order to discriminate between those created synthetically and those which belonged to real videos (unusual movements in the face at the level of the lips or eyes). Therefore, the optical flow represented a vector calculated over two consecutive frames in order to retain the movements between the observer and the scene.



Source: Irene Amerini et al., Deepfake Video Detection through Optical Flow based CNN, 2019.

The results of the calculated optical flows are shown in the figure above. We can observe the flows in the original video (on the left), and those in the deepfake (on the right). The noise around the chin is much more visible in the real video image than in the modified one.

Proposition No. 2
Invest in scientific research on content generation algorithms to be better able to combat them but also to be able to use them in positive applications for society.

However, keep in mind that the more algorithms evolve to detect false content, the more they can be used to generate realistic and undetectable content.

Much work is still in progress, and this section is not exhaustive.

Proposition No. 3
Foster European partnerships, pool resources to provide the means to catch up with China, Russia and the United States in this area.

Private sector

In the area of cyber defense, the ability to detect fake content is becoming a major issue. The company Darktrace, a British unicorn founded in 2013 specializing in cyber defense, recently launched a national advertising campaign aimed at raising public awareness of the issues surrounding images generated by AI.

Darktrace advert, September 2020

This British company, valued at nearly $ 2 billion, applies the principles of AI to the field of cyber defense by developing algorithms capable of detecting numerous computer attacks such as internal attacks, latent vulnerabilities, threats to slow development and state espionage. Its main technology is based on Bayesian probabilities and modeled on the human immune system[103].

Another example is the French startup BusterAI, founded in 2019 by Julien Mardas and Aurélien Cluzeau. Given the amount of critical information disseminated online, they aim to provide AI-based tools to improve the quality of content published online. BusterAI develops algorithms intended to detect fakes in images, texts and videos to assist numerous media outlets, including French television channel TF1[104]. They provide their clients with an API so that the companies are able to detect fake content. The company provides a dashboard for diagnosing content, generating trust scores, identifying bias, or even analyzing emotions.

[103] Darktrace Immune System White Paper, https://www.darktrace.com/en/resources/wp-platform.pdf?utm_source=darktrace&utm_medium=products-enterprise

[104] Challenges.fr, Buster AI: l'antivirus de l'information, *Challenges.fr*, 2020

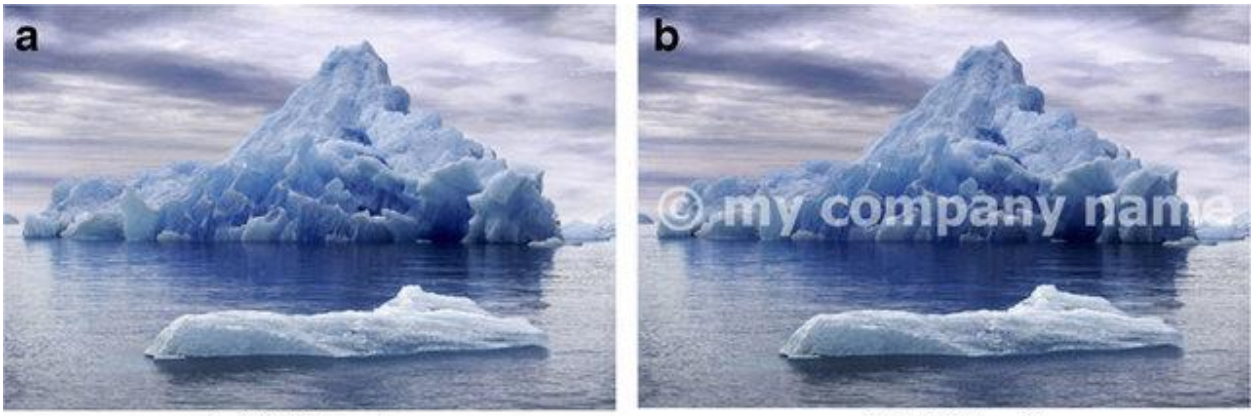
4.2. Tools to protect what's real

4.2.1. Watermarks

Failing to perfectly identify fake content, it is necessary to implement means to authenticate and certify genuine contents. Here watermarks prove to be an essential tool, especially for video and photography. These areas are indeed exposed to piracy issues, but also to content modification. When photos and videos are produced, they can be authenticated through visual (watermarks[105]) or digital (metadata[106]) marking. Metadata tagging can easily be changed with simple software or by taking a screenshot that overrides the metadata. Photographers can use services like Google Image Search or Digimarc to monitor the use of those images.

There are several types of watermarks*:*

- Visible or perceptible watermark: a visible or otherwise perceptible modification of a file. For instance by adding © to the author's name to a photo.



Examples of invisible (left) and visible (right) watermark

[105] A **watermark** is a technique for adding copyright information or other verification messages to an audio, video, image or other digital file or signal.

[106] **Metadata** is data that provides information about other data. Examples of metadata include the size of an image, its creation date and its name.

Source: Sherine M. Youssef et al., 2014[107]

- Invisible watermark: modifies a signal systematically but is imperceptible to the user. This form of watermark can be considered a form of steganography and can be as simple as adding an element like a pixel. It is possible to add invisible watermarks in a photo, video or even in an audio file.
- Fragile watermark: an invisible watermark which becomes visible when the source file has been altered. This helps ensure the authenticity of a file. Universal Music Group is known to apply watermarks to its music files[108].

There are different ways to overcome watermarks: the passive attack which only detects the presence of a watermark, and the active attack which tries to eliminate the watermark. Passive attacks are used in steganography, while an active attack is generally malicious (attempted infringement of copyright, fingerprinting). An active attack can potentially be carried out using a content generation algorithm, which will wipe out watermarks systematically. The aim of an active attack is to use an image, for instance without the knowledge of its author, or to anonymize a document.

Watermarks are not foolproof, whether they are visible or not. Visible defense techniques often have limited effectiveness because they alter the quality of the file. An invisible watermark makes it possible to trace the existence of a file and to guarantee its integrity. It is important to know how to insert a watermark to a file to make it as secure as possible.

However, watermarks aren't just limited to pictures. Indeed, to fight against voice hacking, American company Adobe is actively working on the development of a digital watermark that will reveal if its voice creation software was used.

In order to guarantee the integrity of files, authors (photographers, videographers, etc.) and digital companies developing relevant software (e.g. photo editing software such as the

[107] Sherine M. Youssef et al., Adaptive video watermarking integrating a fuzzy wavelet-based human visual system perceptual model, *Multimedia Tools and Applications*, 2014

[108] Hydrogene Audio, 2019, https://hydrogenaud.io/index.php?topic=117462.0

Adobe suite) are called upon to innovate in order to allow more security and to enforce copyright in particular.

There are two main techniques for marking files: spatially and spectrally. The first is to integrate an element at a specific location in a file. In the case of an image, spatial marking does not hold up well against zoomed-in or cropped sections of an image. However, it is also possible to mark a file with a spectral watermark. This modifies the file in depth (patterns are added within the image without altering it). In this specific case, the whole file is modified in a small way but following the same pattern in order to "encode" a specific pattern. With this pattern, the integrity of the file can be verified by decoding the pattern later.

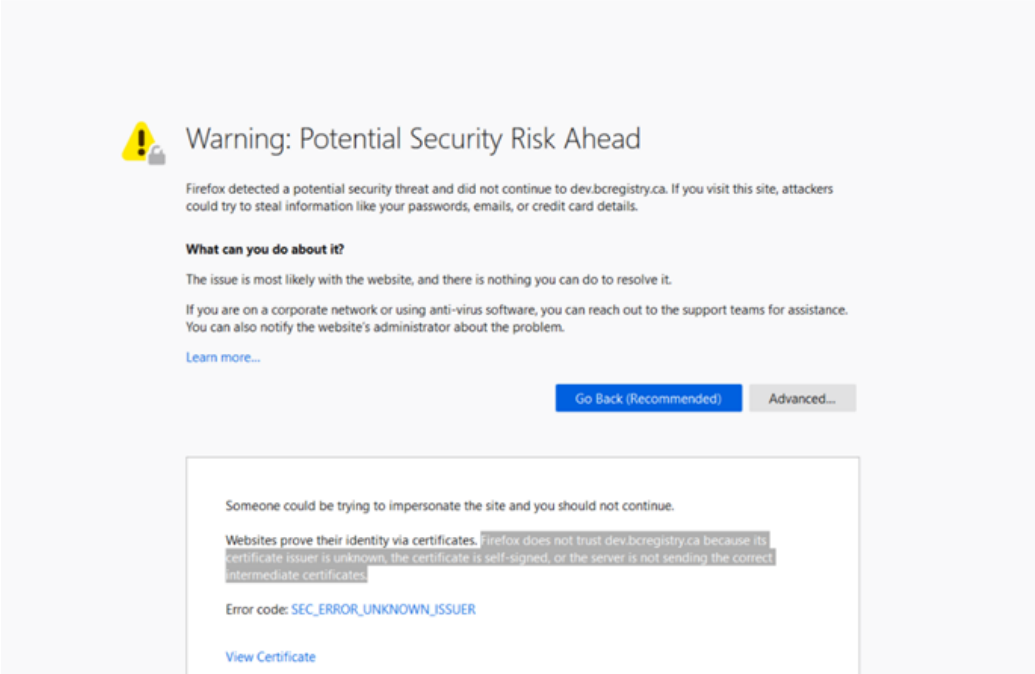
Marking a file at the moment it is created allows to identify the owner of the rights of a work. Marking a file at the moment it is sent or read identifies the recipient or the reading equipment (fingerprint). This can be very useful in detecting a user who is illegally sharing the content. This technique has been used in particular to tag video content and identify users of paid platforms who shared the content on the Internet (streaming). This is for example the case for live broadcasts of boxing matches by HBO.

4.2.2. Digital signatures

Watermarks are useful for tracking content, but another well-known tool can be very useful for authenticating an author: digital signatures, or advanced electronic signatures under the terms of European law.

A digital signature is a unique mathematical code generated from a document and a key (known as private or secret). A twin key (called a public key) will allow anyone in its possession to verify the integrity of the document and the identity of the author. The public keys can be gathered to form a web of trust, public or private, which then constitutes a third party guarantor of identities.

This solution is widely used in IT and on the web to certify the origin of a site. For instance, the display of this message indicates that the visited site is not reliable because its origin could not be verified.



It is a simple but effective mechanism, imposed in France by the eIDAS regulation[109] in 2014, and which allows digital certificates to be given to individuals.

There are three levels of certificate: simple, advanced and qualified with increasing legal validity, depending on use. A simple signature to sign an expense report up to a qualified signature for an authenticated deed at the notary's.

An electronic signature can therefore be useful to check the identity of a sender.

4.2.3. Hash

A hash function is used to generate a digital fingerprint, also called a checksum, in order to identify data. Hash functions are used in computing to authenticate a password or a file.

[109] https://www.ssi.gouv.fr/entreprise/reglementation/confiance-numerique/le-reglement-eidas/

By storing the digital fingerprint of a password, it is possible to avoid storing unscrambled passwords in a database, to ensure confidentiality. They can then be verified by comparing the fingerprints stored in the database with the ones generated from the password entered by the user during subsequent authentications.

Likewise during a file transfer, generating the digital fingerprint of a file before sending, and then generating it again after receipt of the file, allows to ensure the authenticity, and that it has not been altered during transfer. This comparison is necessary in particular in the event of a transfer error, but also to guard against piracy.

Here is an example of a generated digital footprint of a photo available on the Ekimetrics website from the CertUtil utility, available in Windows 10:

In this example, the fingerprint is generated by a SHA-256 encryption algorithm, considered cryptographically secure:

"476aa185797ded1477b45cf47ac7ff381a59194882ba37b6276c46e9dc32ee42"

Hashing therefore makes it possible to ensure the authenticity of a file by verifying its integrity between two steps.

4.2.4. Blockchain

Other innovative solutions such as blockchain are emerging to ensure file integrity.

As detailed in the Praxis Medical Data Report, blockchain technology consists of a network of data records (also called blocks) distributed to each participant. The blocks are chronologically linked to each other. In a blockchain, whether public, private or hybrid, new data can be added to its records via a consensus mechanism, such as mining. Associated with a digital signature mechanism, this consensus could be agreed upon beforehand on the identity of the author[110].

Blockchain has many theoretical advantages:

- **Decentralization**: historically, data records have been centralized by a trusted intermediary (a bank, a hospital, a ministry, etc.) which ensured the security and use of the data. Blockchain and records distributed over the network provide the necessary trust in a data exchange system.
- **Security**: in a blockchain, data records cannot be altered without the consensus of the network. This is the main strength of the system which, unlike a centralized registry on a server, cannot be hacked. On the other hand, it is possible that the password of a user, giving access to his private key, is stolen by fraudulent means[111]. The identity of the victim can then be usurped.
- **Portability**: the authenticity of a participant's data can be verified by several actors if they authorize it. Recording data on a blockchain would allow several entities with contradicting database entries (hospital, pharmacy, etc.) to ensure better traceability of a patient.

Computer security company Amber[112], and its CEO Shamir Allibhai, has developed a technique to ensure the integrity of videos from surveillance cameras and pedestrian cameras (used by the police) using blockchain. This follows the work of security researcher

[110]Healthcare Data Institute, INTERNATIONAL THINK TANK DEDICATED TO BIG DATA IN HEALTHCARE, 2017, https://healthcaredatainstitute.com/wp-content/uploads/2015/02/presentation-hdi_04122017.pdf

[111] Jeff John Roberts, *Bitcoin Theft: 5 Common Threats, Hacks, and Scams*, 2017, http://fortune.com/2017/12/08/bitcoin-theft/

[112] Lily Hay Newman, *A New Tool Protects Videos From Deepfakes and Tampering*, 2019, https://www.wired.com/story/amber-authenticate-video-validation-blockchain-tampering-deepfakes/

Josh Mitchell, who has proven that there are serious security holes in the majority of cameras used by U.S. police. The system relies on the blockchain principle to authenticate video fragments: the platform generates "hashes" (a unique mathematical signature that represents the video fragment) at regular intervals and shares them to a public blockchain. By replaying the video and passing it through the same algorithm, the hashes will be identical (and therefore comparable to those recorded on the blockchain). On the other hand, if the video has been modified, the signature will be different, which therefore protects against manipulation of the video itself.

The computer security company Amber has developed a technique to ensure the integrity of videos from surveillance cameras and pedestrian cameras (used by the police) using blockchain.

Blockchain is a technology for storing and transmitting information, transparent and secure, which operates in a decentralized manner (without a central control body). By extension, a blockchain constitutes a history of all the exchanges carried out between its users since its creation. This history is secure and distributed: it is shared by its various users, without an intermediary, which allows everyone to check the validity of the string. As the mathematician Jean-Paul Delahaye writes, you have to imagine *"a very large notebook that everyone can read freely and for free, on which everyone can write, but which is impossible to erase and indestructible"*. However, we know that there are several examples showing that it is not impossible to reconstruct the whole notebook with immense computing power.

Blockchain is a decentralized alternative to content authentication, which acts as a record connected to a network and secured by the computing power of the network. As with a

watermark, its use would not detect false content, but guarantee its integrity. The files (and their digital signature) inserted into the blockchain have a unique identifier. The blockchain is very useful for authenticating a modification or transaction without necessarily knowing the author.

Computer attacks against the blockchain are very difficult to carry out because an actor or a group must control more than 50% of the guarantors who make up the blockchain in order to corrupt its content. The blockchain would act as a decentralized trusted third party and would ensure the integrity of content.

The implementation of an authentication system requires an image marker that is robust enough to withstand image layout manipulation, and therefore relies a priori on a watermark type system.

We believe it is necessary to take advantage of these technological advances to protect the public. Scientific research is extremely prolific on these essential topics, and the technological tools of watermarks, electronic signatures and blockchain need to be understood well in order to guarantee the integrity of online content.

4.3. Awareness and education

Insofar as the various laws and regulations, including the policies put in place by the platforms, are safeguards but will never be sufficient to guard against the harmful effects of the dissemination of fake news or deepfakes, the development of critical thinking and raising public awareness of these themes is, in our opinion, a major issue.

We suggest setting up a general public awareness campaign against deepfakes. For example by highlighting, in documentaries or training courses, simple tips for detecting false images when the false image is visually detectable. Even though many deepfakes are no longer detectable with the naked eye, there are still many low-quality deepfakes. This type of

training makes it possible to educate the public about deepfakes in a light-hearted way. It seems essential to us to teach methods to protect oneself from this malicious content, and to raise awareness of the impact that this false information can have when it is widely relayed.

Proposition No. 10
Make the fight against deepfakes a European objective for 2021. Set up awareness campaigns to explain and educate about deepfakes, their uses and the associated risks in terms of misappropriation of images, voices and videos, and emphasize the possible applications to manipulate public opinion.

It is also possible to go further and educate the population on how algorithms work. Jean-Baptiste Bouzige, founder and CEO of Ekimetrics, believes that poor understanding of algorithms and the challenges posed by these new tools is at the heart of the problem. It is necessary to build European expertise on the subject and to ensure that we train researchers who will be at the forefront of this subject. In France, computer programming establishments such as the Le Wagon company or even École 42 founded by Xavier Niel have been very successful.

In the education sector, new tools are being developed, such as Scratch, a programming language developed by MIT that allows children in particular to understand how an algorithm works and to "play" with it. It seems important to us to consider these tools to educate the new generations.

However, deepfakes are partly undetectable, and unfortunately everyone can be a victim, including those who are aware of them. It is therefore essential to also invest in more

advanced technologies and techniques capable of detecting even the most advanced deepfakes that it is not possible to detect otherwise.

4.4. Role of authors

The most basic way for a user to know if the content in front of them is genuine or if it has been manipulated using an algorithm would simply be for the author of the content to report if it is. This reporting may be voluntary or enforced, whether by the terms of use of the platforms or by law.

At the time of writing, China is one of the countries with the most stringent deepfake protection laws. Indeed, a law in force since January 1, 2020 stipulates that *"the publication of deceptive video content such as 'deepfakes' created from artificial intelligence must be reported by whoever posted it. [...] With this regulation, the authorities can prosecute not only simple users, but also image and video hosting services if they fail to report deepfake content."* This law, however, presupposes the existence of detection technologies in order to be able to sanction a breach of the obligation to report.

The case of China appears to be isolated at this time, and no other country to our knowledge imposes a reporting obligation on deepfakes. Neither do platforms, which adopt other approaches. As mentioned above, Facebook removes content that it detects as deepfake (except overtly humorous videos) while Twitter wants to report deepfake content that it detects on its part.

However, this posture presupposes a reliable technical means to identify content manipulated using an algorithm. While waiting for the deployment of these technical means (some already exist), users can protect themselves from deepfake content on their own, as some details can indeed betray the manipulation of content.

Proposition No. 11

Increase the authors' responsibilities. Require authors to report any deepfake content.

4.5. Role of platforms

The source of deepfake documents can be difficult to locate, and it can be difficult for authorities to hold accountable and penalize deepfake producers.

In order to achieve this, it is important that the platforms cooperate with the authorities and provide them with any information they have in the context of investigations or legal proceedings.

Proposition No. 6
Require platforms to share with judicial authorities any information they hold on alleged deepfake perpetrators, in the context of criminal investigations, without the option of a legitimate reason for refusal.

Deepfake consumers often participate in the redistribution of such items unknowingly. Without explicit notices that the items are deepfakes, redistributors are not necessarily responsible. The platforms therefore share responsibility and have an important role to play in limiting the harmful effects of deepfakes.

Proposition No. 4
Increase the platforms' responsibilities (for example within the Digital Services Act). Require each platform to set up an internal governance structure for the regulation of their content. Place each governance body under European supervision, drawing inspiration from the regulations put in place in the banking sector following the 2008 crisis.

The platforms are able to set up several tools to fight against deepfakes:

- A penalty for authors publishing false or malicious content in a non-explicit manner (examples of penalties: limitation of posts, temporary limitation of the audience for their posts, implementation of a validation process for the suspicious authors, etc.).
- A clear statement warning the user that the post could not be verified, or even that it is suspicious.
- A limit on the number of recipients of a post.

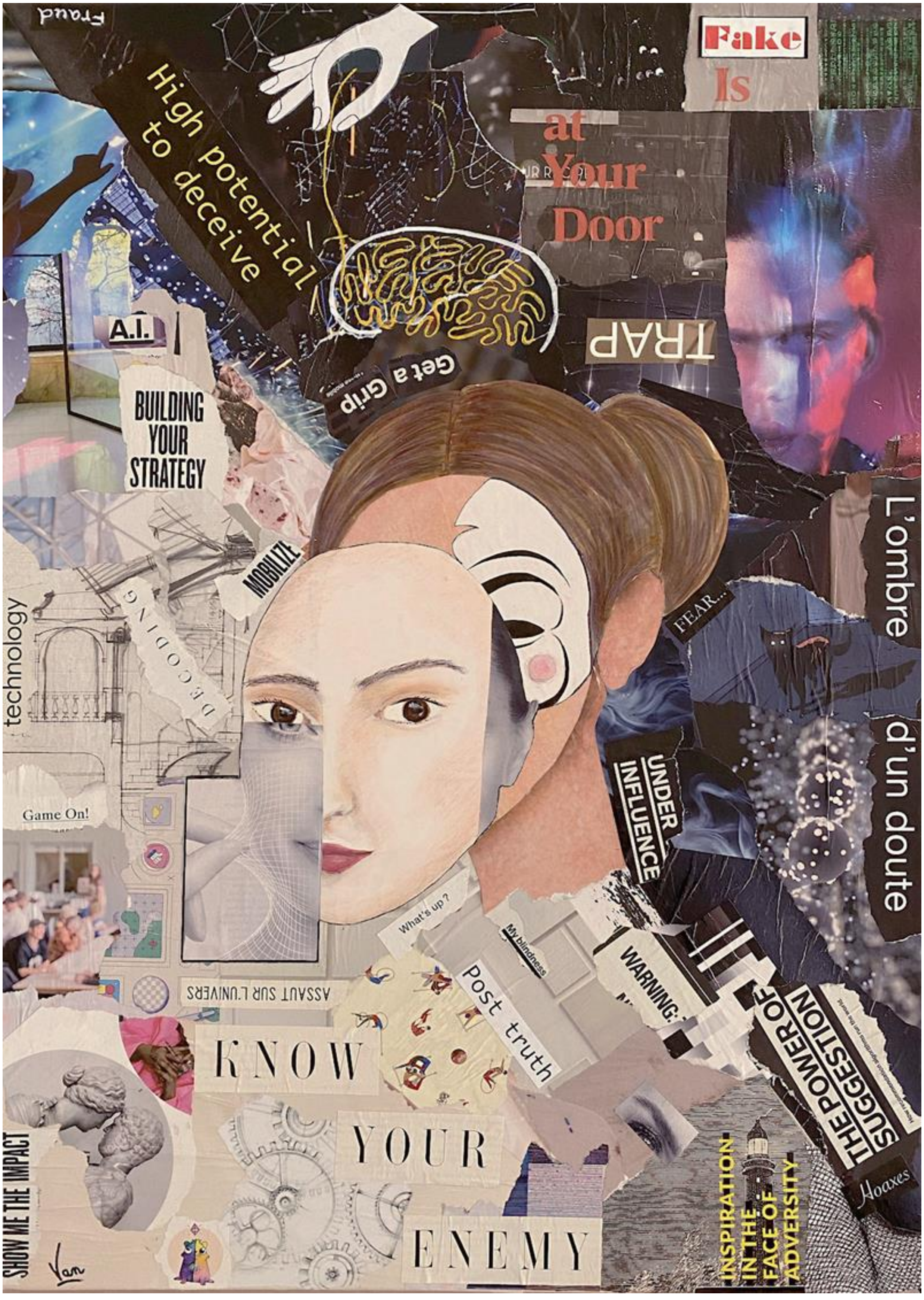
For example, WhatsApp already limits the number of people to whom it is possible to forward a message, as well as the size of groups. In France, group size is limited to 256 people. Since January 2019, to combat fake news, WhatsApp has limited the forwarding of messages to five people (down from 20). WhatsApp also indicates if a message has been forwarded to allow users to better forward these viral messages.

The Telegram app highlights its implementation of a slow mode that prevents users from sharing more than one item per unit of time (for example per minute). However, this is an optional mode that can be selected by the group administrator. Also, the size of a Telegram group can be up to 200,000 people.

Proposition No. 5

Encourage platforms to put in place tools to reduce the impact of deepfakes.

- A penalty for authors posting false or malicious content in a non-explicit manner.
- A clear statement warning the user that the post could not be verified, or even that it is suspicious.
- Limiting the number of recipients of a post.
- Limiting the number of times a post can be forwarded by a user.


Fraud
High potential to deceive
Fake
Is
at
Your
Door
TRAP
Get a Grip
A.I.
BUILDING YOUR STRATEGY
MOBILIZE
DECODING
technology
FEAR...
L'ombre d'un doute
UNDER INFLUENCE
Game On!
What's up ?
My blindness
Post truth
WARNING:
ASSAUT SUR L'UNIVERS
THE POWER OF SUGGESTION
KNOW
YOUR
ENEMY
SHOW ME THE IMPACT
INSPIRATION IN THE FACE OF ADVERSITY
Hoaxes
Van

Conclusion

At a time when, in France, social networks are the main source of information for those under 34, where in Burma the Facebook app is, according to the *New York Times*, pre-installed on all phones and accused of inciting hatred towards the Rohingyas, the impact of false content is more relevant than ever. Especially since the creation of these deepfakes, created by deep learning algorithms (artificial intelligence algorithms inspired by the functioning of the brain) has come within everyone's reach and their dissemination is very often viral.

The objective of our report was threefold. From an educational point of view, it aims to raise awareness of the issue represented by the uncontrolled development of content generation algorithms; from a political point of view, it wants to influence the debate around the digital, regulatory and legal framework for deepfakes as well as the economic and social debate at European level; and finally, from a scientific point of view, it sets out to enable Europe to position itself as a leader in the fight against deepfakes.

Content generation algorithms, like any new technological advance, crystallize many fears. However, it is not a question of banning the technology at the origin of deepfakes, not only because it is not within the reach of any government to permanently hinder the progress of science, but also because these techniques can have useful applications in vast fields which range, as we have seen, from the improvement of epidemiological predictions to disability assistance, including entertainment and artistic creation.

Appendices

Appendix 1: Deepfake Detection Challenge (DFDC)

The generation of content, especially that of deepfakes, has gone from a so-called "sophisticated" status, to a level where anyone is capable of generating it.

In 2019, this collaborative initiative between several leaders in the world of technology as well as in the academic world, was launched around the subject of deepfakes. Thus, the Deepfake Detection Challenge (DFDC) brings together AWS, Facebook, Microsoft and several other academic institutes including MIT with the aim of encouraging researchers to develop new technologies that make it possible to combat the problems of deepfakes.

- **Dataset construction**

A dataset consisting of more than 100,000 videos was created, featuring eight face modification algorithms, such as GANs and refinement methods, in particular overlays and the inclusion of distractors.

Attracting more than 2000 individuals, all participants had to train and test their models on the same dataset. These models were later tested against a black box dataset, made up of new videos and not shared with the participants in advance. Make-up or painting tutorial videos were used to add further complexity to the detector models. Augmentations were also applied to the entire black box dataset, and additional techniques were applied to the tests to increase the level of difficulty.

Source: Devin Coldewey, *The Crunch*, 2020 (Credits: Facebook)

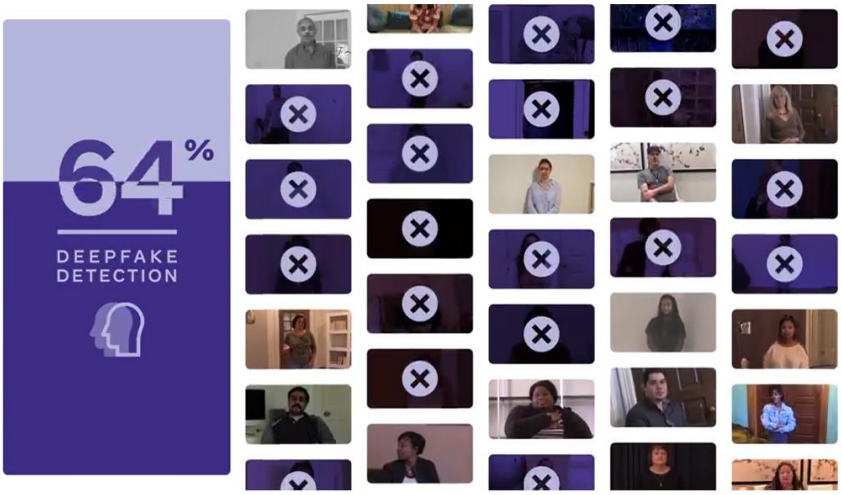
It is sometimes difficult to immediately detect false content with the naked eye. Take, for example, these six videos generated for this competition. Only images 1, 4 and 6 represent real content.

For Mr. Schroepfer, the CTO at Facebook, the variance in the dataset generated for this challenge plays a key role: *"If your dataset does not have the appropriate variance in the appearance of real people, your model won't be able to process this diversity adequately. I think we've gone to great lengths to make sure that this dataset is sufficiently representative"*. To emphasize the importance of representativeness, Mr. Schroepfer adds that several factors, such as age, ethnicity and gender, have been taken into consideration. For him, the developed technology must be able to represent everyone.

- **Competition results**

The first model that was selected, with an accuracy of 85.26%, failed when it was tested against the black box dataset. This shows that the detection of deceivingly lifelike images

is a serious problem. Despite promising results, Facebook admits that the detection of deepfakes remains an unresolved issue.



The most efficient model achieves an accuracy of 65.15%.

Source: Facebook AI

After declaring his satisfaction, Mr Schroepfer announced that Facebook was developing its own technology to deal with the issue of fake content. Nevertheless, the winning algorithms of this competition are available open-source, thus helping other researchers working on this topic. In contrast, Facebook announced that it would keep its own detection technology a secret to prevent it from being reverse engineered.

Appendix 2: The seven ethical principles of the European Commission in the field of artificial intelligence[113]

"Ethics Guidelines for Trustworthy AI" is a document prepared by the High-Level Expert Group on Artificial Intelligence (AI HLEG). This group of independent experts was set up by the European Commission in June 2018.

Based on fundamental rights and ethical principles, the guidelines list seven key requirements that AI systems must meet to be reliable. These requirements apply to the different stakeholders involved in the life cycle of AI systems:

- **Developers**, i.e. those who research, design and / or develop AI systems;
- **"Deployers"**, which are public or private organizations that use AI systems in the course of their business and to deliver products and services to others;
- **The end users**, who engage with the AI system, directly or indirectly;
- **Society** as a whole, which includes all other people who are directly or indirectly affected by AI systems.

These different stakeholder groups have different roles to play in ensuring that the requirements are met:

- Developers must implement and apply requirements to design and development processes;
- "Deployers" must ensure that the systems they use and the products and services they offer meet the requirements;
- End users and society at large should be made aware of these requirements and be able to request that they be met.

[113] For more information see:
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Human%20agency

Below is a more detailed explanation of the seven key principles set out by the European Commission.

Human action and human control

AI systems must support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This implies that AI systems must both act as facilitators of a democratic, thriving, and equitable society by supporting user action and promoting fundamental rights, and enabling human control.

Robustness, technique and safety

Technical robustness, which is closely linked to the principle of damage prevention, is an essential element in achieving trustworthy AI. Technical robustness requires that AI systems be developed with a risk-preventative approach and in such a way that they behave in a reliable and predictable manner while minimizing unintended and unexpected damage, and preventing unacceptable damage. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in antagonistic ways. In addition, the physical and mental integrity of human beings must be guaranteed.

Respect for privacy and data governance

The right to privacy, closely linked to the principle of prevention of harm, is a fundamental right particularly affected by AI systems. The prevention of privacy breaches also requires adequate data governance which covers the quality and integrity of the data used, their relevance to the domain in which the AI systems will be deployed, their access protocols and the ability to process data in a way that protects privacy.

Transparency

This requirement is closely linked to the principle of explainability and encompasses the transparency of the elements relevant to an AI system: the data, the system and the business models.

Diversity, non-discrimination and equity

In order to achieve trustworthy AI, we need to enable inclusion and diversity throughout the life cycle of the AI system. Besides the consideration and participation of all relevant stakeholders throughout the process, this also involves ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked to the principle of equity.

Societal and environmental wellness

In accordance with the principles of fairness and harm prevention, society at large, other sentient beings and the environment should also be considered as stakeholders throughout the lifecycle of the AI system. The sustainability and environmental responsibility of AI systems should be encouraged, and the search for AI solutions in areas of global concern, such as the Sustainable Development Goals, should be encouraged. Ideally, AI systems should be used for the benefit of all human beings, including future generations.

Responsibility

The accountability requirement complements the above requirements and is closely related to the principle of equity. It requires mechanisms to be put in place to ensure accountability for AI systems and their results, both before and after their development, deployment and use.

Appendix 3: European Commission, *White Paper on artificial intelligence. An approach focused on excellence and trust*, 2020

In this white paper, the European Commission reports on the priority axes selected for the future supervision of AI. It thus intends to position the European Union as a world leader in the field of AI.

For this, the Commission considers that an ecosystem of trust is necessary, which implies a regulatory framework adapted to artificial intelligence. The latter is already subject to European legislation on fundamental rights, consumer protection and product safety and responsibility. However, some components of AI (such as opacity) can make the enforcement of these laws ineffective. The Commission further underlines that while AI benefits will be brought in, it can also harm both materially and not materially.

On the dangers linked to fundamental rights, the Commission considers that they may result from flaws in the overall design of AI systems or from the failure to correct for bias. Bias and discrimination are inherent risks in any social or economic activity and human decision-making is not immune to such errors and biases. However, the same bias, when present in AI, could have a much larger effect affecting and discriminating against many people without the social control mechanisms that govern human behavior. Such errors and biases are also likely to appear when the AI system is learning, which can lead to errors in the AI results. In addition, the opacity of AI increases the difficulty of verifying its compliance with existing regulations. Law enforcement authorities and those affected may not have the means to verify how a decision is made and whether basic rules have been followed.

On the security risks and the functioning of the liability regime, AI also presents risks for user safety, when it is integrated into products and services. The European Commission takes the example of self-driving cars, where the incorrect identification of an object on the road can cause an accident resulting in injuries and material damage. The safety risks, even though they are not limited to the AI, can be enhanced with this technology, and can be caused by defects in the design of AI. When these dangers materialize, the challenge will be to prove

the defectiveness of the AI-based product that caused the crash. Being able to understand and trace the decisions taken by the AI that led to the incident is then necessary, in particular to allow potential victims to obtain compensation for the damage suffered.

To take account of these new potential risks, the Commission considers that the legislation should evolve, to deal with the following situations:

- ensure effective application and enforcement of existing Community and national legislation;
- provide a framework for limiting the scope of existing Community legislation;
- take into account the evolution of the functionalities of AI systems;
- remove uncertainty about the distribution of responsibilities between the different economic operators in the distribution chain;
- develop the concept of security.

In order to back it up with a relevant legal framework, it is therefore important to define precisely the scope of this strategy, given the many myths floating around the very concept of artificial intelligence and the lack of scientific consensus on the concept. However, while the Commission insists on the requirement of a definition *"sufficiently flexible to take account of technical progress but precise enough to guarantee the necessary legal certainty"*, it does not intend to create a new framework, which would place a disproportionate burden on SMEs, especially. It therefore recommends that the new regulatory framework only apply to AIs deemed to be "high-risk". A classification that would be based on two cumulative criteria:

- if AI is used in a sector where, taking into account the characteristics of the activities generally carried out, one can expect significant risks, in particular as far as the protection of safety, consumer rights and fundamental rights are concerned. This first criterion ensures that regulatory intervention is targeted at those areas where risks are most likely to occur. The sectors covered will have to be listed in a specific and exhaustive manner in the new regulatory framework. For example, health care,

transport, energy and areas of the public sector. The list should be periodically reviewed and amended if necessary in light of relevant developments in practice;

- and if the use of AI involves significant risks. The assessment of the level of risk of a given use could be based on the impact it has on the parties involved. For example, uses of AI that affect the rights of an individual or a company by carrying a risk of injury, death or significant material or immaterial damage.

Applying these two cumulative criteria would ensure legal certainty, as the area of application would be clearly defined and proportionate. But the design of a future regulatory framework for AI would also require defining the legal obligations of the actors concerned by high-risk AI. It is to that end that the Commission has drawn up a list of obligations that operators need to adhere to in the case of high risk AI:

- ensure that the data sets used for training are exhaustive and representative enough to consider all possible scenarios, limit the occurrence of risks, and avoid discrimination;
- keep a register of the data used to train the algorithms, documenting the characteristics of the data sets, the choices made, design and training techniques and methodologies;
- provide certain information, in particular on the use of AI and its degree of trustworthiness;
- ensure the robustness and accuracy of the data by monitoring the risks of a given AI from the start and keeping an error log throughout the AI's life cycle;
- use of human supervision: making sure the results of the AI are verified by a human, thereby allowing intervention in real time and the possibility to disable the AI.

The European Commission recommends that these obligations be supported by the actor in the chain most likely to be affected by these potential risks.

Appendix 4: Approach and conclusion of the CNIL report, 2017[114]

- **Approach**

The 2016 Digital Republic Bill entrusted the National Commission on Informatics and Liberty (CNIL) with the task of leading the debate on ethical and societal questions raised by the rapid development of digital technologies.

In 2017, the CNIL decided to focus these discussions on algorithms in the age of artificial intelligence. Unbeknownst to us, these are creeping into every corner of our life: results of search engines on the web, financial orders placed by robots in the markets, automated medical diagnoses, distribution of student candidates at universities. In all these areas, algorithms are at work. In 2016, the subject of algorithms rushed onto the public debate scene in an unprecedented way, generating wide media coverage (questions on the algorithm of the online university admission portal "Post-Bac Admission"in the Île-de-France administrative region, the use of artificial intelligence in Trump's election campaign strategy, the role of social media in spreading fake news).

Ethical reflection concerns decisive societal choices. It should not be done without taking into account this pluralist and collective dimension, especially when it concerns such a transversal question, which touches all aspects of our social and personal life. It would simply not be possible to bring together in a single committee all the skills and perspectives necessary to examine the questions raised by algorithms in sectors as diverse as health, education, marketing, culture, defense and security, for example.

Also, rather than directly leading centralized discussions on these subjects, the CNIL has decided to adopt an original position as the leader of an open and decentralized national public debate. During a launch organized on January 23, 2017, it called on all interested actors and organizations (public institutions, civil society, companies) to organize a debate or an event on the subject, which they would then report to the CNIL. The objective was to

[114] For more information see : https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

collect from the actors in the field the ethical issues identified to date as well as their ideas for addressing them.

60 partners presented themselves in response to the appeal launched by the CNIL, from very different sectors and representing various structures. Among them, we can cite the Ligue de l'Enseignement, the French Insurance Federation (FFA), the Ministry of Culture (DGMIC), Open Law (association for reflection on justice) as well as unions such as the CFE -CFC and FO Executives (for recruitment and HR), etc.

They organized 45 events between March and October 2017 in several French cities (as well as abroad thanks to the Future Society at the Harvard Kennedy School), at which some 3,000 people participated. The CNIL ensured a global and coherent coordination of events.

The events organized for the public debate were also an opportunity to raise awareness of issues of utmost democratic and civic importance, not just to experts but to all of us.

Conclusion of the CNIL

The principles and recommendations set out at the end of this report are the result of the synthesis that the CNIL has drawn up from the discussions carried out during the national public debate it hosted from January to October 2017 with the help of 60 partners.

The policy recommendations were expressed very broadly, calling on the greatest possible number of public and private actors. Faced with the challenges posed by algorithms, the whole of civil society (the public, companies and associations) must get involved, be attentive and question themselves in a complex world. The intention was therefore not to say that the law could be the only appropriate means of applying them. On the contrary, most of the recommendations could be understood either through a binding legal framework or through voluntary adoption by stakeholders.

The discussions highlighted two founding principles, and as some of the main ethical challenges posed by AI can be grouped under these principles, they deserve special attention.

Firstly, the substantial principle of equity of algorithms, which is based on the principle already proposed by the French Council of State (see the section "The principle of equity"). This version includes the idea of fairness towards users – not only as consumers but also as citizens – and towards communities whose way of life could be affected by algorithms, whether processing personal data or not.

Secondly, a more methodological principle is that of continuous care and vigilance. This is not a vague mantra, but a substantiated response to three central challenges of digital society. Firstly, the changing and unpredictable nature of algorithms in the “era of machine learning.” Secondly, the silo mentality that affects the way algorithmic chains are organised and encourages working in isolation, indifference to the global impact of the algorithmic system and reduced liability. Thirdly, the risk of an over-reliance on machines, which some form of human cognitive bias leads us to regard as fail-proof and unbiased. The principle of continuous attention and vigilance essentially aims to organize the permanent state of alertness that our societies must adopt with regard to the complex and changing socio-technical objects represented by algorithmic chains or systems. This state of alertness implies a constant subjection to careful scrutiny and critical thought. This first and foremost concerns individuals, who are the link in the algorithm chains : they must be given the means to keep a lookout in a conscious and active manner, while always seeking answers in this digital society. But it also concerns other key players in our society: companies, of course, to allow them to model virtuous algorithmic systems, as well as others.

Thanks to the universal approach that bore them, these principles could be part of a new generation of principles and human rights in the digital age: a generation which, after those

of rights to freedoms, rights to property and social rights, would be that of "system rights" organizing the dimension that underlies our digital world. Are they not worth defending as general principles of the global governance of Internet infrastructure?

At a time when France and Europe are defining their position on AI, the question is relevant.

Appendix 5: Introduction to AI

Statistical learning is a set of methods used to extract information from a database in order to accomplish a task. The tasks vary, and include classification, transcription, translation, anomaly detection, denoising, etc. The resolution of these tasks revolves around three paradigms: supervised[115] or unsupervised[116] learning, the representation of the signal, and the choice of a generative or discriminative model.

- **Supervised or unsupervised learning**

Unsupervised learning corresponds to the search for structure in any database. This is the most common approach. Having a large number of features and examples at our disposal on the learning target, we look for the essential properties. For denoising or restoration tasks, one can, for example, seek to estimate the probability distribution of a signal.. In this category, we also find clustering, which is the search for groups of similar examples.

Supervised learning assumes that each example is associated with a target category. We guide the learning from examples that contain the categorization (that are"annotated"), in order to obtain a prediction of the target on new examples. Thus, we can train a supervised model from a database of photos of cats and dogs, where each photo is associated with one of the two categories. Historically, the recognition of handwritten digits, acquired from a large quantity of photos of handwritten digits, annotated from 0 to 9, has led to progress in supervised learning, and the first convincing results of deep learning models. The success of supervised approaches therefore depends on the existence of annotated databases, where each example contains the target category. The recognition of objects in

[115] Machine-learning task of learning a predictor function from annotated examples.

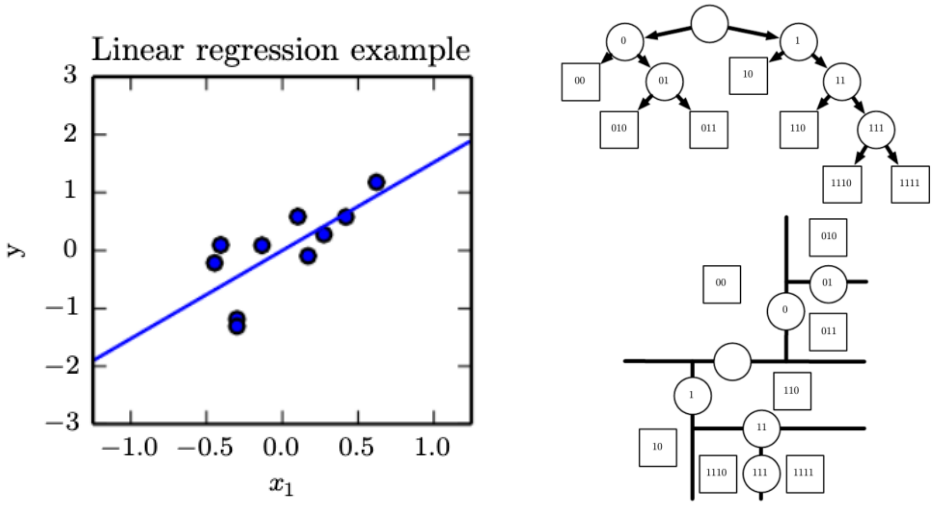
[116] Unsupervised learning involves teaching an artificial intelligence (AI) algorithm information that is neither classified nor labeled, and allowing that algorithm to react to this information without supervision. This type of learning is mainly used to perform classification (clustering).

image processing is done thanks to the provision of gigantic databases: ImageNet, for example, contains 3.2 million images, hierarchically annotated into 5247 categories. The dependence of algorithms on this data has created new professions, thatare sometimes called "AI's microtasks" – see Amazon's Mechanical Turk for example.

In practice, a complete system often oscillates between supervised and unsupervised. This is sometimes referred to as semi-supervised or weakly supervised learning. The acquisition of annotated data is sometimes expensive (some bio-medical data require samples) and always limited in number. Propagating annotations to similar unannotated data, for example, is often a key step.

- **Regression, classification[117], tree, neural network**

Here are some examples of common learning tasks.



Source: *Deep Learning*, Aaron Courville, Ian Goodfellow and Yoshua Bengio

Regression is the prediction of a numeric value based on one or more inputs. In the case of a linear regression, the problem consists in finding the vector of parameters *w* that

[117] Assigning a class / category to each of the observations in a dataset is a classification. This task is done a posteriori, once the data has been recovered.

makes it possible to predict the value *y* according to the inputs *x* by assuming a linear relation *y* = *w.x*. We define the solution as that which minimizes the error between the observations and the prediction. Linear regression is a simple and very limited example, but it helps to understand how a learning algorithm works. We can think of *w* as weights that determine how each entry contributes to the prediction *y*. If the weight *wi* is positive, an increase in the value of the input *xi* implies an increase in the prediction, and vice versa. If the weight of an entry *xi* is zero or almost zero, then the corresponding entry has little influence on the result.

Another example of supervised classification: decision trees. Each node is associated with a region of the input space, and the leaves of the tree break down the space into disconnected sub-regions.

In the specific example above, the tree is at the top of the image and the entry space at the bottom. The nodes of the tree are the circles, and the leaves are the squares. An input value passes through the first node and descends to the leaves. Each node is a binary test that divides the input space in half so that the space is divided into regions by the whole tree, one region per leaf.

The rigidity of the space division obtained limits the generalization capacities of decision trees. Each region requires a full-fledged training data set. The superiority of neural networks comes from the ability to learn from a large set of regions with much fewer points (Bengio and Monperrus, 2005). As opposed to decision trees we can learn 2^k different regions with only *k* training data.

The role of deep neural networks is to assess the transformation function between the inputs and the outputs. They are called neural networks because they are made up of several elementary functions called neurons, which are connected in several layers. The number of layers determines the depth of the network.

- **Detecting structures: representation learning**

The solution for learning problems depends on the chosen representation. An algorithm is not fed with all the information available about a patient, but rather some very specific data, such as the presence of scars, disease, or age. The choice of the variables supplied as input was decisive in the success of the method. Given the computing power of the time, the same algorithm would have been unable to provide the same prediction even from a larger data set, such as images or continuous signal measurements. Feature design is the step of selecting or creating descriptors of the input signals to steer the learning towards the result.

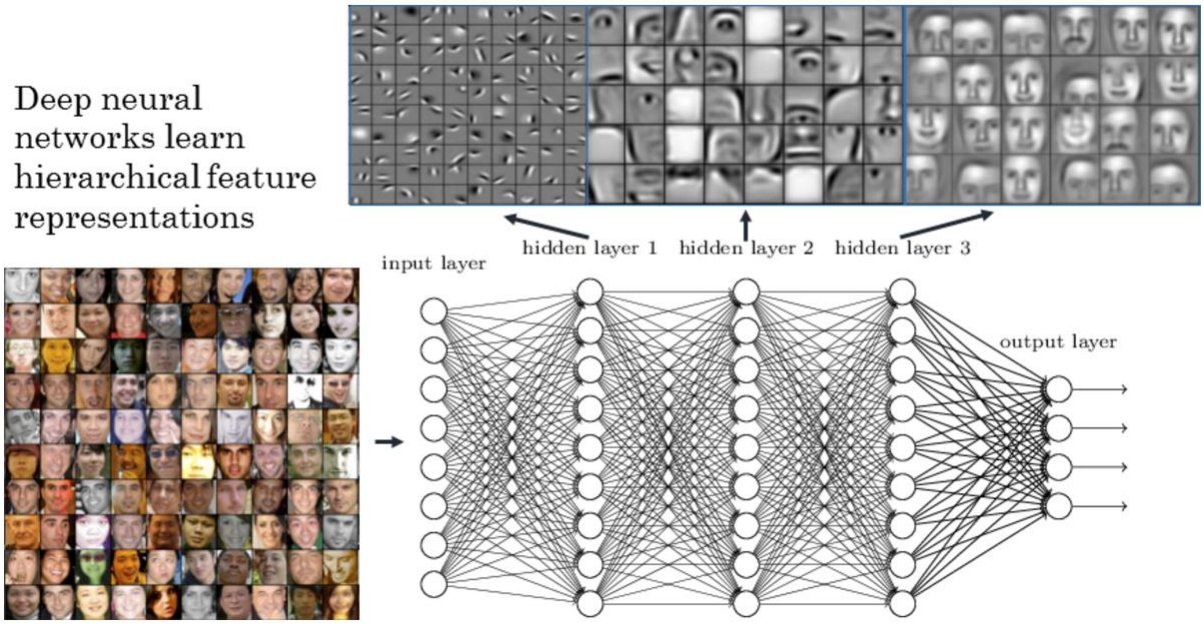
There are many image descriptors: color histograms have been used to detect context (photo non-photo, exterior/interior, nudity). The oriented gradients, by calculating the differences of the pixels with their neighbor, serve as an approximation of the contours of the objects. Combined with shallow classifiers, they can provide satisfactory results. A first *feature* construction step can seek to obtain information such as light direction, foreground/background segmentation, which is then aggregated into vectors of fixed size (for example bag of visual words, or Fisher vectors).[118][119]

In word processing, the counting functions (TF IDF, concurrency,...) are sufficient to classify emails as either spam or non-spam. In sound signal processing, the creation of variables linked to human perception (note pitch, cochleogram, reverberation, general statistics, etc.) are still the basis for efficient speaker recognition algorithms, for example. These

[118] J. Sivic, & A. Zisserman, *Video Google: A text retrieval approach to object matching in videos, 2003*. In *null* (p. 1470). IEEE.

[119] F. Perronnin, J. Sánchez, & T. Mensink, *Improving the Fisher kernel for large-scale image classification. In European conference on computer vision* (pp. 143-156), 2010. Springer, Berlin, Heidelberg.

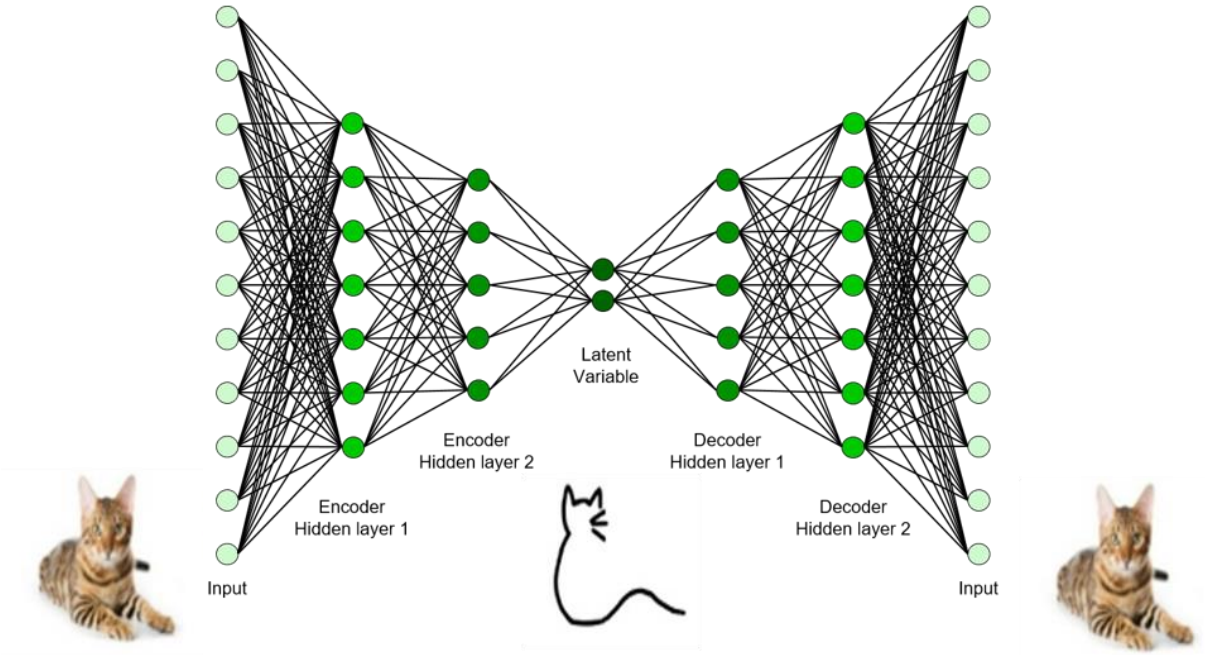
characteristics are sometimes chosen because they are easy to calculate (zero-crossing rate in audio, gradients in an image,...) and do not correspond to optimal separators for classification tasks. They also correspond, depending on the case, to professional knowledge, specific to the nature of the data and to the task to be accomplished. Beyond statistical performance, some still resist feature learning, but they are increasingly the exception. Representation learning replaces the *feature design* step in most modern approaches.



Source: LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444.

For many tasks, the choice of representation is still a difficult problem, especially in images. Object recognition in photos based on careful geometric description never worked. The challenge is to find a robust representation with all the types of angle and setting which comprise the class of the photos of a given object. The advent of neural networks stems from their ability to learn how to represent data. And it is often only once the training is completed that the representation chosen by the network can be identified. Instead of

describing the signal, or the image, we let a neural network learn the representation of the raw signal at the same time as the prediction of the target. Besides, sometimes features can be recognized in the layers of some deep learning networks. In the first layer of image-processing convolutional networks, gradient filters are recognized (recognition of horizontal gradients for example), which is comparable to the processing that occurs in the primary visual cortex (V1) in the human brain. WaveNet, a deep neural network for generating audio, learns directly from waveforms rather than spectrograms. After convergence, the construction of the time-frequency plane can be found in the first layers. This results in greater complexity in terms of the number of parameters, and therefore an increase in computing power requirements.

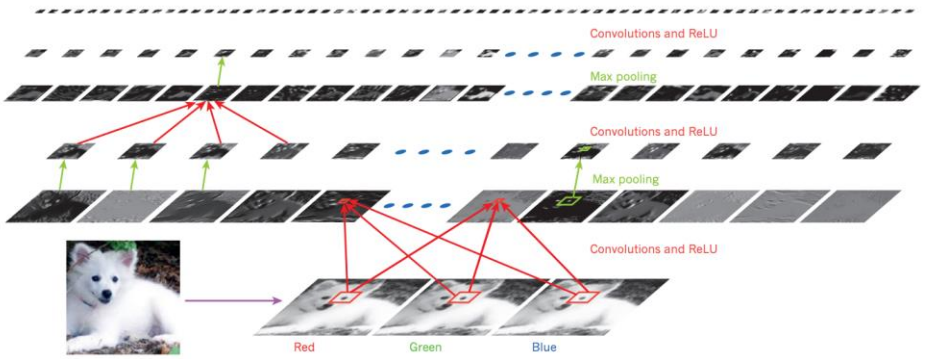


Source: *Autoencoder* by Prof. Seungchul Lee iSystems Design Lab

One way to understand the learning of features is to study autoencoders. An autolinencoder is a neural network that attempts to generate the input signal as an output. Between the two is a layer of parameters, smaller in size than the input signal. The

algorithm is therefore forced to find the minimum criteria to reconstruct the input signal from a smaller number of variables: the idea is that the most relevant structure of the data is concentrated around a space (linear or not) of smaller dimension. In single neuron layers with linear activation, we find a classical reduction of dimension, and the weights of the nodes converge towards a PCA. The advantage lies in the fact that information can be added via the structure of intermediate networks. This way, latent spaces on nonlinear varieties can be found.

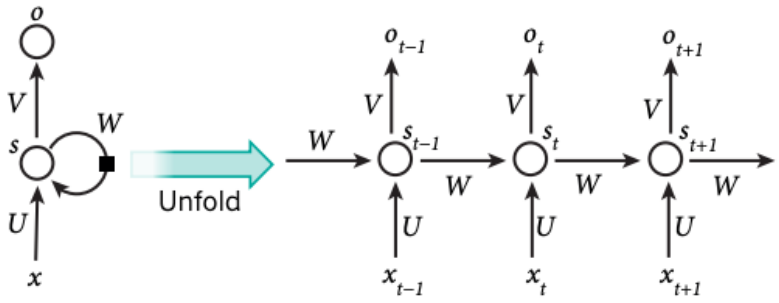
- **Recurrent convolutional networks**



Source: LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444.

Convolutional neural networks are designed to process data structured in matrices. For example, a color image is represented by three 2D pictures. The organization of the network is based on the notion of local connectivity between a large number of layers. Each layer is locally connected to the parameters of the previous layers, which allows a hierarchical organization of information: the layers closest to the input data react directly to the pixels of the images. In order to detect outlines and patterns, the patterns aggregate to form objects, and the higher layers detect global aspects. This hierarchical structure exists in language, where speech sounds form syllables, words, then sentences.

In practice, each layer is typically composed of convolutional nodes, a nonlinear activation function, and a pooling node. Pooling reduces complexity and represents the invariability of the system to local changes.



Source: LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444.

Recurrent networks are built to take into account the sequential aspect of data, particularly in speech and text processing. Recurrent Neural Networks (RNN) operate with a memory: they process input data item by item, while saving previous data in the information nodes. They are particularly effective for word or phrase autocompletion tasks.

- **Discriminative model vs generative model**

Discriminative models learn to distinguish the boundary between target categories, while generative models model the distribution of examples in each category. For example, if we know the Gaussian nature of a signal, an algorithm that learns the parameters of the distribution can be used to generate new examples which appear plausible compared to the input data. Ideally, a generative algorithm should be able to generate examples that cannot be distinguished from the input data.

Appendix 6: How content creation algorithms work

- **Generative adversarial networks (GANs)**

GANs differ from other types of deep networks mainly because their goal is the creation of new content such as images, artistic or scientific text, video, deceptively-real looking paintings, etc. In itself, this goal is more difficult to quantify than, for example, the applications for supervised classification described in previous chapters. However, effective techniques exist and have been implemented with good results. The basic building blocks are still deep networks, but the overall architecture is different.

We will focus in more detail on a category of networks known as GAN (that stands for Generative Adversarial Network) introduced in 2014 by Ian Goodfellow and his colleagues.

The idea is to make two networks interact: a so-called "generator" network (G) which is trained to propose new images, and another, the "discriminator" (D) which has access to a pre-existing database of real (“true”) images and which learns to distinguish the real images from those generated by G. The sequence of interactions is as follows: first G proposes some images which D receives with the label“false” and learns to distinguish them from the images labeled “true” chosen at random from the database. In a second phase, network D is frozen and network G is trained to create images which, once passed through network D (currently frozen) would return the answer "real image" (which it isn’t, since it was created by network G).

Then the process is repeated: D now receives slightly improved images from G with the label "false" and real images with the label "true" and practices to distinguish them automatically. Then it is again the turn of G to be trained and so on. The figure below gives an illustration of this process.

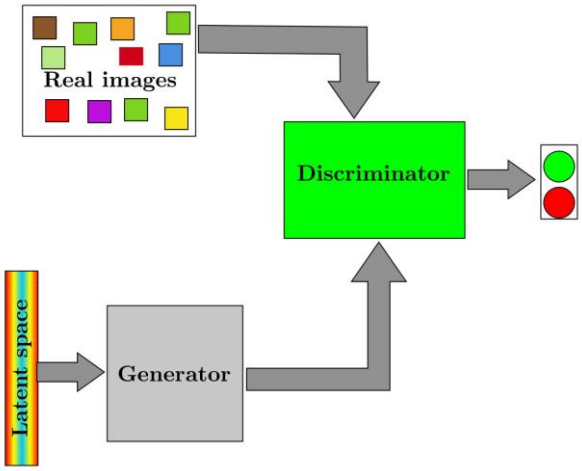
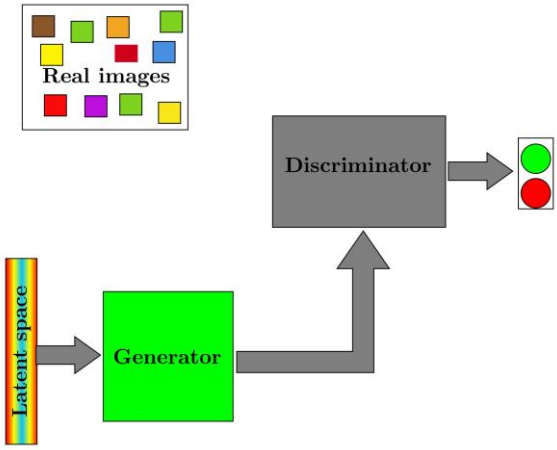
Learning phase for D	Learning phase for G
	

Illustration of how a GAN works. Active components are shown in green, inactive ones in gray. The GAN alternates between two phases: (left picture) a phase in which the discriminator is active and the generator unchanged. The discriminator learns to distinguish between real and generated images.
In the next phase (right picture), it is the generator's turn to be trained with a discriminator unchanged.

A more precise description of the mechanics of the image generation process requires the introduction of the notion of a latent space. In the same way that network D takes an image to render a verdict summarized in a YES/NO answer, the generator's task is the opposite: it takes as input a vector of numbers (in fact: taken at random according to a known distribution, for example the multivariate normal distribution) and transforms them into an image of the expected size. Each image will therefore be determined by the choice of a small set of (real) numbers which form what is called "the latent space". Each generated image has a unique

description in terms of the latent space variables. The dimension of this space is, generally, smaller than the dimension of the resulting image in order:

- on the one hand, to avoid working with a lot of information (all the pixels of an image are not essential or completely independent of each other),
- on the other hand, to still have a dimension that is not too small in order to allow the generation of fairly diverse images.

The user does not see this space and, moreover, few general indications exist to understand the precise role of each element of the latent space in the creation of the final image.

GANs are defined by this interaction between a generator and a discriminator. The possibility of randomly drawing an image or a sound is, at first glance, not surprising for an engineer since it is the essence of digital science. In fact, in a conventional computer all data is represented in binary form. The novelty is that we are now able to “shoot at random”, possibly conditionally in this latent space.

Instead, the way GAN/s work is more surprising to the mathematician: the latent distribution chosen is often a truncated unimodal Gaussian and the networks transforming this Gaussian into an image are continuous. As the image of a connected set of a continuous function is also a connected set, the support of the distribution of the images generated by a *GAN* will be connected sets, too. This assumption seems too strong for natural images. There is therefore a kind of guarantee that some of the images generated will have poor visual quality, even if in practice tricks are put in place to help the network to separate the arrival space into several blocks. A frequent technique is, for example, to teach the GANs conditionally on the subject class to be generated, which contributes to the generation of relatively uncomplicated backgrounds (blurry, solid, plant motif, etc.).

- **A large variety of GANs**

The GAN learning protocol has many variations. Recommendations have been made to improve convergence, the quality of the results, the speed of convergence, etc. This has resulted in the establishment of a whole zoology of GAN-type networks.

Despite impressive results, these networks remain difficult to train: stochastic gradient descent (SGD) algorithms run on a GAN have the reputation of never completely converging and of not always finding a satisfactory minimum. For example, BigGAN, whose training is valued at $50,000 ("public cloud" price) only produces models that are "amenable to truncation" (i.e. which can be sampled at the quality promised in the initial research paper) only 60% of the time.

In practice, when the training goes well, we observe an oscillation that is partly generative and partly discriminative. If, on the other hand, one of the two becomes more efficient than the other, learning often stops in a local minimum. Schematically, if the discriminator has become good at identifying images of real "cats" the generator has two solutions: either improve the production of its cats so that they are more difficult to identify, or never generate images of cats again. This second situation is known under the name of "mode collapse" and occurs frequently when the discriminator is too efficient: its gradients become uninformative because they vary too sharply.

It has become common not to start completely at random. Modern approaches train the discriminator and the generator on different tasks. For example, generating text probably requires the same type of skill as filling in gaps in a sentence. This is exactly the approach applied by a model released by Google called Bidirectional Encoder Representations from Transformers (BERT)[120]. Here, the generator is pre-trained by masking 15% of the words of the initial sentence in order to learn an RNN (bidirectional) capable of guessing the missing

[120] Language model developed by Google in 2018. This method has significantly improved the algorithms for automatic language processing.

words. This task, with the masking module slightly improved over a random selection, is known as the Masked Language Model (MLM).

The pre-training of the discriminator involves a task known as "Next Sequence Prediction." The idea is to build pairs of sentences from a set of texts and to train a model to determine whether or not the two sentences were consecutive (while being careful to provide consecutive phrases in 50% of the examples). Here again it becomes possible to take advantage of the supervised techniques. More generally speaking, learning tasks whose labels can be obtained without human intervention is known as "self-supervised learning" and is considered as one of the most promising avenues for 2020 by many actors in the field, including Yoshua Bengio.

As the structure of the network is also variable, the texts tend to be based on recurrent networks, while images more frequently use convolutional networks such as Radford, Metz, and Chintala in 2016 with DCGAN (Deep Convolutional Generative Adversarial Network). Today transformers are increasingly used for texts, while U-net is becoming more common in images, including in 3D. In all cases, it is common to integrate attention mechanisms into the network. These mechanisms, which have their origin in the observation of the improvement in performance in French / English translation by the reversal of the reading direction at the entrance to the network, have been progressively improved. Today networks are building representations on the basis of words located either before or after, or even automatically ignoring words. Even more surprisingly, it seems that these architectures are now able to rediscover the basic rules of grammar and identify hierarchies of concepts.[121]

There are other notable contributions. A very welcome version that made it possible to improve convergence was the introduction of "Wasserstein GANs" (abbreviated WGAN) by Arjovsky, Chintala, and Bottou in 2017. They propose to adapt a particular technical tool, the distance between laws of probability. The GANs can be seen as the generation of instances of

[121] See: https://pair-code.github.io/interpretability/bert-tree/

a random variable of a law whose unknown probability is given by some observations (examples). Yet, if we have several propositions of laws of probability, we must be able to distinguish which law is closest to the one from which the observations come. The mathematical object which quantifies this distinction is called “statistical distance”, or distance between probability laws. Arjovsky, Chintala, and Bottou propose to use a particular distance, known as the Wasserstein distance. This changes the discriminator (and its drive) slightly by imposing technical constraints on it, but has the potential to improve the diversity of newly generated images. This recommendation was refined a few months later by the contribution of Gulrajani et al. (2017) who introduced “Wasserstein GANs with gradient penalty” (abbreviated WGAN-GP) where a certain flexibility is left to the discriminator (the constraints are imposed in a more flexible way).

A unique architecture (two generators and two discriminators) was tested by Liu and Tuzel in 2016 with good results, leading to coupled GANs (CoGAN).

A little later, Karras et al. from NVIDIA found that if the discriminator is too good the generator becomes “discouraged” and offers only low-quality solutions. In order to correct this, they invented the paradigm of the “Progressive Growing Generative Adversarial Network” (also called “Progressive GAN” or even “ ProGAN ”) where the discriminator and the generator initially act only on very low resolution versions (for example 4x4 pixel images). This allows the generator to offer synthetic images that the discriminator will have a hard time distinguishing from real images. Once the generator is good at this level of resolution, the image quality is increased (8x8) and a new step begins. This is repeated several times until the generator provides high-resolution (1024x1024 or even higher) and good-quality images. It is one of the most efficient algorithms to date. Another GAN developed by NVIDIA, the “Style-Based Generative Adversarial Network” (StyleGAN, Karras, Laine, and Aila, 2019), tackles another aspect of image generation: how to generate images with facial features (pose, shape, hair color, eye color, skin characteristics, etc.) by associating each line with a different level of resolution.

This does not a full list of types of GANs: others that should be included are the "Cycle-Consistent Generative Adversarial Network" (CycleGAN, by Zhu et al., 2017), the "Big Generative Adversarial Network" (BigGAN, by Brock, Donahue and Simonyan, 2019), Pix2Pix (Isola et al., 2018), etc. And research on GANs and their variants continues to advance.

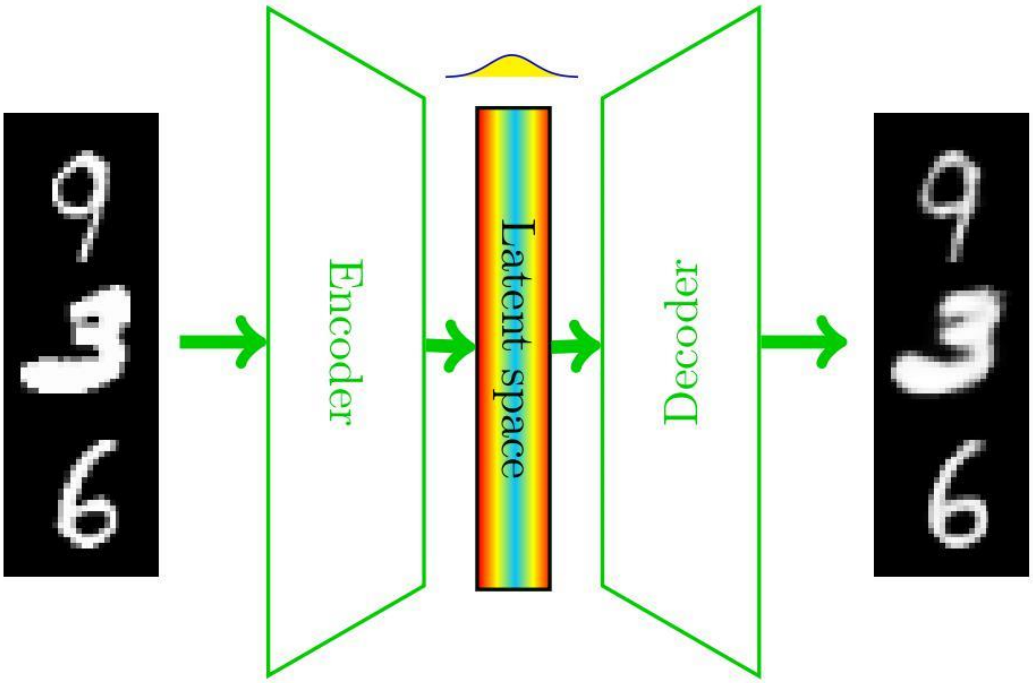
- **Variational Autoencoders[122] *(VAE)***

The quality of a GAN is based on a delicate balance between generator and discriminator. To use an analogy with the world of painters and art critics, we can see that the quality of the pictures will be better when amateur painters have the chance to improve their creations without necessarily being severely reprimanded for their youthful mistakes. Criticism must provide avenues for improvement so that new ideas can be implemented. However, in GANs, it sometimes happens that the criticism (the "discriminator") becomes too strong, thus blocking any progress. The reverse can also happen, when a star painter monopolizes all the attention. In both cases, progress is limited, and in GANs, this leads to cases where the quality of the results is not satisfactory. To overcome this defect, which is inherent in the structure of GANs, several researchers have proposed other architectures capable of generating new content. We will invoke one of them, the "Variational Autoencoders" (VAE). A VAE is a reversed GAN: it begins with a so-called encoding stage which transforms the entry into a (short) list of characteristics, called "latent space." Following this, a "decoder" stage takes a/the sample from the characteristics set and reconstructs the original example. Like GANS, it contains technical ingredients such as "statistical distances."[123]

[122] An autoencoder is an artificial neural network used for the unsupervised learning of discriminative characteristics. The goal of an autoencoder is to learn a representation (encoding) of a set of data, usually with the aim of reducing the size of that set. A VAE is a generative and probabilistic variant of the autoencoding neural network architecture.

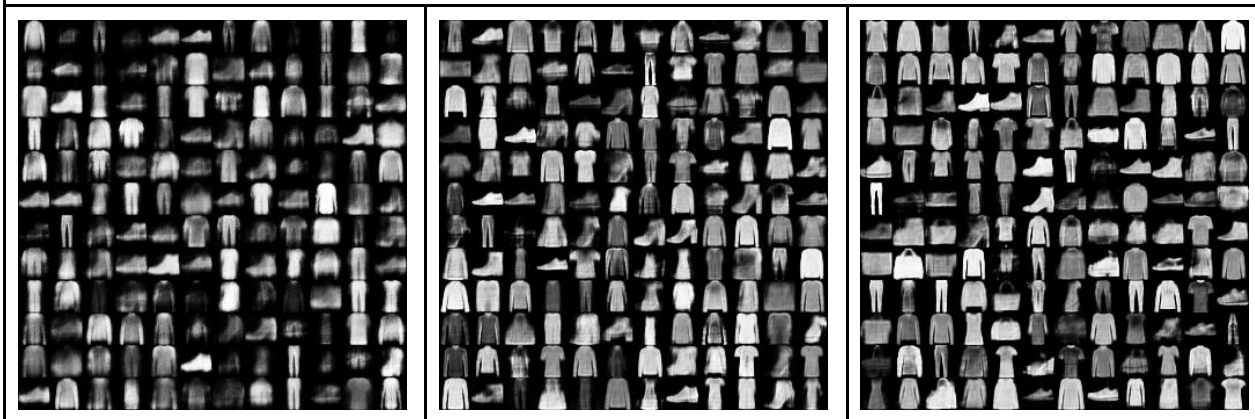
[123] See the work of Gabriel Turinici on statistical distances https://arxiv.org/abs/1911.13135

So far such networks have only been able to achieve information compression at best (similar to taking a digital image of several megapixels and reducing it to a JPEG format photo of a few kilobytes). The VAE goes further by proposing that all the characteristics (the latent space) have a standardized and known structure. If this is ever done, then instead of starting from a real image, we can very well insert in the middle of the VAE, between the encoder and the decoder, a new element of the latent space (we know how to generate it because we have said that this one now has a well-defined structure, or in mathematical language we know its law of probability; we can imagine the latent space as the interval [0,1] with a law of uniform probability, it suffices then to draw at random a number between 0 and 1; in practice it will be a multidimensional Gaussian law. The inserted latent element will be able to pass through the decoder like any other element of the initial database and thus produce new content, similar to the data used as input to the encoder and which is no longer used in the generation step (the data was used to find the optimal network which accurately encodes and decodes the initial base AND the prescribed latent structure).



VAE: illustration of the structure of a variational encoder. The input images are reconstructed at the output, after passing through a “bottleneck” in the middle. The Encoder and Decoder networks are optimized to improve the quality of the reproduction *in fine* but also the concordance with the distribution fixed in advance in the latent space. Credit: Gabriel Turinici 2019.

Convergence of the VAE for a database of images of clothing (called FMNIST). On the left: images generated after a single reading of the database (an “epoch”), in the middle after 20 readings, and on the right after 500. Notice the improvement in the apparent authenticity of the images and their variety (avoiding the danger of concentration specific to GANs).



- **Flow models**

Replacing the latent space made up of a distribution of VAEs with a simple Euclidean space allows us to rediscover an older structure known as autoencoder. These autoencoders are easier to train than VAEs (this is a content compression task), but it is difficult to use them as a distribution. The hope with flow models is to get the best of both worlds.

Learning consists of learning an invertible transformation from the distribution of the data to a distribution known as a normal distribution. Training can therefore be done without having to calculate a derivative in the latent space made up of parameters of laws. Images are generated by drawing according to the known objective law and using the inverse transformation. This idea has been derived in different ways for images leading to Glow and

FFJORD which offer more stable performance than GANs but remain a little behind once the optimization of the hyper-parameters is completed. It is, however, a very active research topic.