**Bart van der Sloot,
Yvette Wagensveld
and Bert-Jaap Koops**

SUMMARY

# DEEPFAKES:

## THE LEGAL CHALLENGES OF A SYNTHETIC SOCIETY

November 2021,
Tilburg Institute for Law, Technology, and Society

Tilburg University

# Deepfakes:the legal challenges of a synthetic society


## Summary

This is the English summary of a report written in Dutch by **Bart van der Sloot, Yvette Wagensveld and Bert-Jaap Koops** (Tilburg University) commissioned by the Scientific Research and Documentation Centre, Ministry of Justice and Security, the Netherlands.


A deepfake is content (video, audio or otherwise) that is wholly or partially fabricated or existing content (video, audio or otherwise) that has been manipulated. Several technologies can be used for this purpose, but the most popular is based on what is known as Generative Adversarial Networks (GAN). GAN has pushed the technological boundaries and has improved the quality and resolution of the material produced, with low cost and time investment. By processing, say, a thousand photos of Donald Trump, a new photo of Trump can be produced that is not an exact copy of either one, but looks authentic. The same applies to audio and video material. A video can be generated within minutes in which a person appears to be saying or doing things that she never did or say. Such video may be indistinguishable from authentic material. The main question of this research project was whether the current legal regime is capable of adequately tackling the negative effects of deepfake technology and if not, what could be options for altering the legal regime. This question was answered through literature review, legal doctrinal analysis,[1] conducting interviews with experts,[2] and analysing existing legislative approaches to deepfakes.[3] Importantly, this research project focused primarily on horizontal relations: citizens producing deepfakes about other citizens.

## Main findings

### Technical possibilities and limitations

Deepfake technology allows a user to manipulate existing material or generate new material. This may involve video, audio or text, but more broadly, may involve any type of signal or information. A simple way to fabricate a deepfake is to take an existing video of a person and superimpose another person's face. In more advanced applications, facial or body features from two or more persons are merged. It is also possible to generate images/sounds of non-existent persons. Although deepfake technology is only a few years old, the technical possibilities have advanced rapidly. A deepfake made by a professional team is already indistinguishable from authentic material. The expectation is that the advanced techniques now in the hands of professional parties will soon appear on the consumer market. Using an app or webservice will make it possible for citizens to generate fake videos or audio clips of themselves or others in a matter of seconds. It is not necessarily that the producer of a deepfake already has access to material that resembles the intended end product; for example, apps already exist that allow citizens to use an image depicting a person fully dressed to generate a fake nude image of her.

Deepfake technology is expected to take off in the coming years. The technology fits the general trend that more and more digital content is manipulated by default. Such often concerns relatively minor manipulations: video call

services that equalise a person's skin tones, audio that loses some of its higher sound registers through compression, photo camera's that filter out red tones when burning forests are captured, because they 'know' that forests are green. Yet even these smaller manipulations can be of great importance, for example in the identification of a suspect or in an online medical consultation with a dermatologist. Experts predict that in about six years' time, more than 90% of all digital content will be synthetic, i.e. material that has been wholly or partly manipulated or generated by digital means.

Experts point out that detection technologies can only pick out 65% of deepfakes. They expect this figure to go down rather than up over time. In addition, such techniques will often only give an 'authenticity percentage': e.g. the chance that this video is authentic/not manipulated is 73%. The best strategy for detecting deepfakes, experts argue, is not through counter-technology, but through human assessment of contextual information: is this something this person would normally say? Are there other sources that confirm the report? Yet, AI is also being used to create fake environments, that is, not just a deepfake video, but also fake news websites that report on it, fake Twitter accounts that discuss the video, fake Insta accounts that generate memes of the video, Wiki pages that are automatically updated or created on the subject matter, fake (CNN, BBC, etc.) new items that report on the issue. Creating a fake environment makes it difficult for both humans and algorithms to distinguish authentic from inauthentic material.

**Major challenges and societal questions**

It is clear that the dangers of deepfake technology exceed the potential benefits in nature and severity. Experts stress that the democratisation of deepfake technology may result in so much fake content that fact and fiction will become almost impossible to disentangle. If 90% or so of digital content will be manipulated, both in terms of time and resources, it will be virtually impossible for the media to systematically check all content for authenticity, assess precisely what has been manipulated in a video/picture/etc. and to what extent that is relevant for the news item. It is inevitable that more and more material will slip through the net. Either 'Main Stream Media' will accept a margin of error, meaning that they will be rightly accused of reporting 'fake news', or they will apply strict rules and procedures, meaning that they are always two steps behind media that immediately post sensational (possibly fake) news items. Fake news coverage might intensify polarisation between groups that increasingly live in their own reality.



The democratisation of deepfake-technology can also put pressure on the rule of law. First, legal proceedings will take longer, because parties can always claim that the evidence produced against them is fabricated. Such a line of defence may require further investigation and increase the role of expert witnesses in the court room. Second, the chance that judges will wrongly take content to be authentic will increase as the percentage of fake material and their apparent authenticity increases. The reverse, i.e. that a judge believes that certain material is (possibly) fake, while this is not the case, may have undesirable consequences as well. Third, a convicted person can always publicly maintain her innocence after a court decision has been issued, claiming that the judge mistakenly took fake material to be authentic. Fourth, with certain crimes, a mere (fake) suggestion can be enough to cause public

outrage and lead to a public conviction, even in absence of a legal conviction by a court of law.

The democratic process may be in jeopardy too. In particular, reference can be made to incidents in which foreign powers appear to be using fake messages and trolls to influence elections. A number of countries and several states in the United States have already passed legislation on this point. It is also clear that national groups are using deepfakes, e.g. to smear their political opponent. Importantly, experts point out that states also try to influence elections and concrete decisions in the Global South. This may have a highly disruptive effect on the international legal order, for example when a country succeeds in putting in place friendly regimes in the Global South, so as to have their support in important votes in the various bodies of the United Nations.

Deepfakes have a big impact on the social safety and societal position of women and young girls. Many of the deepfakes involve non-consensual sexual images or videos of women, leading to the sexualisation of the female body, confirming unrealistic beauty standards and stigmatising women. Slut-shaming and misogynistic remarks are already commonplace offline and certainly online, something that deepfake technology will only exacerbate. One of the existing problems is that private recordings of sexual acts are made public by an ex-partner - so-called revenge porn. With deepfake technology, such private recordings are no longer necessary; any adolescent boy can generate a fake porno of any of his classmates and distribute it among friends or on social networks.

Experts point out that knowing that certain material is inauthentic is only of limited importance. The social consequences of a porno of an adolescent girl can be considerable, even if her classmates know it is a deepfake. Seeing such a film can also affect the self-image of the girl in question; watching herself perform all kinds of explicit actions can have a negative impact on her self-confidence and self-esteem. This also applies to fake news. Even if a news item is later debunked, groups often still maintain that although a specific message may have been fake, the underlying truth was correct. In addition, the initial (often sensational) fake message will often generate significantly more attention than the subsequent nuance or correction. Even if a person has read the correct story, debunking the initial fake news, she is often left with a 'wasn't there something with...' feeling. Finally, a person that knows about the existence of deepfakes or has mistakenly believed in the authenticity of one or more deepfakes in the past, may become guarded or sceptic when seeing authentic news coverage.



Besides these bigger, societal challenges, deepfakes are used for all types of crimes and malignant activities. Inter alia, deepfakes can be used to incite hatred and violence, for example against minorities, can be used to circumvent and undermine the intellectual property rights of artists, can be used to commit fraud or identity theft and can be deployed to smear a person's reputation.

Finally, deepfakes trigger moral questions. For example, a deceased artist can give a tour of a museum; Napoleon can give history lessons at secondary school; family members can see what their great-grandmother would probably have looked and sounded like; a deceased person can speak at her own funeral; a deceased singer can give concerts (Elvis is back); and partners can

stay in touch with their late spouse by using a deepfake of that person. Questions include, but are not limited to: When historical figures lecture schoolchildren, wouldn't that contribute to habituation to the post-truth world; will continuing to converse with a deceased partner lead to psychological problems of its own; did great-grandma really want to be brought back to life? Similar questions arise in other deepfake applications, such as when the police use fully fictitious persons to track down child pornography networks, women traffickers and organised crime. How far can and should the police go? And when a politician addresses a national minority in their own language, using deepfake technology, is that desirable in terms of inclusion or is it a form of deception?

**Limited interests in horizontal relations**

There are also many positive applications of deepfake technology. These include the previously mentioned possibilities for the police to use fakes to infiltrate criminal networks, to anonymise witnesses and, for example, to catch pederasts through producing fake child pornography. Fake child pornography can also be used for the treatment of convicted paedophiles. There are also medical applications that run on deepfake technology, for example for people who have a distorted self-image. A crime scene can be reproduced by means of a deepfake, a realistic deepfake avatar of a person can be used in a game, a store can give a deepfake impression of a customer's house with a brand-new kitchen in it, a deepfake of an actor can perform dangerous stunts, women working in the sex industry can use a deepfake to perform certain (extreme) activities and deepfakes can be used in the retail sector, for example by showing a deepfake of a person with new clothes or glasses on. Deepfake technology can also be used to ensure that the audio of movies are translated into English, while at the same time synchronising the actors' lips so that it looks as if they are indeed speaking English. This application can also be used by politicians who want to address minorities, by celebrities who make appeals for charity in every language of the world and in video calls between, for example, French and Chinese employees.

Most of these positive use cases play a role in professional settings, such as when used by the retail sector, for business conversations, in the entertainment industry, for giving guided tours in museums and when politicians use the technology to give speeches in multiple languages. This study has identified only one common positive application of deepfake technology in horizontal relationships and that is its use for satire (e.g., videos have been produced in which Nicolas Cage appears to play in literally every movie ever made), though it cannot be excluded that in the (near) future, other positive use cases in horizontal relations will also become commonplace.

**Technology is not neutral**

Sometimes it is argued that technology itself is neutral. Deepfake technology itself is neither good nor bad, so the argument goes, it is what people do with the technology. Hence, it is suggested, it is not the technology itself that should be regulated, but the harmful ends they are put to use for in concrete cases. This understanding of technology can be heard most often in the United States and is summarised in the NRA's credo: guns don't kill people, people kill people. Others, however, suggest that technology is never neutral; technologies are developed and designed with a specific purpose in mind.

Though it is not exclude that technologies are used for other use cases than for which they were designed, mostly, other technologies are more suitable for those purposes. For example, a vegetable peeler has been so designed to make it suitable for peeling fruit and vegetables, a hammer is not. Consequently, more than 99% of the cases in which a vegetable peeler is used concern the peeling of vegetables and fruit. This point is relevant because research shows that more than 95% of the deepfakes concern so-called non-consensual porn. The term deepfake was initially used exclusively for this practice. That is why experts have suggested to include this use case in the definition of deepfakes. In addition, what is intrinsic to deepfake-technology is that it increases the confusion between the authentic and the inauthentic, between fact and fiction. It is also in this sense, experts say, that deepfake-technology is not neutral.

At the same time, it is important to emphasise that the way in which deepfakes have been used ties into broader social trends. Deepfake porn films are in fact a consequence of the disrespect for women and objectification of the female body that is rampant offline and certainly online. The rise of deepfake misinformation is but an emblem of the post-truth era. The use of deepfakes for political purposes dovetails with an increase in interstate hostilities via digital means. Fraud and identity theft have been committed for centuries and deepfakes are but the next means of introducing false evidence in court cases. Moreover, the fear of a 'post-truth' world has been around for centuries and resurfaces with the introduction of every new technology, such as with the printing press, the Internet and Virtual Reality. The introduction of a new technology is always followed by a period of chaos, after which legal, social and institutional norms are developed to steer the use of the technology in the right direction. In this sense, deepfakes are nothing new.

The novelty and potential danger of deepfake-technology lies in two aspects, one qualitative and the other quantitative in nature. On the one hand, deepfakes seem so real that they are more likely to be taken for granted. People have a so-called 'truth-bias', they assume something is true unless there are contraindications. This certainly applies to video images. The second, and perhaps more important, difference is the democratisation of the technology. The expectation of all those interviewed for this study was that the technology would be in the hands of ordinary citizens within two or three years and that it would be used enthusiastically. Free apps are already available and, in their opinion, these apps would only get better and faster. The production of a very realistic deepfake could then be done in no time by almost any citizen in the world.



**Enforceability**

Perhaps the most important insight regarding the current legal framework is that although amendments are possible and perhaps desirable on specific points, such would not tackle the main problem with regard to deepfakes in horizontal relationships and, more generally, to breaches of privacy in horizontal relationships. In the first, second and third place, the problem is one of enforceability. Producing pornographic material of another person without her consent is already prohibited; generating child pornography of a fictitious child is already prohibited; committing fraud and deception by means of a deepfake is already prohibited; introducing false evidence in a court case is already prohibited; inciting hatred or violence between groups is already prohibited;

exploiting someone's image or likeness or creative works without permission is already prohibited; causing (economic) harm by means of identity theft or reputational harm by fake messages can already be dealt with under tort law; etc.

The legal framework applicable to deepfakes is not the primary problem; the problem is the enforcement of the existing and any additional legal rules. There are a number of obstacles. First, technology is developing rapidly, so that technology-specific rules will become outdated quickly. Second, it is also difficult to define any technology for legislative purposes, because a too narrow definition leaves too much room for undesirable applications or for techniques that are modified in such a way that they do not fall under the definition, while too broad definitions negatively impact useful technologies or positive use cases. Third, due to the cross-border nature of data-driven technologies, parties are often subject to multiple legal regimes, and tend to locate in the jurisdiction with the lowest regulatory burden. Fourth, it is difficult to impose and adequately enforce the rules of one jurisdiction on parties located in other countries. Fifth, there is often a complex web of parties involved with the production and distribution of deepfakes, all sharing partial responsibility. Sixth, it is often easy to circumvent rules of a particular jurisdiction, for example by using a VPN connection.

The democratisation of data-driven techniques, including deepfake technology, challenges the current regulatory framework which places emphasis on ex-post regulation (the development, distribution and possession of technologies are often left unregulated, while content is only checked for legitimacy after it has been made public). The choice for ex-post regulation and the democratisation of technology means that per day, millions of pictures, video's and audio fragments are put online. It is impossible for any governmental organisation to assess their lawfulness. That is why their attention and energy almost exclusively goes to the more extreme violations of law, leading to a normalisation of minor (privacy) violations.

Although every citizen has various rights, it is by no means always clear to someone that her data are or have been collected or that a deepfake of her has been distributed on the internet (for example on a porn site). Even if she does know or learns about it, it is not always clear who can be held accountable. In order to find out the identity of the perpetrator, the cooperation of Internet intermediaries is often necessary, while they are not always willing to cooperate (without court order) because of the privacy interests of the person who posted the material. This means that two lawsuits are often necessary, one to find out the identity of the perpetrator and another to sue her. If this includes a request for removal from the platform or from any copies published elsewhere, a third, fourth and subsequent lawsuit may be necessary. This requires time, money and energy that citizens often lack; the amount of compensation awarded by a European court if a citizen is successful is generally low, typically some hundred euro's.



### Regulatory options

This study has identified several regulatory options. A number of caveats apply:

- 1. The regulatory options are not recommendations, but options; their desirability and feasibility will have to be subject to further research and political/societal discussion.

2. Some of the options can be implemented immediately, others require structural changes to the legal system, and still other options are controversial or have major potential negative consequences and require further research.

3. The regulatory options should be considered in relation to each other. Several options address the same underlying problem; if one option is implemented, others may be omitted. The various options discussed with respect to procedural law all deal with the same underlying problem; they can partly be seen as complementary, but introducing all of them will presumably be too much. Regulatory options 1 and 5 address essentially the same underlying problem, namely the fact that processing personal data of others in the private sphere is currently left unregulated. Again, both options could be introduced and considered complementary, but introducing one of them may also suffice.

4. Many of the problems described and potential solutions offered are related to general, societal trends. Sometimes it is possible to adopt specific rules for deepfakes; often, however, it seems advisable to address the underlying problem as such.

The regulatory options are divided in three groups: amendments to substantive law, amendments to procedural law and amendments to ensure the effective enforcement of existing rules.

## Substantive law

### Criminal law

Substantive criminal law is applicable to most harmful use cases of deepfakes, such as when they are used for identity theft, fraud or the distribution of non-consensual porn. However, when deepfake sex videos are not distributed, but are made for purely personal use, this does not fall under a penal provision. A criminal provision could be introduced to that end, which could possibly also apply more broadly to any deepfake that could be deemed intrinsically harmful.

Regulatory Option 1:

Consider whether the making or possession of "intrinsically" criminal (morally reprehensible) deepfakes should be criminalized.

In addition, in Dutch law, a potential loophole concerns the gap between Article 231a Criminal Code (CC), which criminalizes identity fraud with biometric data that are processed for purposes of identification, and Article 231b CC, which criminalizes harmful identity fraud with non-biometric data. Consideration could be given to amending article 231b CC by deleting the clause 'not being biometric personal data', so that it would also include deepfakes that are used for identify fraud in situations where biometric data have no identification purpose.

Regulatory option 2:

Consider amending article 231b CC by deleting the clause 'not being biometric data'.

### Privacy and data protection

Post-mortem privacy has been discussed for decades, but especially in the last few years, the debate has gained momentum. In principle, if a person dies, her data do not fall under the regime of the EU General Data Protection Regulation (GDPR). Yet many citizens do not want their data to be released after their death; for example, they want their emails destroyed and prevent them

from being commercialised by private parties or from falling in the hands of their heirs. But because the GDPR no longer applies, companies often claim they own the data and continue to process and use them while next of kin may want to exploit the data instead, for example when the deceased is a famous artist.

Deepfakes take this discussion to a new dimension, both morally and commercially. In principle, there is nothing in privacy law to prevent a deceased person from being brought back to life, whether she wanted to or not, for example by having her speak at her own funeral or communicate with her next of kin on a daily basis, long after she has gone. The legal regime in most countries does offer protection to the rights/interests of the dead, for instance, by regulating in detail what can and cannot be done with a body. Such rules concern the physical body, but there are no rules on the virtual body or the realistic reproduction of a person's psyche. There are numerous applications that could be the subject of political debate. For example, is it desirable and permissible to have long gone historical figures teach in schools? Is it desirable and permissible to have deceased artists give a tour of a museum? Is it desirable and permissible to have deceased actors feature in films? Is it desirable and permissible to have a deceased person star in a porno film? Is it desirable and permissible to have deceased artists still give concerts?

Regulatory option 3:

Develop laws or regulations regarding post-mortem privacy

The creation of non-existent persons through AI also raises numerous ethical dilemmas. The police may infiltrate a criminal network using a fake person, pederasts can be traced using fake child pornography and traffickers in women can be identified using fake customer profiles. For these applications as well as for others, more clarity is needed as to what is or is not permitted in terms of the creation and deployment of fictional but highly realistic characters, not only by the police but also within the entertainment industry, within the porn industry or for medical applications. For example, there are therapies for the treatment of paedophiles through the display of fake child pornography, but is that desirable? What are the moral boundaries to producing fake personalities and the activities that they can perform?

Regulatory option 4:

Develop laws or regulations on the use of fully AI-generated individuals

Finally, an adjustment is possible with respect to the household exemption. The exemption, which dates back to the 1995 EU Data Protection Directive, was already under discussion when the GDPR, replacing the Directive, was crafted, yet was left virtually unchanged. The exemption provides that when personal data are processed in the private sphere, the data protection rules do not apply. When the 1995 Directive was adopted, the primary example for the need of a household exemption was keeping an address book. Such does concern processing personal data of third parties, but only involves their name, address, and telephone number. Keeping such data is socially accepted and usually desired by the third parties in question. Currently, however, citizens have access to a wealth of information and can use various advanced data processing

technologies. This means that the type of data that can be processed about others with the use of a private computer is incomparable to those thought of when drafting the 1995 Directive. In addition, the household exemption made sense in a world in which the private sphere was more or less closed off from the public sphere. In the current data-driven environment, the boundary between the two spheres has become increasingly blurry. A click of a button is enough to disseminate thousands of photos or videos stored on a private computer online through social media or digital platforms.

The household exemption raises the following problem. Suppose an ex-partner stores private photographs of his ex-girlfriend on his computer, with which he then produces a deepfake in which she performs all kinds of perverse sexual acts. He tells his friends about it, who also communicate this to her. This is just one of the many possible examples of deepfake applications that cannot be addressed under the GDPR. The production of compromising material and the possession of it, is not covered by the GDPR. Once the material is on the internet or distributed to large groups of friends it is, but by then it is too late. The damage has already been done; compromising videos can attract thousands or millions (in the case of celebrities) of viewers within a few hours. It may often be impossible to take that video down permanently, because of the ease with which a copy of the video can be produced.

Consequently, it could be considered to limit the household exemption. However, it should be borne in mind that although this reduces the problem of enforcement because the production of malicious content and its distribution to a large audience could be countered, it also raises a new enforcement problem. How can it be ensured that all standards are enforced effectively in the private sphere? If the government indeed would endeavour to do so, wouldn't the cure be worse than the disease? Hence, a societal and political debate is necessary before regulatory steps are taken on this point.

Regulatory option 5:

Consider whether and to what extent revising the household exemption is desirable

### Freedom of expression and the right to reputation

Deepfakes produced by citizens can broadly be divided into two groups. Deepfakes that depict themselves or people they know, such as their partner, children, parents, neighbours and friends, and deepfakes that target public figures, such as actors, singers, politicians, civil servants, royalty, and journalists. In general, citizens are more likely to have access to (private) material of persons in the first group while, in general, they are more likely to obtain the material used for creating deepfakes about persons in the second group from public sources on the Internet. In general, consent can be sought more easily from persons in the first group, while it is generally more difficult with respect to those in the second group. Generating deepfakes about persons without their consent may be legitimate under the GDPR if the deepfake is unharmful and/or serves an important interest. The GDPR prohibits the production of deepfakes without consent when 'sensitive data' (data concerning a person's medical condition, political believes, sexual activities, race or ethnicity, etc.) are processed, but an exception may apply in light of the freedom of expression.

Under the European Convention of Human Rights, public figures (as well as ordinary citizens) can invoke their right to privacy in order to protect their name, honour and reputation, even when they actively seek the limelight. Yet the European Court of Human Rights has also ruled that public figures must tolerate greater intrusion into their private lives than ordinary citizens and must accept that they will be mocked and ridiculed. The relationship between the freedom of expression of citizens and the right to privacy of public figures is assessed by the Court on a case-by-case basis. Consequently, there are few general rules and prohibitions on expressions about public figures. This means that public figures have little legal certainty when going to court over potentially unlawful expressions. The result is that legal action is rarely taken, resulting in the normalization of extreme expressions.

Increasingly many qualified men and especially women refrain from entering politics or leave public administration for precisely this reason. Deepfakes will only deepen this problem. Politicians may appear to be giving a speech drunk, mayors may appear to make racist remarks, police officers may appear to be using excessive force, civil servants may appear to be breaking the rules that they are supposed to oversee, a head of state may appear to give an obscene Christmas speech, and female public figures will be depicted in explicit movies. In time, this may have a detrimental effect on the functioning of the government and public administration. Therefore, it could be considered to adopt rules in order to protect the reputation and honour of public figures more effectively, while the importance of a free and open societal debate should not be lost out of sight.

Regulatory option 6:

Develop laws or regulations on the protection of the reputation and honour of public figures

One fear regarding deepfakes is that they will take the post-truth era to the next level. An untrue, inaccurate or misleading statement can be addressed under the current legal regime, but only if damage has been caused, for example to personal interests (under tort law) or to certain social interests (under criminal law). This raises three issues. First, it can be difficult to substantiate the causal relationship between an untrue, false, or misleading statement and the (foreseeable) harm it causes (e.g. the hatred a deepfake has incited against minority groups). Second, untrue, inaccurate, or misleading expressions can be problematic because they blur the line between fact and fiction, even if they do no concrete harm. Third, there are untrue, inaccurate, or misleading expressions that do cause harm, but that are very difficult to link to a specific legal provisions. For example, fake satellite images may be produced in which Russia appears to move its nuclear missiles near the Latvian border, creating political tensions. Or, fake news may be distributed on Covid-vaccines, leading to a decline in people that want to get vaccinated. Or, a political leader may distribute a video, making it look like there are thousands of supporters at her rallies, while there are a handful only in truth.

These developments may force the government to choose between Scylla and Charybdis, between staying clear from these complicated issues, which may mean that the problem of misinformation will grow, and adopting regulation on untrue, false or misleading statements. Orwell warned for such governments, the EU tried to play a

more active role in addressing online fake news, which backfired more quickly than even sceptics had predicted. Yet the government need not take an active role on this point, even if it would adopt regulation. It could, for example, leave the ultimate decision to a court of law. Neither would a judge need to enter into complex questions regarding what is true and what is not per sé. It could err in favour of dubious statements and only address obvious untruths such as those that were spread during the Corona crisis. Yet obviously, adopting such regulation is a sensitive political issue and should be subject to debate.

Regulatory option 7:

Develop laws or regulations regarding manifestly untrue, false, or misleading statements

Several states of the U.S.A. have passed laws on the dissemination of misinformation during elections. Those laws are linked to the phenomenon of deepfakes, but have a broader material scope. Spreading of misinformation by foreign actors is difficult to address through legal action, yet when civilians or national groups engage in such activities, either civil law actions (e.g. by a politician or political party being portrayed in a deepfake) or a criminal law action could be considered. Such rules could be extended to influencing political decisions as such. Potentially, rules could also be adopted for attempting to influence foreign elections or political decisions.

Regulatory option 8:

Develop laws or regulations on influencing or attempting to influence elections or political decisions through the production or dissemination of misinformation

**Procedural law**

The assumption in most legal procedures is that material is authentic, unless there are contraindications. In practice, this means that it is mainly up to the defendant in a criminal case, or, in a civil law case, the opposing party, to state and make plausible that evidence was manipulated or fabricated. This may be problematic. On the one hand, it entails a privatization of a general problem. On the other hand, citizens will not always be able to challenge the veracity and accuracy of evidence (e.g. suspects convicted in absentia, persons with mental disorders). In addition, it may be costly to obtain the technical expertise necessary to demonstrate that evidence is or may be inauthentic, which may cause problems for those in economically disadvantaged positions. Although most jurisdictions already have a rule that require parties to only introduce authentic evidence in court proceedings, it became clear during this study that these provisions are only marginally applied in practice and seldom lead to substantial sanctions. Consequently, it could be considered to impose obligations or duties of care on parties other than the citizen, to ensure that the public interest in having legal cases decided on the basis of authentic material does not depend solely on assertive and well-to-do citizens. Three parties could potentially play a bigger role: the lawyer, the police/public prosecutor, and the judge.

- The lawyer has a professional duty of care to ensure that only authentic evidence is introduced in court proceedings. At the same time, she must put forward her client's interests and present the client's version of the truth. This will not always be the version that the court will ultimately accept. This means that the question to what extent a lawyer may bring forward arguments that later prove to be untrue cannot be answered unequivocally; this also applies to the question to what extent she should have an obligation to check the authenticity of the evidence, for example, that she has received from her client. Such a duty, moreover, would entail substantial investments in terms of time and resources, which would make a legal case more costly for citizens that already face substantial economic barriers. Finally, not in all cases are citizens represented by an attorney. Nevertheless, the option of imposing further duties on lawyers was mentioned several times during the interviews, particularly because the lawyer, as a professional party, is considered to play a central role in the proper conduct of legal proceedings. Therefore, one option is to impose further duties of care on lawyers, another is to look at how the current obligations can be enforced more effectively, and finally, the Bar Association could draw up guidelines on the verification of evidence submitted in court proceedings by its members.
- In addition, a bigger role for the police and/or the public prosecutor could be explored in criminal cases. This may be desirable because a judicial investigation or prosecution for criminal offenses in itself can have a major impact on the suspect, her personal life, and social status. For example, an obligation can be imposed that the police may only start investigations and/or the public prosecutor may only start legal proceedings against a person when evidence against her has been checked on authenticity by an independent expert or institution.
- Similarly, in light of the fact that in time, more than 90% of all digital content may be manipulated, an obligation could be introduced for judges to have any material introduced in a legal proceeding checked on authenticity.
- Finally, it has been suggested that tougher sanctions could be introduced for citizens that introduce manipulated material either as a party in a civil law case or as a defendant in a criminal case. Such could entail a specific prohibition on the introduction of evidence which the citizen knows or should have known to be manipulated through the use of deepfake-technology. Alternatively, it could be assessed how existing legal provisions to only introduce authentic evidence in court proceedings could be enforced more effectively.



To facilitate this process, two suplementary options have emerged during this study, namely, first, to set up an institute that can check material for authenticity and, second, to work with an obligation on all litigants to supply only watermarked material.

- Concerning the first option, many countries already have an independent institute that can provide expert opinions on the reliability of technical, biometric and digital evidence (e.g. DNA). Such an institute could be given more powers and resources to also assess digital evidence for manipulation and fabrication. It should be borne in mind that such an institute will most likely only give 'authenticity' percentages - for example, the chance that this evidence has been manipulated is 29%. This may give rise to new legal questions and standards; experts predict that this will have the effect that supporting evidence becomes more important.

- As to the second option, the suggestion has been made several times to only allow evidence in court rooms if it is accompanied by proof of authenticity. This regulatory option requires further elaboration and can only be introduced in the medium term. One example that was suggested is that emails are very easy to manipulate or fabricate. It is conceivable to develop an e-mail system that can produce a non-manipulable reproduction of one or more emails with an authentication stamp. Only such emails could then be added as evidence in a court case. By analogy, the same approach could be adopted with respect to other digital material. Such systems are, however, unavailable at the moment. In addition, it is likely that if such systems would be developed, they would be offered by private parties, which raises numerous questions about their responsibility, the standards chosen, and the reliability of these private parties.

Regulatory option 9:

Assess to what extent laws, regulations or policies should be developed or amended to combat (deep)fake evidence in court proceedings

**Enforcement and oversight**

It is clear that a choice to make adjustments to substantive/material law only will be equivalent to accepting that most unlawful deepfakes will continue to be produced, disseminated and consumed. That is why several options could be considered in in order to ensure more effective oversight on and enforcement of existing and future legal provisions.

**Banning**

Banning technologies and products is, quite rightly, a sensitive legal and political issue. Every technique has positive applications, it is only when a technology is frequently used that new, previously unforeseen possibilities are discovered and both citizens and companies generally want to have access to new technologies. This will be no different for deepfakes. Nevertheless, banning deepfake-technology for horizontal relations is a serious option, given the societal challenges involved with the democratisation of this technology and the fact that the only positive use case for this technology in horizontal relations is its use for satire. Such a ban would mean that the technology could be deployed within professional settings, such as by the police, the film industry and the retail sector, in the medical domain and for business-to-business applications as deemed desirable. Only citizens/consumers would be excluded from access. The question is who would be the primary norm addressee of such a ban:

- A ban on the production of deepfake technology seems difficult to apply, if only because the technology is developed around the world and such a ban would be impossible to uphold.
- A second option could be to prohibit providers from selling or making available deepfake technology or applications to consumers. Still, this option too raises numerous questions. Would such a ban be targeted at all parties around the world or only at the major app stores and service providers? Would such a ban target apps or services only that exclusively offer deepfake functionalities or any app or service that offers such as one of the many features? How will deepfake technology be defined/delineated and will the definition include reference to specific technologies (e.g. GAN)

or more generally to technology that allows the manipulation of material; what would this mean for the risk of under- or overregulation? And how about open access deepfake technology hosted from third countries?

- Alternatively, access providers could be obliged to block specific sites or services, yet this would presumably only result in a cat-and-mouse-game because sites and services would simply relocate to different domains.
- Finally, the regulatory burden could be placed on the citizen, for example, by prohibiting the downloading, possession, or use of deepfake technology or technology that can be used to manufacture deepfakes. Two issues should be noted, however. First: is it desirable to place the regulatory burden on the citizen? If a deepfake app is accessible in an app store, for example, the average citizen will assume that it is legitimate to download and use the app. Second: will such a ban be effective?

Regulatory Option 10:

Develop laws or regulations that prohibit the production, distribution, possession or use of deepfake technology

**Ex-ante legitimacy test**

Alternatively, instead of blocking the production or use of deepfake technologies, rules could be developed that impose an obligation to do an ex-ante test of legitimacy before content may be distributed among friends or put online. Again, the question is on which party such a burden should be placed:

- The most obvious candidates would be internet providers who host or distribute deepfakes. More and more services and websites already prohibit (certain forms of) deepfakes from their platforms, as made clear in their Terms and Conditions. A specific duty could be imposed on providers to use deepfake detection technologies. These techniques would not filter out all deepfakes, but a substantial part in any case. Two questions are important in this respect. First, if a detection system suggests that material could be fake, should a company automatically block such content, or are some deepfakes allowed? If the latter, should the legislator provide further clarification on what type of deepfakes are or are not permitted, or is that left to providers, with the possible consequence that different providers apply different sets of rules? Second, must there be human involvement/assessment when material is blocked? The requirement for human intervention can put a substantial financial burden on companies, but would not be unique, as a company like Google had to invest substantially to handle all the right to be forgotten requests. Automatic content selection has three main drawbacks. First, that such a system is always both under- and over-inclusive, and second, that detection techniques, mostly give a 'truth percentage' only and this raises the question at what 'truth' or 'authenticity' percentage content should be allowed. Finally, it may also block content that is authentic, but has been manipulated in such a way that artefacts from deepfake technology are put on it.
- Ex ante legitimacy tests could potentially also be performed by the Data Protection Authority (DPA). To this end, citizens would have to submit their deepfake or particular application to the DPA before distributing or publishing it; the DPA could then check the content for compliance with the GDPR. However, it is unsure whether all citizens will actually adhere to such an obligation should it be introduced and it is equally unsure whether the DPA has the manpower (or the will)

to check all deepfakes for GDPR compliance. It could only work if, as an effect of the introduction of such an obligation, citizens would stop producing deepfakes on a large scale.

- Alternatively, an obligation could be imposed on citizens to carry out a Data Protection Impact Assessment (DPIA) themselves. In the event of an identified high risk, they would be obliged to inform the DPA and request its advice on whether or not deepfakes may be distributed. This would mean that the DPA and/or judge ruling on the legitimacy of a deepfake after a complaint could base its decision on the DPIA. If citizens cannot produce a DPIA, this in itself would mean a violation of the GDPR. Thus, this would relieve them from too much administrative burden. Yet again, it is both questionable whether it is desirable to impose such obligations on citizens and whether such an obligation will prove to be effective in practice.

Regulatory option 11:

Develop laws or regulations regarding mandatory and prior verification of content by hosting providers, citizens and/or DPA's

**Awareness**

Finally, an awareness-raising campaign could be launched.

First, a public campaign could highlight new or existing moral and legal standards for the use of deepfake technology. For example, it could be made explicit to young men in particular, but perhaps to men in general, that producing and distributing fake-porn is inadmissible and unlawful. Secondly, positive applications of deepfake technology could be highlighted. Third, attention could be drawn to the possibilities for victims to protect themselves against deepfakes. Women in particular could be informed about possibilities for removing deepfakes and legal actions. However, experts caution for 'victim-blaming' and the privatisation of a societal problem, by making victimised women responsible for removing harmful content. Fourth, attention could be drawn to the existence of deepfakes and manipulated content in general. Journalists and judges should be conscious of how realistic manipulated material may appear. Citizens may also be warned against dangerous fake material, for example by suggesting as a standard that one source is no source when they see sensational news or are asked to transfer money to relatives.

Regulatory option 12:

Launch a public campaign providing information on the dangers of deepfakes, making new social and legal norms explicit and highlighting best practices

## Footnotes summary

- 1 European privacy and data protection law and the regime for freedom of expression, the EU Digital Services Act and the EU AI Act and the rules contained in Criminal Law, Intellectual Property Law and Tort Law as well as Criminal and Civil Procedural Law were studied.
- 2 Judit Altena-Davisen (Criminal law; Netherlands); Margreet Ashmann (Civil law; Netherlands); Ruth de Bock (Civil law; Netherlands); Jacquelyn Burkell & Chandell Gosse (Information & Media Studies; Canada); Manon den Dunnen (National Law Enforcement Agency; Netherlands); Serena Iacobucci (Behavioral Economics; Italy); Tyrone Kirchengast (Criminal law; Australia); Andrei Kwok Onn Jui (Management; Malesia); Hao Li (Computer Scientist; United States); Sophie Maddocks (Media & Communication; United States); Emma Perot (Commerical law; Trinidad & Tobago); Lonneke Stevens (Criminal Law; Netherlands); Aya Yaldin (Politics & Communication; Israel); Mika Westerlund (Technology Innovation Management; Canada); Christopher Whyte (Political Science; United States).
- 3 In particular, two country reports were written, both in English, one on the regulation of deepfakes in China by Bo Zhao and one on the regulation of deepfakes in the United States by Andrew Roberts.

Design by Julia Janssen

Summary

Bart van der Sloot, Yvette Wagensveld and Bert-Jaap Koops